MATH 7339 - Machine Learning and Statistical Learning Theory 2-He Wang

Section Bayesian Statistics

1. Basics of Bayesian Statistics
2. Bayes Classifier
3. Bayesian Inference
4. Bayesian decision theory
5. Bayesian Regression (OLS, Ridge, Lasso
6. Bayesian logistic regression/Laplace Approximation
7. Bayesian Model Selection

Thomas Bayes
1701–1761

➢ Review Central Limit Theorem, Hypothesis Tests, and Confidence Interval. (Frequentist Point of View)

The **sample mean** of test statistics $x^{(1)}, \ldots, x^{(n)}$ is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

Central Limit Theorem (CLT): Assume that the distribution of test statistics $x^{(1)}, \ldots, x^{(n)}$ is drawn independently from the same distribution (iid) with fixed mean μ and variance σ^2, then the sample mean follows normal distribution:

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \xrightarrow{n \to \infty} N(0,1)$$

or

$$\bar{x} \xrightarrow{n \to \infty} N\left(\mu, \frac{\sigma^2}{n}\right)$$

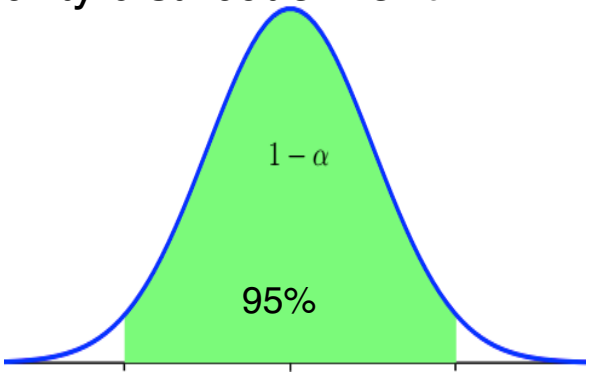
Confidence Interval

Suppose we know the distribution for θ.

The $1-\alpha$ confidence interval for any parameter/statistic θ is a set $I = [a, b]$ such that

$$P(a < \theta < b) = 1 - \alpha$$

Probability distribution for θ



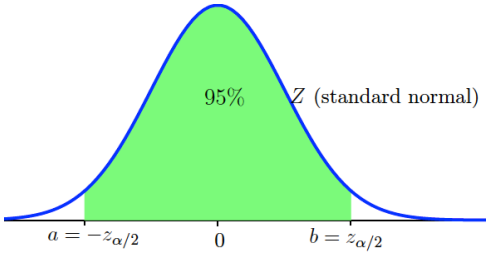
"There is a 95% probability that $[\tilde{a}, \tilde{b}]$ calculated from a given future sample will cover the true value of the population parameter."

"Repeated the procedure on numerous samples, the proportion of calculated $[\tilde{a}, \tilde{b}]$ that encompassed the true value of the population parameter would tend toward 95%."

NOT Mean: "For a given realized interval $[a, b]$, there is a 95% probability that the population parameter lies within the interval"

Example,

By CLT, when n is large, $\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$



The $\mathbf{1-\alpha}$ **confidence interval** for μ is a set

$$\left[\overline{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),\ \overline{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

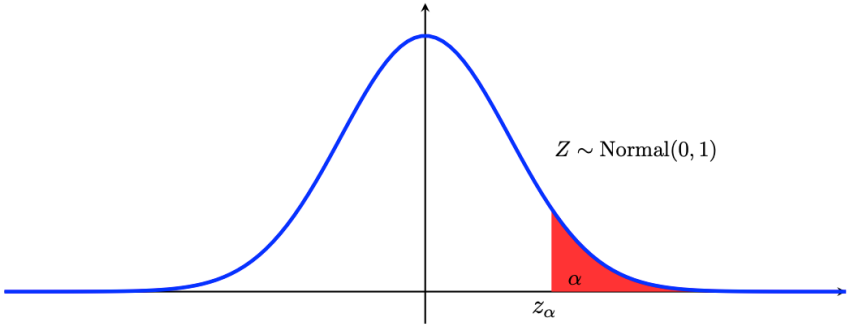
Usually, σ is unknown, then, we need to use sample variance and t-distributions.

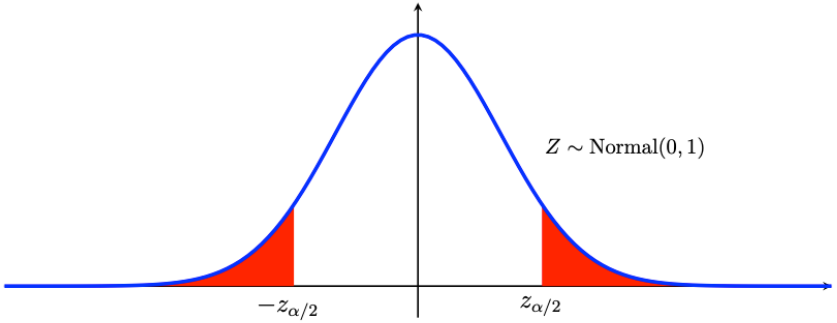
Another method of estimation of confidence interval: **Bootstrapping**.

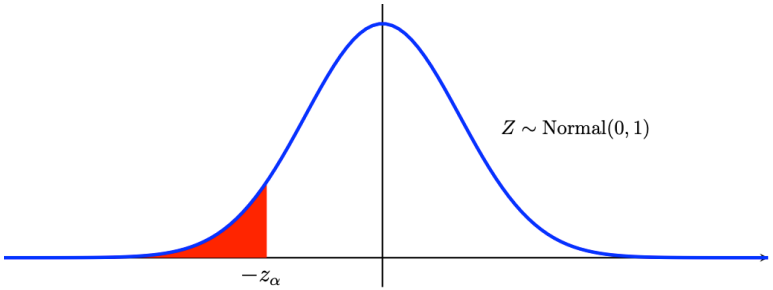
Hypothesis Tests H_0 v.s. H_1

Null hypothesis H_0: no difference or no relationship or no effect. ($\mu = \mu_0$)

Alternative hypothesis H_1: opposite of H_0 ($\mu > \mu_0, or\ \mu \neq \mu_0$ or $\mu < \mu_0$)







Type I and Type II Errors in Hypothesis Testing

When we make the decision, it is possible to have errors.

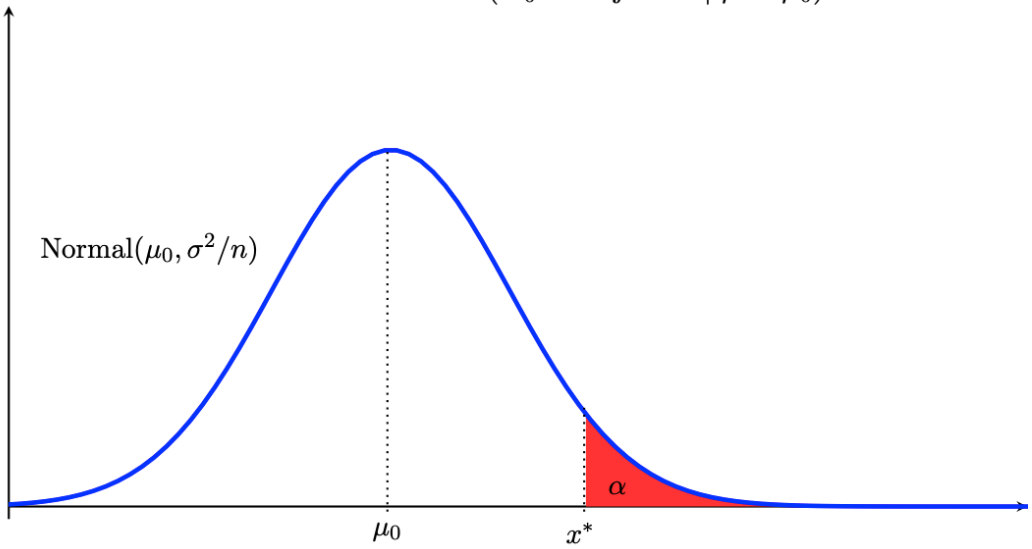
Facts / Decisions	H_0 is True	H_1 is True
Reject H_0	**Type I Error**	Correct Decision
Fail to reject H_0	Correct Decision	**Type II Error**

P(Type I error): $\alpha = P(H_0 \text{ is Rejected} \mid H_0 \text{ is True})$

P(Type II error) $\beta = P(H_0 \text{ is Accepted} \mid H_1 \text{ is True})$

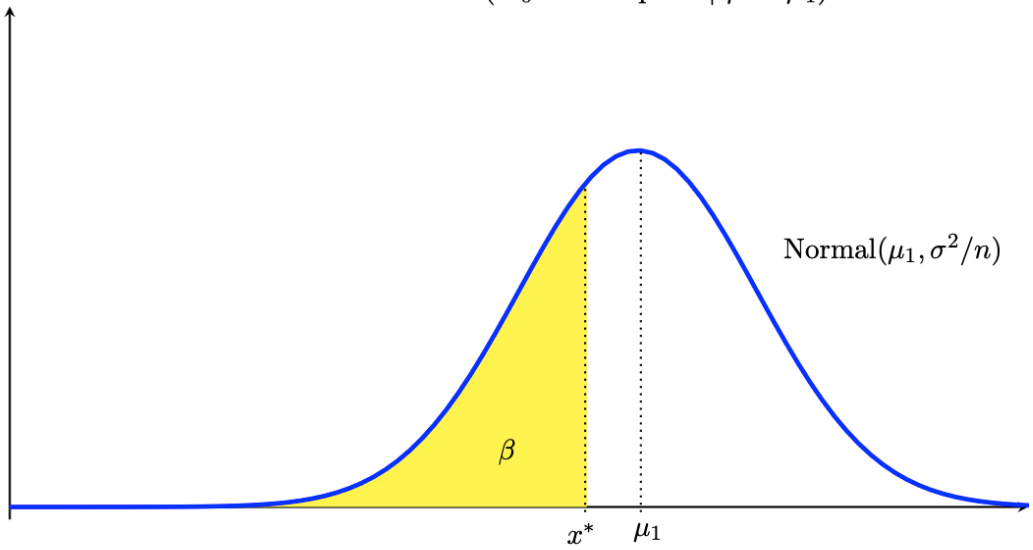
Type I error:

$$\begin{aligned}\alpha &= P(H_0 \text{ is Rejected} \mid H_0 \text{ is True}) \\ &= P(H_0 \text{ is Rejected} \mid \mu = \mu_0)\end{aligned}$$



Type II error:

$$\begin{aligned}\beta &= P(H_0 \text{ is Accepted} \mid H_1 \text{ is True}) \\ &= P(H_0 \text{ is Accepted} \mid \mu = \mu_1)\end{aligned}$$



The Rules of Probability

Sum Rule: $$p(X) = \sum_{Y} p(X, Y)$$

Product Rule: $$p(X, Y) = P(Y|X)P(X)$$

Bayes' rule:

$$\mathrm{p}(Y|X) = \frac{p(X,Y)}{p(X)} = \frac{p(X,Y)}{\sum_{Y'} p(X,Y')} = \frac{p(X|Y)p(Y)}{\sum_{Y'} p(X|Y)p(Y')}$$

By **Bayes Rule:**

$$P(Y = k \,|\vec{X} = \vec{x}) = \frac{P\left(\vec{X} = \vec{x}\middle|Y = k\right)P(Y = k)}{P(\vec{X} = \vec{x})}$$

Here, $P\left(\vec{X} = \vec{x}\right) = \sum_{all\ i} P\left(\vec{X} = \vec{x}\middle|Y = i\right)P(Y = i)$

Bayes Rule using words:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

$$\text{Goal} = \frac{\text{known} \times \text{Guess}}{\text{constant}}$$

Some examples of Bayes' rule application

1. Covid-19 test.
2. Monty Hall Problem.
3. Bayesian inference example

Example: Testing for Covid-19.

