Bayesian Statistics

AI Friends Seminar

Ganguk Hwang

Department of Mathematical Sciences
KAIST

Introduction

Bayesian Inference

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available.

Bayesian updating is particularly important in the dynamic analysis of a sequence of data.

Bayesian inference has found application in a wide range of activities, including science, engineering, philosophy, medicine, sport, and law.

(in Wikipedia)

Thomas Bayes and Bayes' Theorem

Thomas Bayes (1701 1761) was an English statistician, philosopher and Presbyterian minister who is known for formulating a specific case of the theorem that bears his name: Bayes' theorem.

Bayes never published what would become his most famous accomplishment; his notes were edited and published after his death by Richard Price.

(in Wikipedia)

Bayes' Theorem

For a partition $\{E_i\}$ of the sample space Ω and an event F,

$$P\{E_i|F\} = \frac{P\{E_i \cap F\}}{P\{F\}} = \frac{P\{F|E_i\}P\{E_i\}}{\sum_{i=1}^{N} P\{E_i\}P\{F|E_i\}}$$

Distribution Function and Parameter

The *distribution function* $F(x)$ of a random variable X is defined by

$$F(x) = P\{X \leq x\}.$$

The *parameter* of a distribution function is the value that determines the distribution.

For instance, when we consider a Bernoulli random variable X with $P\{X = 1\} = p, P\{X = 0\} = 1 - p$, the value p determines the distribution of X and is the parameter of the distribution of X.

Likelihood Function

Likelihood Function
Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed (i.i.d.) random variables with parameter θ. The probability mass (or density) function of X_i is given by $p(x|\theta)$. Then, the likelihood $\mathcal{L}(x_1, x_2, \cdots, x_n|\theta)$ is defined by

$$\mathcal{L}(x_1, x_2, \cdots, x_n|\theta) = p(x_1|\theta)p(x_2|\theta)\cdots p(x_n|\theta)$$

Maximum Likelihood Estimator (MLE) of θ

- The MLE is the value of θ that maximizes the likelihood function.
- Here, we consider the likelihood function as a function of θ.
- To compute the MLE, we usually use $\log(\mathcal{L}(x_1, x_2, \cdots, x_n|\theta))$.

Bayesian models

Bayesian models

In Bayesian models we use the Bayes' rule to obtain unknown probability.

In Bayesian models with two random varaibles X and Y, the following are initially given.

- The data generating distribution: $X|Y \sim p(x|y)$
- The prior distribution $Y \sim p(y)$

We then obtain a sample value $X = x$ and want to estimate the distribution (the posterior distribution) of Y, given $X = x$, i.e., $p(y|x)$.

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x,y)\ dy}.$$

We frequently use $p(y|x) \propto p(x|y)p(y)$.

Bayesian models

Consider a random variable X having distribution function $F(x)$ with unknown parameter θ.

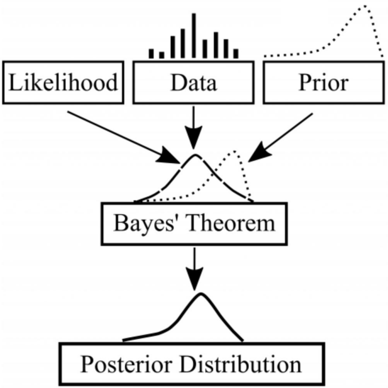
In Bayesian models, the unknown parameter θ is considered *stochastic*. So we believe that $\theta \sim p(\theta)$ where $p(\theta)$ is called a *prior* distribution before sampling. After sampling, using Bayes' rule we obtain so-called a *posterior* distribution as follows:

$$p(\theta|x) = \frac{p(x,\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x,\theta)\ d\theta}.$$

Bayesian models

- $P(\theta)$ is the *prior* which is our belief of θ without considering the data (evidence) $\mathcal{D}$
- $P(\theta|\mathcal{D})$ is the *posterior* which is a refined belief of θ with the evidence $\mathcal{D}$
- $P(\mathcal{D}|\theta)$ is the *likelihood* which is the probability of obtaining the data $\mathcal{D}$ as generated with parameter θ
- $P(\mathcal{D})$ is the *evidence* which is the probability of the data as determined by considering all possible values of θ

Bayesian models



Courtesy: http://jason-doll.com/wordpress/?page_id=127

Figure: Concept of Bayesian models

Maximum A Posterior Estimator

Consider $X \sim p(x|\theta)$ with a prior distribution of $p(\theta)$. Here, θ is the parameter of the distribution of X.