$D = 1$ Infected by disease. ($D = 0$ not infected.)

$Y = 1$ **Test** positive. ($\mathrm{Y} = 0$. Test negative) –binary classification for Y

Test **Sensitivity (True-Positive Rate**): $= P(Y = 1|D = 1)$

Test **Specificity (True-Negative Rate)**:$= P(Y = 0|D = 0)$

Prevalence of the disease $= P(D = 1)$

Suppose $P(D = 1) = 0.01$.

$P(Y = 1|D = 1) = 0.875$

$P(Y = 0|D = 0) = 0.975$

Then suppose a person test positive, what is the chance that the person really infected?

$$P(D=1|T=1) = \frac{P(Y=1|D=1)P(D=1)}{P(Y=1|D=1)P(D=1) + P(Y=1|D=0)P(D=0)}$$

$$= \frac{0.875 * 0.01}{0.875 * 0.01 + 0.025 * 0.99}$$

$$= 0.26$$

Similarly, we also know the chance that a person infected given test negative:

$$P(D=1|T=0) = \frac{P(Y=0|D=1)P(D=1)}{P(Y=0|D=1)P(D=1) + P(Y=0|D=0)P(D=0)}$$

$$= 0.0013$$

Example

Table 1: Ratings of 109 CT images by a single radiologist vs. true disease status

	CT rating					
True disease status	Definitely normal (1)	Probably normal (2)	Questionable (3)	Probably abnormal (4)	Definitely abnormal (5)	Total
Normal	33	6	6	11	2	58
Abnormal	3	2	2	11	33	51
Total	36	8	8	22	35	109

Table 2: **Sensitivity** and **specificity** of the radiologist's ratings according to different test-positive criteria based on the data in above table

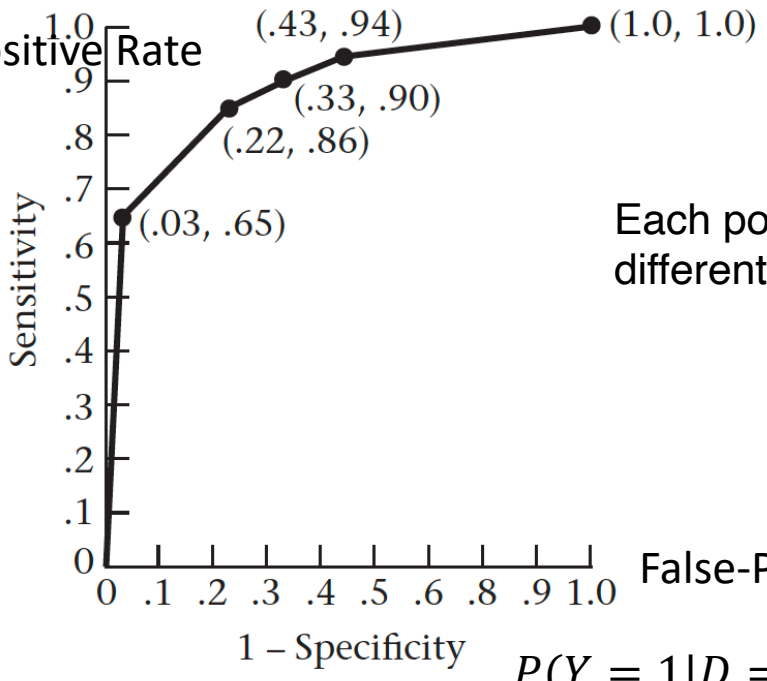
Test-positive criteria	Sensitivity	Specificity
1 +	1.0	0
2 +	.94	.57
3 +	.90	.67
4 +	.86	.78
5 +	.65	.97
6 +	0	1.0

ROC Curves

A **Receiver Operating Characteristic** (ROC) curve is a plot of the **sensitivity** (on the y-axis) versus (**1 – specificity**) (on the x-axis) of a screening test, where the different points on the curve correspond to different cutoff points used to designate test positive.

$P(Y = 1|D = 1)$

True-Positive Rate



Each point represents (1 – specificity, sensitivity) for different test-positive criteria.

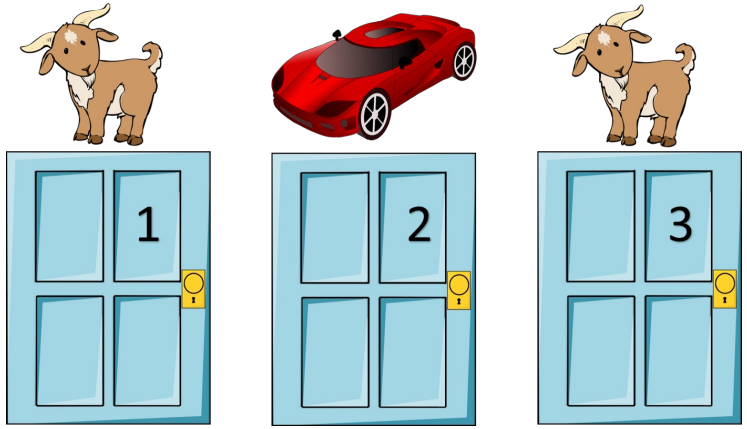
False-Positive Rate

$P(Y = 1|D = 0)$

The area under the ROC curve is a reasonable summary of the overall diagnostic accuracy of the test. (AUC means Area Under Curve.)

Monty Hall Problem (https://en.wikipedia.org/wiki/Monty_Hall_problem)

In a game show, there are three doors with is a big prize behind only one door. You choose one of them, then one of the left is opened and there is no prize behind the opened door. You have a chance to switch your choice. Will you switch?



C: 1st choice is correct. $P(C) = 1/3$
S: win after switch.
T: win without switch.

$$P(S) = P(S|C)P(C) + P(S|C^c)P(C^c) = 0(\frac{1}{3}) + 1(\frac{2}{3}) = \frac{2}{3}$$

$$P(T) = P(T|C)P(C) + P(T|C^c)P(C^c) = 1(\frac{1}{3}) + 0(\frac{2}{3}) = \frac{1}{3}$$

Conclusion: you should switch.

Bayesian Inference

I tell you that I can toss coin such that it always comes up Heads. You are 95% certain that I am lying. I tossed a coin 5 times in front of you and comes up Head every time. How certain are you now that I am lying

H_1: I am lying (the coin if fair).
$H_2 = H_1^c$: I can always toss Heads.
D=Data:= { I tossed 5 times and got 5 heads }
Prior probabilities: $P(H_1) = 0.95$ and $P(H_2) = 0.05$
Posterior probabilities: (after experiments) $p_1 = P(H_1|D)$ updated probability for H_1 given data D. By Bayes's Theorem,

$$P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)} == \frac{(0.5)^5 0.95}{(0.5)^5 0.95 + 1(0.05)} = 0.37$$

Summary: based on the data D, your new degree of certainty that I am lying is 37%.
We used independence to calculate $P(D|H_1) = P(HHHHH) = P(H)^5 = 0.5^5$.

Further examples:
1. Naive Bayes spam filtering https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering
2. Bayesian poisoning https://en.wikipedia.org/wiki/Bayesian_poisoning

Bayesian Method Example-Binomial distribution

Suppose there is a coin that may be biased with unknown probability θ of giving a "heads."

Suppose we flip the coin n times and observe x "heads."

The probability of this observation given the value of θ, comes from binomial distribution:

$$P(x|n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Classical (Frequentist) method:

The frequentist approach is to construct an estimator for θ, which in theory can be any function of the observed data $\hat{\theta}(x,n)$ and show that $\hat{\theta} \to \theta$ as $n \to \infty$.

The classical estimator in this case is the empirical frequency (use MLE)

$$\hat{\theta} = \frac{x}{n}$$

❑ Bayesian approach

Frequentist approach ignores all prior information.

Bayesian approach choose a **prior distribution** $p(\theta)$. A convenient prior in this case is the Beta distribution:

$$p(\theta \mid \alpha, \beta) = \mathcal{B}(\theta;\ \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\text{with } \Gamma(z) = \int_0^{\infty} t^{z-1}e^{-t}dt$$

$$Or\ write\ \mathcal{B}(\theta;\ \alpha, \beta) = \frac{1}{B(\alpha, \beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\text{with normalizing constant } B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\,\mathrm{d}\theta$$

A **prior probability distribution** $p(\theta)$ to θ, representing your **degree** of belief with respect to θ.

Given our observations $D = (x, n)$, we can now compute the **posterior distribution** of θ by Bayes Theorem:

$$p(\theta \mid x, n, \alpha, \beta) = \frac{P(x|n, \theta)p(\theta|\alpha, \beta)}{P(x|n, \alpha, \beta)}$$

$$= \frac{P(x|n, \theta)p(\theta|\alpha, \beta)}{\int_0^1 P(x|n, \theta)p(\theta|\alpha, \beta)\, d\theta}$$

$$= \ldots$$

$$= \frac{1}{B(\alpha + x, \beta + n - x)} \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}$$

$$= \mathcal{B}(\theta;\ \alpha + x, \beta + n - x)$$

If we assign a different prior distribution, then we arrive at a different posterior distribution. (α, β are hyperparameters.)

Beta distribution Calculator: https://homepage.divms.uiowa.edu/~mbognar/applets/beta.html

Summary of Bayesian approach

$P(\mathcal{D} \mid \vec{\theta})$: the **likelihood** for the data. The parameters of interest $\vec{\theta}$ (unknown)

$P(\vec{\theta})$: density associated with the **prior distribution**. (The **degree of belief** of the distribution before the data.)

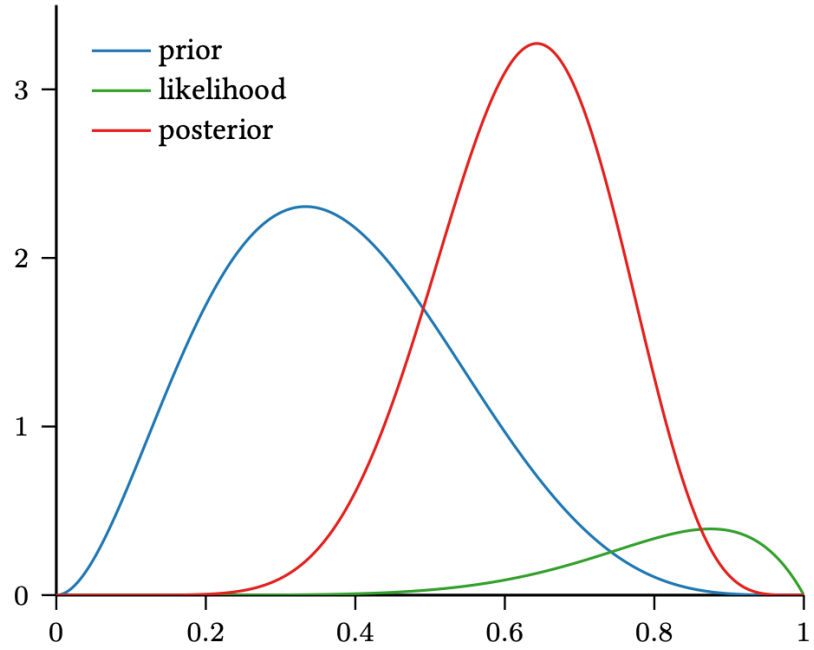
Bayes' theorem converts a prior probability into a **posterior probability**

$$P(\vec{\theta}|\mathcal{D}) = \frac{P(\mathcal{D} \mid \vec{\theta})\, P(\vec{\theta})}{P(\mathcal{D})} = \frac{P(\mathcal{D} \mid \vec{\theta})\, P(\vec{\theta})}{\int P(\mathcal{D} \mid \vec{\theta}')\, P(\vec{\theta}')\, d\vec{\theta}'}$$

Posterior distribution $P(\vec{\theta}|\mathcal{D})$ is the updated degree of belief with respect to θ, based on the data $\mathcal{D}$. The **new degree of belief** is called the posterior probability distribution of θ.

The rather convenient fact that the posterior remains a beta distribution is because the beta distribution satisfies a property known as **conjugacy** with the binomial likelihood.

α, β serve as "pseudocounts," or fake observations we pretend to have seen before seeing the data. They are hyperparameters we need to determine.



$$(\alpha, \beta) = (3, 5)$$

$$(x, n) = (5, 6).$$

We also get **another** "test and confidence interval" (**credible interval**), for example.

$$P\left(\theta < \frac{1}{2} \mid x, n, \alpha, \beta\right) = \int_0^{1/2} p(\theta | x, n, \alpha, \beta) d\theta$$

Once we derive the posterior distribution $p(\theta \mid \mathcal{D})$ using Bayesian approach, we can study the posterior mean and posterior variance.

Posterior mean

$$E[\theta|\mathcal{D}]$$

Posterior variance

$$Var[\theta|\mathcal{D}]$$

For example, the posterior distribution of the probability θ of giving head in coin is

$$E[\theta|\mathcal{D}] = \frac{\alpha + x}{\alpha + \beta + n}$$

$$Var[\theta|\mathcal{D}] = \frac{(\alpha + x)(\beta + n - x)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}$$

In addition, the Mode

$$\text{Mode [highest point in density curve]} = \frac{\alpha+x-1}{\alpha+\beta+n-2}$$

Prior v.s. Posterior

Relation between the prior and posterior mean and variance is give by the probability theorems:

law of iterated expectations:

$$E(\theta) = E_{\mathcal{D}}\left[E_{\theta|\mathcal{D}}[\theta\ |\mathcal{D}]\right]$$

Prior mean of θ = Average posterior mean of θ over data distribution.

law of total variance:

$$Var(\theta) = E[Var(\theta|\mathcal{D})] + Var(E(\theta|\mathcal{D}))$$

Posterior variance of θ is, on average, less than prior variance of θ

➢ Credible Interval

Recall that Bayesian approach derive the posterior distribution:

$$p(\theta \mid \mathcal{D})$$

A interval $[a, b]$ is called 95% **credible interval for $\boldsymbol{\theta}$,** if the posterior probability

$$P(\theta \in [a, b] \mid \mathcal{D}) = \int_a^b p(\theta|\mathcal{D})\, d\theta = 95\%$$

Bayesian Hypothesis testing

The hypothesis testing is similarly. Suppose we have a null hypothesis H_0

$$P(\theta \in H_0 \mid \mathcal{D}) = \int_{H_0} p(\theta|\mathcal{D})\, d\theta$$

A hypothesis is simply a subset of the parameter space $H_0 \subseteq \Theta$ in the Bayesian decision theory.

Example: Coin Flip continue

From our coin flip example, we know that

$$p(\theta \mid x, n, \alpha, \beta) = \mathcal{B}(\theta;\ \alpha + x, \beta + n - x)$$

Assume $p(\theta \mid \alpha, \beta) = \mathcal{B}(\theta; 1, 1)$

Suppose we flip the coin independently $n = 50$ times and observe $x = 30$ heads. After gathering this data, we wish to consider whether the coin is fair H_0: $\theta = \frac{1}{2}$?

$$P(\theta \in H_0 | \mathcal{D}) = p(\theta = \frac{1}{2} \mid x, n, \alpha, \beta) = \mathcal{B}(\theta; 31, 21)$$

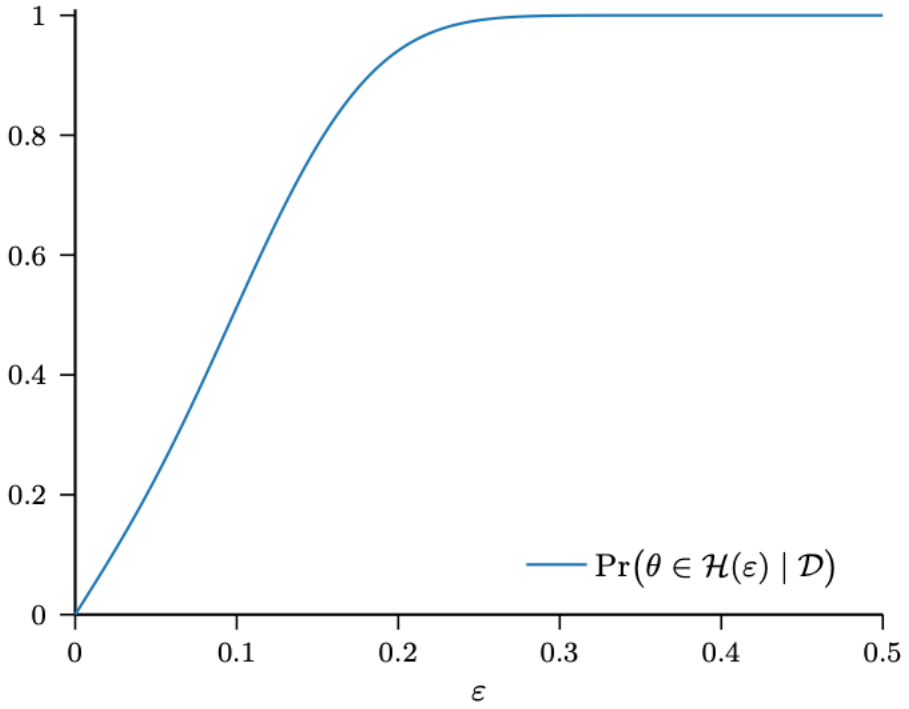
This is **zero** since it is continuous distribution.

Consider a parameterized family of hypotheses of the for

$$H_\epsilon := \left(\frac{1}{2} - \epsilon\ , \frac{1}{2} + \epsilon\right)$$

The posterior probability of the hypotheses H_ϵ for $1 < \epsilon < 1/2$

$$P(\theta \in H_\epsilon | \mathcal{D}) = \int_{\frac{1}{2}-\epsilon}^{\frac{1}{2}+\epsilon} p(\theta|\mathcal{D})\, d\theta$$



Python/MATLAB/R Code calculating the posterior of the binomial likelihood :

```
success_prob = 0.3
data =random.binomial(n=1, p=success_prob, size=1000) # sucess is 1, failure is 0.
# Domain θ
theta_range = linspace(0, 1, 1000)

# Prior P(θ)
a = 2
b = 8
theta_range_e = theta_range + 0.0001
prior = beta.cdf(x = theta_range_e, a=a, b=b) - beta.cdf(x = theta_range, a=a, b=b)

# The sampling dist. aka Likelihood P(X|θ)
likelihood = binom.pmf(k = np.sum(data), n = len(data), p = theta_range)

# Posterior
posterior = likelihood * prior
normalized_posterior = posterior / sum(posterior)
```

Conjugate prior

$$P(\vec{\theta}|\mathcal{D}) = \frac{P(\mathcal{D} \mid \vec{\theta})\, P(\vec{\theta})}{P(\mathcal{D})}$$

For some likelihood functions $P(\mathcal{D} \mid \vec{\theta})$, if you choose a certain prior $P(\vec{\theta})$, the posterior $P(\vec{\theta}|\mathcal{D})$ ends up being in the same distribution as the prior. Such a prior then is called a **Conjugate Prior** of the likelihood function.

For example, the Beta distribution $\mathcal{B}(\theta;\ \alpha, \beta)$ is the conjugate prior of the binomial distribution $Binomial(x; n, \theta)$.

The conjugate prior of the normal distribution $N(\mu, \sigma^2)$ with fixed σ is also normal distribution, but with different parameters.

If the likelihood function belongs to the exponential family, then a conjugate prior exists, often also in the exponential family.

A Table of conjugate distributions can be found in Wikipedia:

https://en.wikipedia.org/wiki/Conjugate_prior

How does the Conjugate Prior help?

When you know that your prior is a conjugate prior, you can skip the computation.

$$\text{posterior} \propto \text{likelihood} * \text{prior}$$

During the modeling phase, we already know the posterior will also be a beta distribution.

Therefore, after carrying out more experiments, **you can compute the posterior simply by adding the number of acceptances and rejections to the existing parameters α, β respectively**, instead of multiplying the likelihood with the prior distribution**.**

Maximum A Posteriori estimation (MAP) v.s. MLE

Suppose $\vec{\theta}$ are the model parameters, and $\mathcal{D} = \left\{(\vec{x}^{(i)}, \vec{y}^{(i)})\right\}_{i=1}^{N}$ the observed data.

$P(\mathcal{D} \mid \vec{\theta})$: the **likelihood** for the data.

The **Maximum Likelihood Estimate (MLE)** of is

$$\hat{\vec{\theta}}_{MLE} \coloneqq \underset{\vec{\theta}}{\operatorname{argmax}}\, P(\mathcal{D} \mid \vec{\theta}) = \underset{\vec{\theta}}{\operatorname{argmax}}\, \log P(\mathcal{D} \mid \vec{\theta})$$

Most of the model we have are using MLE, e.g., logistics regression, linear regression, generalized linear regression,

From Bayesian statistics, based on **prior** $p(\vec{\theta})$, we have calculated the **posterior** distribution:

$$P(\vec{\theta}|\mathcal{D}) = \frac{P(\mathcal{D} \mid \vec{\theta})\, P(\vec{\theta})}{P(\mathcal{D})}$$

The Maximum A Posteriori estimation (MAP) is

$$\hat{\vec{\theta}}_{MAP} := \underset{\vec{\theta}}{\operatorname{argmax}}\, P(\vec{\theta}|\mathcal{D}) = \underset{\vec{\theta}}{\operatorname{argmax}}\, P(\mathcal{D} \mid \vec{\theta})\, P(\vec{\theta})$$

$$= \underset{\vec{\theta}}{\operatorname{argmax}}\, \log P(\mathcal{D} \mid \vec{\theta}) + \log P(\vec{\theta})$$

Example (MAP for μ in normal distribution.)