The Maximum A Posterior (MAP) estimator is given by

$$\theta_{MAP} = \mathrm{argmax}_\theta \log p(\theta|X).$$

Note that $p(\theta|X)$ is the posterior distribution of θ and

$$p(\theta|X) = \frac{p(\theta, X)}{p(X)}.$$

Since $p(X)$ is not a function of θ, we have

$$\theta_{MAP} = \mathrm{argmax}_\theta \log p(\theta, X).$$

Maximum A Posterior Estimator

Recall that the Maximum Likelihood Estimator (MLE) is defined by

$$\theta_{MLE} = \mathrm{argmax}_\theta \log p(X|\theta)$$

and the MAP estimator is given by

$$\theta_{MAP} = \mathrm{argmax}_\theta \log p(\theta, X) = \mathrm{argmax}_\theta \log p(X|\theta)p(\theta).$$

So, the MAP estimator can use the prior information, but the MLE cannot.

Bayesain model: an example

Let $X_1, X_2, \cdots, X_n$ be indepedent and identically distributed (i.i.d.) Bernoulli random variables where

$$P\{X_1 = 1|\theta\} = \theta, \qquad P\{X_1 = 0|\theta\} = 1 - \theta.$$

Here, θ is usually called the parameter of Bernoulli distribution. First, observe that

$$\begin{aligned} & P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\theta\} \\ &= \prod_{i=1}^{n} P\{X_i = x_i|\theta\} = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}. \end{aligned}$$

The prior distribution of θ is given by a uniform distribution over $[0, 1]$, i.e.,

$$p(\theta) = 1 \text{ for } 0 \leq \theta \leq 1.$$

After obtaining sample values $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$, we are interested in the posterior distribution of θ.

$$\begin{aligned} & p(\theta | X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\ & \quad \propto P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n | \theta) p(\theta) \\ & \quad \propto \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}. \end{aligned}$$

Considering the normalization constant, we obtain

$$
\begin{aligned}
&p(\theta|X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\
&= \frac{\Gamma(n+2)}{\Gamma(1+\sum_{i=1}^n x_i)\Gamma(1+n-\sum_{i=1}^n x_i)} \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}
\end{aligned}
$$

which is a Beta distribution with parameters $a = \sum_{i=1}^n x_i + 1$ and $b = n - \sum_{i=1}^n x_i + 1$.

c.f.

$$
\text{Beta}(a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}
$$

Note that, when $a = b = 1$, Beta(a, b) is in fact a uniform distribution over $[0, 1]$. That is, the prior and the posterior of θ are both Beta distributions.

Example 1

Suppose we want to buy something from Amazon.com, and there are two sellers offering it for the same price. Seller 1 has 90 positive reviews and 10 negative reviews. Seller 2 has 2 positive reviews and 0 negative reviews. Who should we buy from? (from Machine Learning by K.P. Murphy)

Let θ_1 and θ_2 be the unknown reliabilities of the two sellers. Since we don't know much about them, we will endow them both with uniform priors, $\theta_i \sim \text{Beta}(1,1), i = 1,2$. The posteriors are

$$p(\theta_1|\mathcal{D}_1) = \text{Beta}(91,11), \qquad p(\theta_2|\mathcal{D}_2) = \text{Beta}(3,1).$$

Hence,

$$\begin{aligned} P(\theta_1 > \theta_2|\mathcal{D}_1,\mathcal{D}_2) &= \int_0^1 \int_0^1 I_{\{\theta_1>\theta_2\}} \text{Beta}(\theta_1|91,11)\text{Beta}(\theta_2|3,1)\ d\theta_1 d\theta_2 \\ &= 0.710. \end{aligned}$$

This concludes that we are better off buying from seller 1.