Suppose we have iid data $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$ observed from normal distribution $N(\mu, \sigma^2)$ with known σ.

$$p(x^{(i)}|\mu) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}(x^{(i)} - \mu)^2\right)$$

The **MLE** for μ is

$$\hat{\mu}_{MLE} = \underset{\mu}{\operatorname{argmax}} \log P(\mathcal{D}\,|\mu) = \underset{\mu}{\operatorname{argmax}} \log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}(x^{(i)} - \mu)^2\right) =$$

$$\dots = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$$

Now find the **MAP** estimate of μ.

The conjugate prior of normal distribution is normal, there is a closed-form solution analytically.

$$\hat{\mu}_{MAP} = \underset{\mu}{\operatorname{argmax}} \log P(\mu|\mathcal{D}) = \underset{\mu}{\operatorname{argmax}} \log P(\mathcal{D}\,|\mu)\,P(\mu)$$

Notice that

$$P(\mu)P(\mathcal{D}\,|\mu) = \frac{1}{\sqrt{2\pi}\,\sigma_0}\exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right)\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2\sigma^2}\left(x^{(i)}-\mu\right)^2\right)$$

Hence,

$$\begin{aligned}\hat{\mu}_{MAP} &= \underset{\mu}{\operatorname{argmax}} \log P(\mathcal{D}\,|\mu)\,P(\mu) \\ &= \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{N}\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 + \frac{1}{2\sigma^2}\left(x^{(i)}-\mu\right)^2 \\ &= \frac{\sigma_0^2\left(\sum_{i=1}^{N}x^{(i)}\right)+\sigma^2\mu_0}{\sigma_0^2 N+\sigma^2}\end{aligned}$$

The MAP estimate $\hat{\mu}_{MAP}$ is a linear combination between the prior mean μ_0 and the sample mean $\bar{x}$ weighted by their respective covariances.

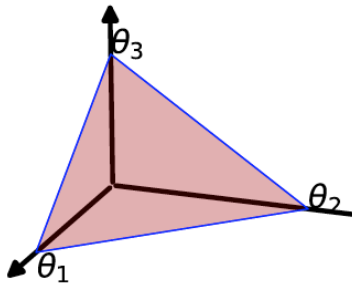
$\hat{\mu}_{MAP} \to \hat{\mu}_{MLE}$ when $\sigma_0 \to \infty$

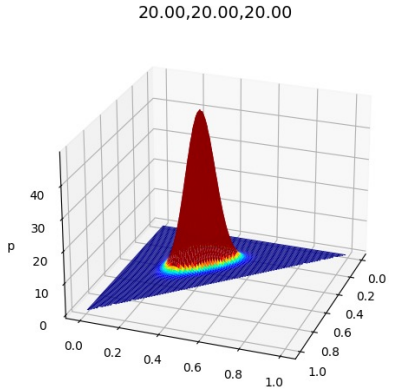
Extra Example: **The Dirichlet-multinomial model**

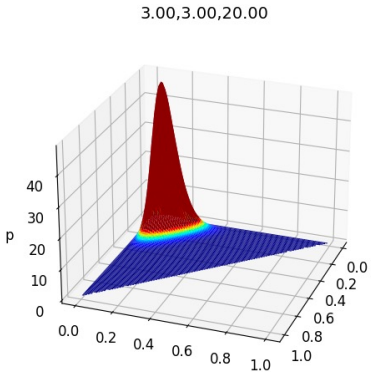
Likelihood: Categorical distribution.

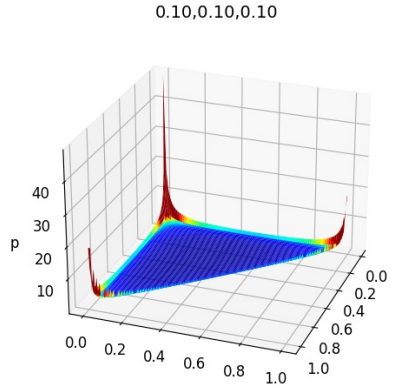
Prior: Dirichlet distribution

Posterior: *Dirichlet* distribution









See Murphy 1: Sec 4.6.3

➢ Bayesian decision theory

Design classifiers to recommend **decisions** that **minimize** some total expected "**risk**".

– The simplest risk is the classification error (i.e., costs are equal).

– Typically, the risk includes the cost associated with different decisions.

❑ General Decision Problem

1. **Parameter space** (also called a **state space**) Θ, with an unknown value $\theta \in \Theta$. These are the underlying state of nature.

2. **Sample space** $\mathcal{X}$: potential observations $\mathcal{D} \in \mathcal{X}$ we could theoretically make.

3. **Action space:** $\mathcal{A}$: representing the potential actions we/decision maker/agent may select from.

Finally, we will have a **likelihood** function $p(\mathcal{D} \mid \theta)$ linking potential observations to the parameter space.

After conducting an experiment and observing data $\mathcal{D}$, we are compelled to select an action $a \in \mathcal{A}$.

Decision rule is a function $\delta: \mathcal{X} \rightarrow \mathcal{A}$ that selects an action a given the observations $\mathcal{D}$

In general, this decision rule can be any arbitrary function. (In Machine Learning, we will usually put restrictions on $\mathcal{A}$ to ensure we have enough data to learn them.)

- Use loss/cost function to select which decision rule to use:

loss function $L\colon \Theta \times \mathcal{A} \to \mathbb{R}$

Loss function $L(\theta, a)$ summarizes "how bad" an action a was if the true value of the parameter was revealed to be θ. For example,

$$L(\theta, a) = \begin{cases} 0 & \text{If } a(x) = \theta \text{ (correct decision)} \\ 1 & \text{If } a(x) \neq \theta \text{ (incorrect decision)} \end{cases}$$

Ideally, we would select the action a that **minimizes** this loss $L(\theta, a)$, but we don't know the exact value of θ.

Frequentist use likelihood $p(\mathcal{D} \mid \theta)$ to establish decision theory. (Only from experiments)

Bayesian use the posterior $p(\theta|\mathcal{D})$ establish decision theory, which also includes prior belief $p(\theta)$.

❑ Bayesian decision

Bayesian decision theory use the posterior $p(\theta|\mathcal{D})$, which is the current belief about the unknown parameter θ, given the observed data $\mathcal{D}$.

Given a potential action a, define the **posterior expected loss/risk** of a by averaging the loss function over the unknown parameter θ :

$$l(p(\theta|\mathcal{D}), a) = E_\theta[L(\theta, a)|\mathcal{D}] = \int_\Theta L(\theta, a)\, p(\theta|\mathcal{D}) \mathrm{d}\theta$$

When there is an action minimizing the posterior expected loss, choose a **Bayes action/estimator**:

$$\delta^*(\mathcal{D}) = \underset{a \in \mathcal{A}}{\operatorname{argmin}}\, l(p(\theta|\mathcal{D}), a)$$

$$= \underset{a \in \mathcal{A}}{\operatorname{argmin}} \int_\Theta L(\theta, a)\, p(\mathcal{D} \mid \theta) p(\theta) \mathrm{d}\theta$$

❑ **Frequentist decision theory** (no prior distribution)

Use likelihood $p(\mathcal{D} \mid \theta)$ on a given parameter θ.

The **frequentist risk(error)** **(Expected loss)** of a decision function δ is defined by the expected loss incurred when repeatedly using the decision rule δ on different datasets $\mathcal{D}$ as a function of the unknown parameter θ:

$$R(\theta, \delta) = E_{\mathcal{D}|\theta}\left[L\left(\theta, \delta(\mathcal{D})\right)\middle|\theta\right] = \int_{\mathcal{X}} L(\theta, \delta(\mathcal{D}))p(\mathcal{D}|\,\theta)\; d\mathcal{D}$$

For two decision rules $\delta_1 and\ \delta_2$, we say δ_1 $\boldsymbol{dominates}$ δ_2 if

$R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all $\theta \in \Theta$ and

$R(\theta, \delta_1) < R(\theta, \delta_2)$ for at least one θ

If there is a decision rule δ that is not dominated by any other rule, it is called **admissible**.

❑ Bayes risk(error)

Bayes risk of a decision δ

$$r(p(\theta), \delta) = E_\theta\big(R(\theta, \delta)\big) = \int_\Theta \int_\mathcal{X} L(\theta, \delta(\mathcal{D}))p(\mathcal{D}|\,\theta)\, p(\theta) d\mathcal{D} d\theta$$

Any decision δ *minimizing Bayes risk* is called a **Bayes rule/decision**.

- Every Bayes rule is admissible
- Every admissible decision rule is a generalized Bayes rule for some (possibly improper) prior $p(\theta)$.

$$\underset{a}{\operatorname{argmin}}\, r(p(\theta), a) = \underset{a}{\operatorname{argmin}} \int_\Theta L(\theta, a) p(\theta|\mathcal{D}) d\theta = \delta^*(\mathcal{D})$$ **Bayes action.**

Here, $a = \delta(\mathcal{D})$

➢ Point Estimation Example

Problem: Estimate the value of a parameter θ. (e.g., unknown bias of a coin.)

Review Bayesian method:

1.) Select a **likelihood** function $p(\mathcal{D} \mid \theta)$, explaining how observed data $\mathcal{D}$ are expected to be generated given the value of θ.

2.) Select a **prior** distribution $p(\theta)$ reflecting our initial beliefs about θ.

3.) Conduct an experiment to gather data and use Bayes' theorem to derive the **posterior** $p(\theta \mid \mathcal{D})$.

Bayesian inference need to use $p(\theta \mid \mathcal{D})$ to answer some questions. For example, we might be compelled to choose a single value $\hat{\theta}$ to serve as a **point estimate** of θ. To a Bayesian, the selection of $\hat{\theta}$ is a **decision**, and in different contexts we might want to select different values to report. (Find $\hat{\theta}$ which has the smallest loss.)

In the example, the action is choosing a parameter, so $\mathcal{A} = \Theta$.

The decision rule $\delta: \mathcal{X} \to \mathcal{A}$ is written as $\hat{\theta}(\mathcal{D})$.

1. Selecting a **loss function**, e.g.,

$$L(\theta, \hat{\theta}) := (\theta - \hat{\theta})^2$$

2. Find the **posterior expected loss** at every point.

$$l(\hat{\theta}) := E[L(\theta, \hat{\theta})|\mathcal{D}] = \int_\Theta L(\theta, \hat{\theta})\, p(\theta|\mathcal{D}) \mathrm{d}\theta = \int_\Theta (\theta - \hat{\theta})^2\, p(\theta|\mathcal{D}) \mathrm{d}\theta$$

$$= \int_\Theta \theta^2\, p(\theta|\mathcal{D}) \mathrm{d}\theta - 2\hat{\theta} \int_\Theta \theta\, p(\theta|\mathcal{D}) \mathrm{d}\theta + \hat{\theta}^2 \int_\Theta p(\theta|\mathcal{D})\, \mathrm{d}\theta$$

$$= \int_\Theta \theta^2\, p(\theta|\mathcal{D}) \mathrm{d}\theta - 2\hat{\theta} \mathrm{E}[\theta|\mathcal{D}] + \hat{\theta}^2$$

3. A **decision rule** $\hat{\theta}$ that **minimizes** posterior expected loss for every possible set of observations $\mathcal{D}$ is called a **Bayes estimator**.

By calculus $\dfrac{\partial l(\hat{\theta})}{\partial \hat{\theta}} = -2E[\theta|\mathcal{D}] + 2\hat{\theta} = 0$

$\hat{\theta} = E[\theta|\mathcal{D}] = \operatorname{argmin} l(\hat{\theta})$, since the second derivative positive.

Bayes estimator in the case of squared loss is the **posterior mean**:

$$\hat{\theta}(\mathcal{D}) = E[\theta|\mathcal{D}]$$

Recall that $\theta|\mathcal{D} \sim \mathcal{B}(\theta;\ \alpha + x, \beta + n - x)$

$E[\theta|\mathcal{D}] = \dfrac{\alpha + x}{\beta + n - x}$ Median$[\theta|\mathcal{D}] \approx \dfrac{\alpha+x-1/3}{\beta+n-x-2/3}$ $\operatorname{mode}[\theta|\mathcal{D}] = \dfrac{\alpha + x - 1}{\beta + n - x - 2}$

Similarly,

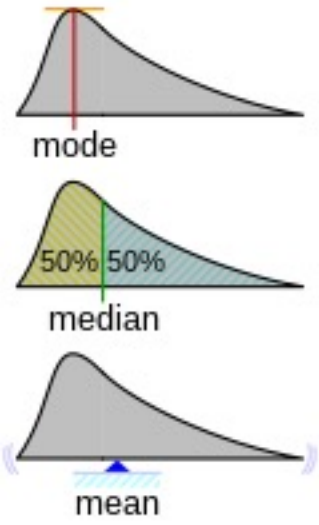
- The **Bayes estimator** for the absolute deviation loss $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ is the **posterior median**.

- The **Bayes estimators** for a relaxed $0-1$ loss:

$$L(\theta, \hat{\theta}) = \begin{cases} 0 & |\theta - \hat{\theta}| < \epsilon \\ 1 & |\theta - \hat{\theta}| \geq \epsilon \end{cases}$$

converge to the posterior **mode** for small ε.

mode: value that appears most often.



More over, we know $\hat{\theta}_{MAP} = \text{mode}[\theta|\mathcal{D}]$

$$\hat{\theta}_{MLE} = \frac{x}{n}$$

Example: Classification with 0–1 loss

Suppose our observations are of the form (x, y), where x is an arbitrary input, and $y \in \{0, 1\}$ is a binary label associated with x.

Goal: predict the label y' associated with a new input x'.

The Bayesian approach: derive a model giving probability (not conditioned on θ)

$$P(y' = 1 | x', \mathcal{D})$$

Use joint distribution with θ, then take marginal:

$$P(y' = 1 | x', \mathcal{D}) = \int P(y' = 1, \theta | x', \mathcal{D})\, d\theta = \int P(y' = 1 | x', \mathcal{D}, \theta)\, P(\theta | \mathcal{D}) d\theta$$

The prediction of a label is actually a decision.

Action space is $\mathcal{A} = \{0, 1\}$, enumerating the two labels we can predict.
Parameter space is the same: the only uncertainty we have is the unknown label y'.

Suppose the $0-1$ loss function for this problem:

$$L(y', a) = \begin{cases} 0 & a = y' \\ 1 & a \neq y' \end{cases}$$

We pay a constant loss for every mistake we make. In this case, the **expected loss** of each possible action is simple to compute:

$$E(L(y', a = 1)|x', \mathcal{D}) = P(y' = 0|x', \mathcal{D})$$
$$E(L(y', a = 0)|x', \mathcal{D}) = P(y' = 1|x', \mathcal{D})$$

The **Bayes action (classifier)** is then to predict the class with the highest probability.

$$\hat{f}(\vec{x}) = \begin{cases} 1 & \text{if } P(y' = 1|x', \mathcal{D}) > 0.5 \\ 0 & \text{if } P(y' = 1|x', \mathcal{D}) < 0.5 \end{cases}$$

Notice that if we change the loss to have different costs of mistakes (so that $L(0,1) \neq L(1,0)$), then the Bayes action might compel us to select the less-likely class to avoid a potentially high loss for misclassification.

➢ Bayesian Perspective of Modeling in Machine Learning

Suppose $\vec{\theta}$ are the model parameters, and $\mathcal{D} = \left\{(\vec{x}^{(i)}, \vec{y}^{(i)})\right\}_{i=1}^{N}$ the observed data.

Assume $\vec{\theta}$ are also drawn from a distribution. It is convenient to characterize the weighting schemes $\vec{\theta}$ in terms of distributions and their associated densities.

Even though there is some 'true' value of " $\vec{\theta}$ (i.e. the values that drive the data generating mechanism), they are treated as random variables.

$P(\vec{\theta})$: density associated with the prior distribution. (The best guess of the distribution before the data.)

$P(\mathcal{D} \mid \vec{\theta})$: the likelihood for the data.

Bayes' theorem converts a prior probability into a **posterior probability**

$$P(\vec{\theta}|\mathcal{D}) = \frac{P(\mathcal{D} \mid \vec{\theta})\, P(\vec{\theta})}{P(\mathcal{D})} = \frac{P(\mathcal{D} \mid \vec{\theta})\, P(\vec{\theta})}{\int P(\mathcal{D} \mid \vec{\theta}')\, P(\vec{\theta}')\, d\vec{\theta}'}$$

In words:

posterior $\propto$ **likelihood** $\times$ **prior**

The denominator $P(\mathcal{D})$ is the normalization constant.

Notation $\propto$ means "proportional to", since we we are ignoring the constant.

- **Prediction**

  Once we have computed the posterior over the parameters, we can make prediction of Y, y given a **new** input $\vec{x}$ by computing the **posterior predictive distribution**

$$p(y|\vec{x}, \mathcal{D}) = \int p(y|\vec{x}, \vec{\theta}) p(\vec{\theta}|\mathcal{D})\, d\vec{\theta}$$

  Here, we assume given **fixed** location $\vec{x}$

The posterior predictive distribution is essentially a weighted average
Likelihood: weight potential $P(\mathcal{D} \mid \vec{\theta})$ according to the posterior distribution $P(\vec{\theta}|\mathcal{D})$.

Automatically accounts for uncertainty in the estimation of $\vec{\theta}$

$p(y|\vec{x},\mathcal{D})$ is the density for a distribution and, hence, also automatically provides a measure of variability for the prediction itself.

The Bayesian approach adopts the attitude that, while the true $\vec{\theta}$ is unknown, one can does have access to the posterior, $p(\vec{\theta}|\mathcal{D})$, which provides relative weight to potential values of $\vec{\theta}$

If we knew $\vec{\theta}$, we could base our prediction on the likelihood $p(\vec{\theta}|\mathcal{D})$

- this is the frequentist approach
- i.e., replace $\vec{\theta}$ with an estimate $\hat{\vec{\theta}}$
- uncertainty in $\hat{y}$ typically doesn't account for uncertainty in $\hat{\vec{\theta}}$

➢ The Bayesian Framework for general linear model

For a given parameter $\vec{\theta}$, suppose the distribution of a random dataset $\mathcal{D} = \{(\vec{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ is the following model:

$$y^{(i)} = f(\vec{x}^{(i)}) + \epsilon_i = \sum_{i=1}^{p} \theta_j h_j(\vec{x}^{(i)}) + \epsilon_i = \vec{h}^T(\vec{x}^{(i)})\vec{\theta} + \epsilon_i$$

where ϵ_i are iid $N(0, \sigma^2)$ random variables, and

$$\vec{h}^T(\vec{x}) = [h_1(\vec{x}) \quad h_2(\vec{x}) \quad \ldots \quad h_p(\vec{x})]$$

with the right side consisting of the spline basis elements. (When $h_i(\vec{x}) = x_i$, we have the linear regression model.)

Thus, **given** $\vec{\theta}$, and the **fixed** location $\vec{x}$, the probability distribution $y^{(i)}$ is

$$(y^{(i)}|\vec{\theta}, \vec{x}^{(i)}) \sim N(\vec{h}^T(\vec{x}^{(i)})\vec{\theta}, \sigma^2)$$

so that $$p\left(y^{(i)}\middle|\vec{\theta},\vec{x}^{(i)}\right)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)}-\vec{h}^T\left(\vec{x}^{(i)}\right)\vec{\theta}\right)^2\right)$$

Note that conditioning on $\vec{x}^{(i)}$ at the end means only that we are treating the $\vec{x}^{(i)}$ as fixed in the calculation.

MLE for linear regression:

We already know that the Ordinary Least Squares(OLS) method

We maximize the **likelihood** $P(\mathcal{D}|\vec{\theta})$, or equivalently

$$\hat{\vec{\theta}}_{MLE} \coloneqq \underset{\vec{\theta}}{\operatorname{argmax}}\, P(\mathcal{D}\mid\vec{\theta}) = \underset{\vec{\theta}}{\operatorname{argmin}} -\log P(\mathcal{D}\mid\vec{\theta})$$

Equivalently, $$= \underset{\vec{\theta}}{\operatorname{argmin}}\ RSS(\vec{\theta})$$

$$= (X^TX)^{-1}X^T\vec{y}$$

Bayesian method setup:

The logic is essentially that we are assuming a model for the unknown parameter $\vec{\theta}$ as having a probability distribution.

Before we see any data in the dataset $\mathcal{D} = \left\{(\vec{x}^{(i)}, y^{(i)})\right\}_{i=1}^{N}$ only our prior knowledge can give us an idea of this **distribution** $p(\vec{\theta})$ **for** $\vec{\boldsymbol{\theta}}$, which is therefore called the **prior distribution**.

In this case we give a relatively naïve prior where we do not assume too much by assuming that $\vec{\theta} \sim N(0, \Sigma)$ with **prior** covariance matrix Σ, that is

$$p(\vec{\theta}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{\theta})^T \Sigma^{-1}(\vec{\theta})\right)$$

Remark: We can also assume $\vec{\theta} \sim N(\vec{\mu}, \Sigma)$. Calculation is similarly.

❑ Prior distributions of $f(\vec{x})$ and y

Suppose there are two **fixed** locations $\vec{x}$ and $\vec{x}'$.

$$f(\vec{x}) = \sum_{i=1}^{p} \theta_j h_j(\vec{x}\,) = \vec{h}^T(\vec{x}\,)\vec{\theta} \qquad \text{and} \qquad f(\vec{x}') = \sum_{i=1}^{p} \theta_j h_j(\vec{x}'\,) = \vec{h}^T(\vec{x}'\,)\vec{\theta}$$

They are linear combinations of independent normal **random variables** θ_i with (non-random) coefficients $\vec{h}^T(\vec{x}\,)$.

So, the mean $E_{\vec{\theta}}\big(f(\vec{x})\big) = E\big(\vec{h}^T(\vec{x}\,)\vec{\theta}\big) = \vec{h}^T(\vec{x}\,)E\big(\vec{\theta}\big) = 0$

Similarly, $E\big(f(\vec{x}')\big) = 0$

So, (prior) covariance $Cov\big(f(\vec{x}), f(\vec{x}')\big) = E\big(f(\vec{x})f(\vec{x}')\big) = E[\vec{h}^T(\vec{x}\,)\vec{\theta}\vec{\theta}^T h(\vec{x}')]$

$$= \vec{h}^T(\vec{x}\,)E[\vec{\theta}\vec{\theta}^T]\, h(\vec{x}') = \vec{h}^T(\vec{x}\,)\Sigma\, h(\vec{x}')$$

A Useful Lemma: If $\vec{z} = A\vec{x}$, then $\mathrm{Cov}(\vec{z}) = A\ \mathrm{Cov}(\vec{x})\ A^T$

Recall that $y^{(i)} = f\left(\vec{x}^{(i)}\right) + \epsilon_i = \sum_{i=1}^{p} \theta_j h_j\left(\vec{x}^{(i)}\right) + \epsilon_i = \vec{h}^T\left(\vec{x}^{(i)}\right)\vec{\theta} + \epsilon_i$

It follows that

$$\vec{y} = H\vec{\theta} + \vec{\epsilon}$$

where

$$H = \begin{bmatrix} \vec{h}^T\left(\vec{x}^{(1)}\right) \\ \vec{h}^T\left(\vec{x}^{(2)}\right) \\ \vdots \\ \vec{h}^T\left(\vec{x}^{(N)}\right) \end{bmatrix} \qquad \vec{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

So, $E(\vec{y}) = 0$ and $Cov(\vec{y}) = H\Sigma H^T + \sigma^2 I$

Since a sum of independent Gaussians is Gaussian) we know $\vec{y}$ is Gaussian

$$\vec{y} \sim N(0, H\Sigma H^T + \sigma^2 I)$$

Here, $\vec{x}$ is fixed and $\vec{\theta}$ is the random variable.

Our convention $p(x)$ is density function of random variable x; $p(x|y)$ is conditional density function of random variable x given value of random variable y.

Note that we always view $\{\vec{x}^{(i)}\}_{i=1}^{N}$ as fixed, and for our **prior** distributions we have

$$\vec{\theta} \sim N(0, \Sigma)$$

with $$\vec{y}|\vec{x}^{(i)} \sim N(0, H\Sigma H^T + \sigma^2 I)$$

Given $\vec{\theta}$, the probability distribution $y^{(i)}$ is

$$\left(y^{(i)}\middle|\vec{\theta}, \vec{x}^{(i)}\right) \sim N\left(\vec{h}^T\left(\vec{x}^{(i)}\right)\vec{\theta}, \sigma^2\right)$$

❑ Posterior distribution

The **posterior distribution** for $\vec{\theta}$ (i.e., our distribution for $\vec{\theta}$ given the new information in $\mathcal{D}$) is

$$p(\vec{\theta}|\mathcal{D}) = \frac{p(\mathcal{D} \mid \vec{\theta})\, p(\vec{\theta})}{p(\mathcal{D})}$$

First, the **likelihood** $p(\mathcal{D} \mid \vec{\theta})$ can be computed by

$$p(\mathcal{D} \mid \vec{\theta}) = \prod_{i=1}^{N} p\left(y^{(i)}|\vec{\theta}, \vec{x}^{(i)}\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)} - \vec{h}^T\left(\vec{x}^{(i)}\right)\vec{\theta}\right)^2\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^N \exp\left(-\frac{1}{2\sigma^2}\left(\left\|\vec{y} - H\vec{\theta}\right\|\right)^2\right) \qquad \text{where } H_{\text{ij}} = h_j(x^{(i)})$$

If we want we can even write down the prior distribution of our dataset $\mathcal{D}$ as

$$p(\mathcal{D}) = \int p\left(\mathcal{D}|\vec{\theta}\right) p\left(\vec{\theta}\right) d\vec{\theta}$$

The **posterior density** for $\vec{\theta}$ (i.e., its new probability density given the new information in $\mathcal{D}$) is

$$p(\vec{\theta}|\mathcal{D}) = \frac{p(\mathcal{D} \mid \vec{\theta})\, p(\vec{\theta})}{p(\mathcal{D})}$$

$$= \frac{1}{p(\mathcal{D})}\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^{N} \exp\left(-\frac{1}{2\sigma^2}\left(\|\vec{y} - H\vec{\theta}\|\right)^2\right) \frac{1}{\sqrt{(2\pi)^d\,|\Sigma|}} \exp\left(-\frac{1}{2}\left(\vec{\theta}\right)^T \Sigma^{-1}\left(\vec{\theta}\right)\right)$$

$$= \frac{1}{p(\mathcal{D})}\left(\frac{1}{\left(\sqrt{2\pi}\right)^{N+d}\,\sigma^N\sqrt{|\Sigma|}}\right) \exp\left(-\frac{1}{2\sigma^2}\left(\|\vec{y} - H\vec{\theta}\|\right)^2 - \frac{1}{2}\left(\vec{\theta}\right)^T \Sigma^{-1}\left(\vec{\theta}\right)\right)$$

We now rearrange the exponent and rewrite the density as

$$= \mathrm{K} \exp\left(-\left(\vec{\theta} - \frac{\mathrm{B}^{-1}C}{2}\right)^{\mathrm{T}} \mathrm{B}\left(\vec{\theta} - \frac{\mathrm{B}^{-1}C}{2}\right)\right)$$

where K is the Normalization constant (no dependence on $\vec{\theta}$)

$$\mathrm{K} := \frac{1}{p(\mathcal{D})}\left(\frac{1}{\left(\sqrt{2\pi}\right)^{N+d}\sigma^{N}\sqrt{|\Sigma|}}\right)\exp\left(-A + \frac{(B^{-1}C)^{T}B(B^{-1}C)}{4}\right)$$

Here: $A = \frac{\vec{y}^{T}\vec{y}}{2\sigma^{2}}$, $B = \frac{H^{T}H}{2\sigma^{2}} + \frac{\Sigma^{-1}}{2}$ and $C^{T} = \frac{\vec{y}^{T}H}{\sigma^{2}}$

$\vec{\theta}|\mathcal{D}$ is also *a normal distribution with* ***mean***

$$E\left(\vec{\theta}|\mathcal{D}\right) = \frac{\mathrm{B}^{-1}C}{2} = (H^{T}H + \Sigma^{-1}\sigma^{2})^{-1}H^{T}\vec{y}$$

The **covariance matrix** is

$$Cov\left(\vec{\theta}|\mathcal{D}\right) = (2B)^{-1} = (H^{T}H + \sigma^{2}\Sigma^{-1})^{-1}\sigma^{2}$$

Posterior distribution of $f(\vec{x})$ for fixed $\vec{x}$ (For prediction)

Now fixing $\vec{x}$ and recalling from that

$$f(\vec{x}) = \vec{h}^T(\vec{x}\,)\vec{\theta}$$

with fixed basis functions vector $\vec{h}^T(\vec{x}\,)$.

So, $f(\vec{x})|\mathcal{D}$ is a multivariate normal with mean

$$E[f(\vec{x})|\mathcal{D}] = \vec{h}^T(\vec{x}\,)E\left[\vec{\theta}\middle|\mathcal{D}\right] = \vec{h}^T(\vec{x}\,)(H^TH + \Sigma^{-1}\sigma^2)^{-1}H^T\vec{y}$$

And covariance matrix

$$Cov[f(\vec{x})|\mathcal{D}] = \vec{h}^T(\vec{x}\,)(H^TH + \sigma^2\Sigma^{-1}\,)^{-1}\sigma^2\vec{h}\,(\vec{x}\,)$$

❑ Connection to ridge regression

Q: How to choose covariance matrix Σ for the prior distribution of $\vec{\theta}$?

This is a big question in Bayesian inference. We can take the prior covariance to be $\tau\Sigma$, for correlation Σ, e.g., τI

$\vec{\theta}|\mathcal{D}$ has ***mean***

$$E(\vec{\theta}|\mathcal{D}) = \frac{\mathrm{B}^{-1}C}{2} = \left(H^T H + \frac{1}{\tau}\Sigma^{-1}\sigma^2\right)^{-1} H^T \vec{y}$$

The **covariance matrix** is

$$Cov(\vec{\theta}|\mathcal{D}) = (2B)^{-1} = \left(H^T H + \frac{1}{\tau}\sigma^2\Sigma^{-1}\right)^{-1} \sigma^2$$

Notice that as $\tau \to \infty$, this becomes a constant prior and in turn reduces to the linear regression. We again see that Bayesian reasoning leads to ridge regression, and in turn that another way to understand the λ of ridge regression is as an inverse variance in the parameters.

Actually, if we use Maximum a Posteriori estimator, it is equivalent minimize ridge loss.

Summary of the Bayesian Method

There are four main steps to the Bayesian approach to probabilistic inference:

1. Likelihood. First, we construct the likelihood (or model), $p(\mathcal{D}\,|\vec{\theta})$. This serves to describe the mechanism giving rise our observations D given a particular value of the parameter of interest $\vec{\theta}$.

2. Prior. Next, we summarize our prior beliefs about the parameters $\vec{\theta}$, which we encode via a probability distribution $p(\vec{\theta})$.

3. Posterior. Given some observations $\mathcal{D}$, we obtain the posterior distribution $p(\vec{\theta}|\mathcal{D})$ using Bayes' theorem.

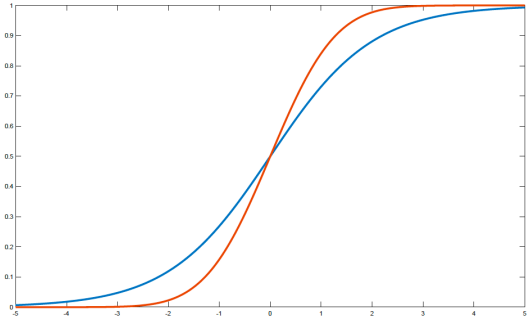
4. Inference. We now use the posterior distribution to draw further conclusions as required. (open-end)

➢ Bayesian Logistic Regression/The Laplace Approximation

Consider the GLM for independent Bernoulli observations $y^{(i)} \sim Ber\left(\mu^{(i)}\right)$ for $i = 1, \ldots, N$

Recall Binary Classification Model Assumption

$$\mu =: E(Y|X) = P\left(Y = 1 \middle| \vec{X}\right) = \sigma\left(\vec{\theta}^T \vec{X}\right)$$



Logistics Model: $\mu(\vec{x}) = \dfrac{1}{1 + e^{-\vec{x}^T \vec{\beta}}}$ when $\sigma(u) = \dfrac{1}{1+e^{-u}}$

Probit model $\mu(\vec{x}) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{\vec{\beta}^T \vec{x}} \exp\left(-\frac{u^2}{2}\right) du$ when $\sigma(u)$ is cdf of normal.

Traditional approach to logistic regression: **MLE**

$$\hat{\vec{\theta}} = \underset{\vec{\theta}}{\operatorname{argmax}}\, P\left(\vec{y} \middle| X, \vec{\theta}\right) = \underset{\vec{\theta}}{\operatorname{arg}min}\, -\log P\left(\vec{y} \middle| X, \vec{\theta}\right)$$

➢ Bayesian logistic regression

Select a **prior** distribution for the parameter $\vec{\theta}$

Assuming that $\vec{\theta} \sim N(0, \Sigma)$ with **prior** covariance matrix Σ, that is

$$p(\vec{\theta}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{\theta})^T \Sigma^{-1}(\vec{\theta})\right)$$

Derive the **posterior distribution** for $\vec{\theta}$ by Bayes' theorem

$$p(\vec{\theta}|\mathcal{D}) = \frac{p(\mathcal{D} \mid \vec{\theta})\, p(\vec{\theta})}{p(\mathcal{D})} = \frac{P(\mathcal{D} \mid \vec{\theta})\, P(\vec{\theta})}{\int P(\mathcal{D} \mid \vec{\theta}')\, P(\vec{\theta}')\, d\vec{\theta}'}$$

Unfortunately, the product of the Gaussian prior on $\vec{\theta}$ and the likelihood (for either choice of sigmoid) does not result in a posterior distribution in a nice parametric family that we know. The normalization constant (the evidence) $p(\mathcal{D}) = p(\vec{y} \mid X)$ is intractable as well.

Two main approaches:

- Use a deterministic method to find an approximation to the posterior (that will typically live inside a chosen parametric family). E.g., **Laplace approximation.**

- Derive an algorithm to draw samples from the posterior distribution, which we may use to, for example, make Monte Carlo estimates to expectations.

❑ Laplace Approximation to the Posterior

Suppose we have an arbitrary parameter prior $p(\vec{\theta})$ and an arbitrary likelihood $p(\mathcal{D}|\vec{\theta})$, and wish to approximate the posterior

$$p(\vec{\theta}|\mathcal{D}) = \frac{1}{Z} p(\mathcal{D}|\vec{\theta})p(\vec{\theta}),$$

where Z is the unknown normalization constant. Define

$$\phi(\vec{\theta}) := \log p(\mathcal{D}|\vec{\theta}) + \log p(\vec{\theta})$$

The **Laplace approximation** is based on a **Taylor expansion** to $\phi(\vec{\theta})$ around its maximum $\hat{\theta}$.

$$\phi(\vec{\theta}) \approx \phi(\hat{\theta}) - \frac{1}{2} (\vec{\theta} - \hat{\theta})^T H(\vec{\theta} - \hat{\theta})$$

where H is the Hessian matrix of $\phi(\vec{\theta})$ at $\hat{\theta}$.

Find **maximum a posteriori (MAP),** use gradient $\nabla\phi(\vec{\theta})=0$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \phi(\vec{\theta}) = \underset{\theta}{\operatorname{argmax}}\, p(\vec{\theta}|\mathcal{D})$$

"optimization is easier than integration".

Then we have

$$p(\vec{\theta}|\mathcal{D}) \propto \exp \phi(\hat{\theta}) \exp\left[-\frac{1}{2}\left(\vec{\theta} - \hat{\theta}\right)^T H\left(\vec{\theta} - \hat{\theta}\right)\right]$$

Proportional to **Gaussian** distribution!

So, $p(\vec{\theta}|\, D) \approx N(\hat{\theta}, H^{-1})$.

So we also have the normalizing constant Z,

$$Z = \exp\left(\phi(\hat{\theta})\right)\sqrt{(2\pi)^d |H^{-1}|} = \exp\left(\phi(\hat{\theta})\right)\sqrt{\frac{(2\pi)^d}{|H|}}$$

❑ Making Predictions

Suppose we already have $p(\vec{\theta}|\mathcal{D})$, e.g., by Laplace $\approx N(\hat{\theta}, H^{-1})$.

Given a test input $\vec{x}_*$, in the Bayesian approach, we marginalize the unknown parameters to find the the predictive distribution:

$$P(y_* = 1 \mid \vec{x}_*, \mathcal{D}) = \int P(y_* = 1 \mid \vec{x}_*, \mathcal{D}, \vec{\theta})\, p(\vec{\theta}|\mathcal{D}) d\vec{\theta}$$

$$= \int \sigma(\vec{x}_*^T\, \vec{\theta}) p(\vec{\theta}|\mathcal{D}) d\vec{\theta}$$

When σ is logistic function, the integral can not be evaluated.
When $\sigma(u) = g(u)$ is **cdf normal function**, the integral can be evaluated

$$P(y_* = 1 \mid \vec{x}_*, \mathcal{D}) = \int g(\vec{x}_*^T\, \vec{\theta}) p(\vec{\theta}|\mathcal{D}) d\vec{\theta}$$

$$a := \vec{x}_*^T\, \vec{\theta}$$

$$= \int_{-\infty}^{\infty} g(a) p(a|\mathcal{D})\, da$$

From $\mathbb{R}^d$ to $\mathbb{R}$

$$p(a|\mathcal{D}) = N(\vec{x}_*^T\ \hat{\vec{\theta}}, \vec{x}_*^T H^{-1} \vec{x}_*^T)$$

$$P(y_* = 1 \mid \vec{x}_*, \mathcal{D}) = E\big(g(a)\big) = g\left(\frac{\vec{x}_*^T\ \hat{\vec{\theta}}}{\sqrt{1 + \vec{x}_*^T H^{-1} \vec{x}_*^T}}\right)$$

The Bayesian Information Criterion (BIC)

Given a set of models $\{\mathcal{M}_i\}$, and observed data $\mathcal{D}$ with N data points , we compute the following statistic for each:

$$BIC_i := \log p(\mathcal{D}|\widehat{\theta}_i) - \frac{d}{2}\log N$$

$$\log \text{true predicted goodness} = \log Likelihood - penalty$$

Very roughly approximate

Reason for BIC:

$$p\left(\vec{\theta}\middle|\mathcal{D},\mathcal{M}\right) = \frac{1}{Z}\, p(\mathcal{D}|\vec{\theta},\mathcal{M})p\left(\vec{\theta}|\mathcal{M}\right) \text{ Implies } Z = p(\mathcal{D}|\mathcal{M})$$

From Laplace Approximation to the posterior distribution

$$\log p(\mathcal{D}|\mathcal{M}_i) = \log Z \approx \log p(\mathcal{D}|\widehat{\theta}_i) + \log p\left(\widehat{\theta}_i\right) + \frac{d}{2}\log 2\pi - \frac{1}{2}\log|H|$$

Here, $\widehat{\theta}_i$ is the **maximum a posteriori (MAP)** estimate.

BIC penalizes model complexity more heavily than AIC. $(-\frac{d}{N})$

➢ Bayesian Model Comparison/Selection

Given a finite set of models $\{\mathcal{M}_i\}_{i=1}^{L}$, and observed data $\mathcal{D}$. A model $\mathcal{M}_i$ is probability distributions with parameters $\vec{\theta}_i$. Suppose the data $\mathcal{D}$ is generated from one of these models but we don't know which.

Give a data $\mathcal{D}$, we wish to evaluate the **posterior**:

$$P(\mathcal{M}_i|\mathcal{D}) = \frac{p(\mathcal{D}\,|\mathcal{M}_i)\,p(\mathcal{M}_i)}{p(\mathcal{D})} = \frac{p(\mathcal{D}\,|\mathcal{M}_i)\,P(\mathcal{M}_i)}{\sum_j p(\mathcal{D}\,|\mathcal{M}_j)\,P(\mathcal{M}_j)}$$

$P(\mathcal{M}_j)$: **Prior** distribution over models that we have selected. E.g., uniform distribution. That is assume all models are given equal prior probability.

$p(\mathcal{D}\,|\mathcal{M}_i)$: **likelihood**. (**model evidence**)

Averaging over all possible parameters. (joint distribution with $\vec{\theta}_i$, then marginalize.)

$$p(\mathcal{D}\,|\mathcal{M}_i) = \int p\left(\mathcal{D}\middle|\vec{\theta}_i, \mathcal{M}_i\right) p(\vec{\theta}_i|\mathcal{M}_i)\, d\vec{\theta}_i$$

Model Selection:

Suppose now that we have exactly two models for the observed data that we wish to compare: $\mathcal{M}_1$ and $\mathcal{M}_2$.

$$\frac{P(\mathcal{M}_1|\mathcal{D})}{P(\mathcal{M}_2|\mathcal{D})} = \frac{p(\mathcal{D}\,|\mathcal{M}_1)\,p(\mathcal{M}_1)}{p(\mathcal{D}\,|\mathcal{M}_2)\,p(\mathcal{M}_2)}$$

posterior odds

If we assume all models are given equal prior probability, then posterior odds Is the same as **Bayes factor:**

$$\frac{p(\mathcal{D}\,|\mathcal{M}_1)}{p(\mathcal{D}\,|\mathcal{M}_2)}$$

Example: Coin head ratio

Suppose I am presented with a coin and want to compare two models for explaining its behavior.

$\mathcal{M}_1: P(head) = \frac{1}{2}$ No parameter in this model.

$\mathcal{M}_2$: Assume the heads probability is fixed to an unknown value $\theta \in (0, 1)$, with a uniform **prior** on θ: $p(\theta \mid \mathcal{M}_2) = 1$ (this is equivalent to a Beta prior on θ $with$ $\alpha = \beta = 1$).

For simplicity, we choose a uniform model prior: $P(\mathcal{M}_1) = P(\mathcal{M}_1) = \frac{1}{2}$.

Suppose we flip the coin $n = 200$ times and observe $x = 115$ heads.

Which model should we prefer in light of this data?

https://en.wikipedia.org/wiki/Bayes_factor#Example

The model evidence for $\mathcal{M}_1$ is quite straightforward, as it has no parameters:

$$p(\mathcal{D}\,|\mathcal{M}_1) = \text{Binomial}\left(n, x, \frac{1}{2}\right) = \binom{200}{115}\left(\frac{1}{2}\right)^{200} \approx 0.005956$$

The model evidence for $\mathcal{M}_2$:

$$p(\mathcal{D}\,|\mathcal{M}_2) = \int p\left(\mathcal{D}\middle|\vec{\theta}\,, \mathcal{M}_2\,\right) p(\vec{\theta}\,|\mathcal{M}_2)\, d\vec{\theta}$$

$$= \int_0^1 \binom{200}{115} (\theta)^{115}(1-\theta)^{85} d\theta = \frac{1}{201} \approx 0.004975$$

The Bayes factor in favor of $\mathcal{M}_1$ is approximately 1.2, so the data give very weak evidence in favor of the simpler model $\mathcal{M}_1$.

An interesting aside here is that a frequentist hypothesis test would **reject the null hypothesis** $\theta = \frac{1}{2}$ $at\ the\ \alpha = 0.05\ level.$

The probability of generating **at least** 115 heads under model $\mathcal{M}_1$ is approximately $P(x \geq 115\ |\mathcal{M}_1) \approx 0.02$

(similarly, the probability of generating at least 115 tails is also 0.02), so a two-sided test would give a **p-value** of approximately 4%.

However, that a non-uniform prior (for example one that reflects the fact that you expect the number of success and failures to be of the same order of magnitude) could result in a Bayes factor that is more in agreement with the frequentist hypothesis test.

One spin on Bayesian decision theory is that it automatically gives a preference towards simpler models, in line with Occam's razor: "entities should not be multiplied beyond necessity". Consider $p(\mathcal{D}\ |\mathcal{M})$, more complex models can explain more datasets, so the support of this distribution is wider in the sample space. But note that the distribution must normalize over the sample space as well, so we pay a price for generality.

❑ Model selection for Bayesian linear regression

Suppose $\mathcal{M}_{[s]}$corresponds to order s −polynomial regression.

Assume $y^{(i)} = \vec{h}_s^{\,T}\left(\vec{x}^{(i)}\right)\vec{\theta} + \epsilon_i$ where ϵ_i are iid $N(0, \sigma^2)$ random variables, and

$\vec{h}_s^{\,T}(\vec{x}\,) = [h_1(\vec{x}\,) \quad h_2(\vec{x}) \quad \ldots \quad h_{p(s)}(\vec{x})]$ where $\vec{h}_i(\vec{x})$ give the degree i $monomials$

From Bayesian linear regression, assume **prior** $\vec{\theta} \sim N(\vec{\mu}, \Sigma)$ we have

$$p(\mathcal{D}\,|\mathcal{M}_i) = p(\vec{y}\,|X, \sigma^2, \mathcal{M}_i) = N(H\vec{\mu}, H\Sigma H^T + \sigma^2 I),$$

Again, however, the simpler model will be preferred due the Occam's razor effect.

❑ Bayesian Model Averaging

When making predictions, we should theoretically use the sum rule to marginalize the unknown model $\mathcal{M}_i$, e.g. Bayesian model averaging:

$$p(y_* \,|\vec{x}_*, \mathcal{D}) = \sum_i p(y_*|\vec{x}_*, \mathcal{D}, \mathcal{M}_i) P(\mathcal{M}_i|\mathcal{D})$$

Both model averaging and model selection are used in practice.

MATLAB: https://www.mathworks.com/matlabcentral/fileexchange/29326-bms-toolbox-for-matlab-bayesian-model-averaging-bma

Python: https://www.kaggle.com/code/billbasener/bayesian-model-averaging-regression-tutorial/notebook
https://www.kaggle.com/code/billbasener/bayesian-model-averaging-logistic-regression

Papers: https://www.stat.colostate.edu/~jah/papers/statsci.pdf

https://arxiv.org/pdf/1509.08864.pdf

Bayesian Statistics v.s. Frequentist Statistics

Frequentist: The parameter $\vec{\theta}$ is not a random variable.
Bayesian: The parameter is a random variable.

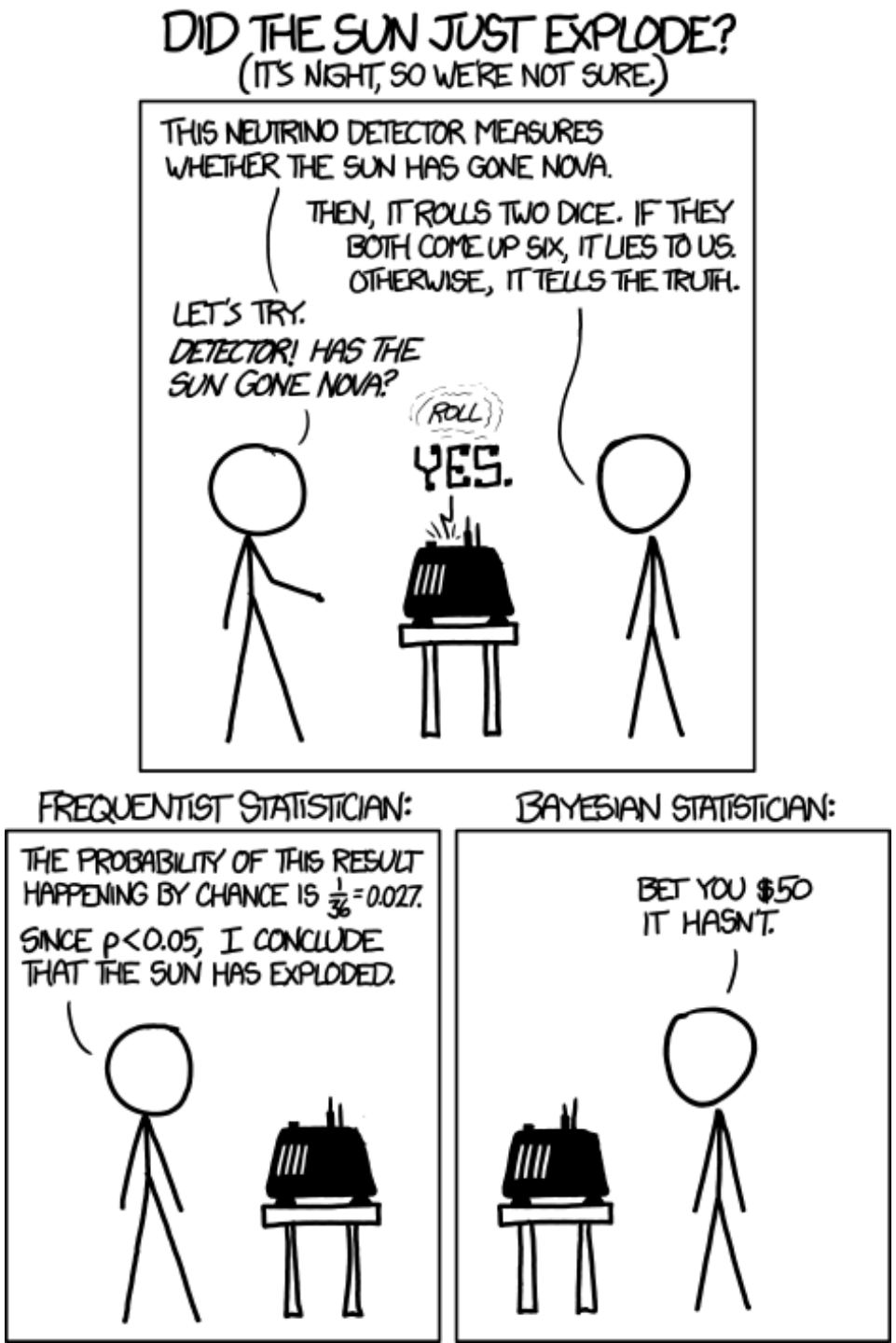
Frequentist statistics never uses or calculates the probability of the hypothesis. $p(\vec{\theta}\,)$

Bayesian uses probabilities of data and probabilities of both hypothesis.

Frequentist methods do not demand construction of a prior and depend on the probabilities of observed and unobserved data.

On the other hand, Bayesian methods depend on a prior and on the probability of the observed data

	Bayesian	Frequentist
Parameter	Random	Fixed
Inference	Based on posterior	Based on likelihood
Background knowledge	Yes	No
Representative algorithm	Gibbs sampling	Restricted maximum likelihood
Point estimation	Many point estimates from posterior (e.g., posterior mean, maximum a posteriori, posterior median)	One point estimate by a specific estimator (e.g., restricted maximum likelihood estimate)
Interval estimation	Credible interval	Confidence interval



https://www.explainxkcd.com/wiki/index.php/1132:_Frequentists_vs._Bayesians

https://xkcd.com/1132/

References:

- Textbooks:

✓ **Hastie**: Sec 8.2

✓ **Bishop**: Sec 1.2, 3.3, 3.4, 4.4, 4.5

✓ **Murphy 1**: Sec 2.3, 4.1-4.6, 5.1-5.4

✓ **Osvaldo Martin**: <Bayesian analysis with python>

✓ *Bayesian Reasoning and Machine Learning* by David Barber. Geared

✓ Gaussian Processes for Machine Learning by Carl Rasmussen and Christopher William

✓ Information Theory, Inference, and Learning Algorithms by David J. C. Mackay.

- Online resources

✓ https://statswithr.github.io/book/the-basics-of-bayesian-statistics.html

✓ https://metacademy.org/roadmaps/rgrosse/bayesian_machine_learning

A PhD thesis: Bayesian Methods and Machine Learning for Processing Text and Image Data
https://dc.uwm.edu/cgi/viewcontent.cgi?article=2638&context=etd

Limitations of Bayesian Methods

The challenges of specifying an appropriate model and prior and of performing the required Bayesian computations are not always easy to meet. Here are some currently di cult situations for Bayesian methods:

Problems requiring specific priors in vague situations.
An example: We have a sample of points that we know come from a convex polyhedron, whose volume we wish to estimate. A Bayesian method will need a prior over possible polyhedrall which could be done, but probably requires a lot of thought. But a simple non-Bayesian estimate based on cross validation is (usually) available.

Problems where the likelihood has an intractable normalizing constant.
Boltzmann machines are an example even maximum likelihood is hard, and Bayesian inference seems out of the question at the moment.

Problems with complex, unknown error distributions.
We can try to model the error, but it may be difficult. A bad model may lead to “overfitting” data where the model thinks the error is less than it is. A cross-validation approach to regularization can sometimes work better in such situations.

Misguided “Bayesian” Methods

Some attempts at Bayesian inference are just mis-guided I either falling prey to various problems, or failing from the start because they don't take the Bayesian framework seriously. Here are some commonly observed errors:

Improper posterior distributions: Priors that are “improper” e.g., uniform over $(-\infty, \infty)$ can sometimes be convenient, but not if the posterior ends up improper.

Ridiculous priors: Surprisingly many people use a gamma(0.001,0.001) prior for inverse variances, even though it's absurd.

Relying on MAP estimation: Using the mode of the posterior (MAP estimate) can be a crude but useful approximation. But if a method “works” only because of this approximation, it's not Bayesian.

Meaningless marginal likelihoods: Often based on priors that aren't well considered, or computed using hopelessly inaccurate methods.

Fear of unidentifiability: Irrationally worried that some transformations of the model (e.g, permutations of hidden units) leave the probability of the data unchanged, some people impose constraints that destroy interpretability, introduce arbitrary asymmetries in the prior, and hinder MCMC convergence.

Successes of Bayesian Methodology

Bayesian neural network models and Gaussian process models have been applied to many practical problems, with excellent results. See Lampinen and Vehtari (2001) for examples.
Bayesian neural networks produced winning results in the NIPS*2003 feature selection challenge (http://www.nipsfsc.ecs.soton.ac.uk), with some help from Dirichlet diffusion tree models.

Dirichlet process mixture models are widely used in the statistical literature, and they and their hierarchical extensions are becoming popular as models of documents for information retrieval.
Bayesian methods have increased in popularity in statistics since MCMC methods were popularized around 1990. Complex hierarchical models, with many layers of parameters are often used, and are the only viable approach to some problems.

Further references:

- J. M. Bernardo, A. F. M. Smith (1994) Bayesian Theory, Wiley.
- A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin (2003) Bayesian Data Analysis, 2nd edition, Chapman&Hall/CRC.
- J. S. Liu (2001) Monte Carlo Strategies in Scienti c Computing, Springer-Verlag.
- R. M. Neal (1993) Probabilistic Inference Using Markov Chain Monte Carlo Methods. http://www.cs.utoronto.ca/~radford/review.abstract.html
- R. M. Neal (1996) Bayesian Learning for Neural Networks, Springer-Verlag.
- D. J. C. MacKay (2003) Information Theory, Inference, and Learning Algorithms. http://wol.ra.phy.cam.ac.uk/mackay/itila/book.html

Bayesian Modeling and Computation in Python, By Osvaldo A. Martin, Ravin Kumar, Junpeng Lao https://bayesiancomputationbook.com/welcome.html
Code:
https://github.com/BayesianModelingandComputationInPython/BookCode_Edition1

- Tom Griffiths' reading list: http://www-psych.stanford.edu/~gruffydd/bayes.html
- MCMC Preprint Service: http://www.statslab.cam.ac.uk/~mcmc
- The BUGS software for MCMC: http://www.mrc-bsu.cam.ac.uk/bugs
- Software for flexible Bayesian modeling:
- http://www.cs.toronto.edu/~radford/fbm.software.html
- The new on-line journal Bayesian Analysis: http://ba.stat.cmu.edu