Example 2

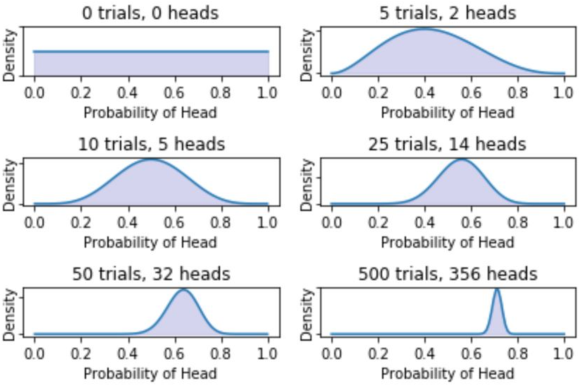


Figure: Bayesian Experiment for Bernoulli distribution with $p = 0.7$

More generally, if we use Beta(a, b) as a prior distribution of θ, we have

$$
\begin{aligned}
&p(\theta|X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\
&\qquad \propto P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\theta)p(\theta) \\
&\qquad \propto P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\theta)\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\
&\qquad \propto \theta^{a+\sum_{i=1}^n x_i - 1}(1-\theta)^{b+n-\sum_{i=1}^n x_i - 1}.
\end{aligned}
$$

Hence, we see that

$$p(\theta|X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \text{Beta}(a + \sum_{i=1}^n x_i, b + n - \sum_{i=1}^n x_i).$$

In this case, the Beta distribution is a *conjugate prior*.

From now on, we use the notation $p(\theta|x_1, x_2, \cdots, x_n)$ or $p(\theta|X)$ for simplicity.

More on Conjugate Distributions

In Bayesian probability theory, if the posterior distributions $p(\theta|x)$ are in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called *conjugate distributions*, and the prior is called a *conjugate prior* for the likelihood function.

As shown before, if we use a conjugate prior, we can obtain a closed-form expression for the posterior. This also shows how the likelihood function updates a prior distribution.

Conjugate Prior for Normal distribution

Assume that X is distributed according to a normal distribution with unknown mean μ and variance $1/\tau$ (or precision τ), i.e.,

$$X \sim \mathcal{N}(\mu, \tau^{-1})$$

and that the prior distribution on μ and τ, (μ, τ), has a Normal-Gamma distribution

$$(\mu, \tau) \sim \mathsf{NormalGamma}(\mu_0, \lambda_0, \alpha_0, \beta_0),$$

for which the density $p(\mu, \tau)$ satisfies

$$p(\mu, \tau) \propto \tau^{\alpha_0 - \frac{1}{2}} \exp[-\beta_0 \tau] \exp\left[-\frac{\lambda_0 \tau (\mu - \mu_0)^2}{2}\right].$$

Suppose that

$$X_1, \ldots, X_n \mid \mu, \tau \sim \text{i.i.d. } \mathcal{N}\left(\mu, \tau^{-1}\right),$$

i.e., the components of $X = (X_1, \cdots, X_n)$ are conditionally independent given μ, τ and the conditional distribution given μ, τ is normal with expectation μ and variance $1/\tau$.

The posterior distribution of μ and τ given this dataset X, is given by

$$P(\tau, \mu \mid X) \propto \mathcal{L}(X \mid \tau, \mu)\; p(\tau, \mu),$$

where $\mathcal{L}$ is the likelihood of the data given the parameters.

Since the data are i.i.d., the likelihood of X is given by

$$\begin{aligned}
\mathcal{L}(X \mid \tau, \mu) &\propto \prod_{i=1}^{n} \tau^{1/2} \exp\left[\frac{-\tau}{2}(x_i - \mu)^2\right] \\
&\propto \tau^{n/2} \exp\left[\frac{-\tau}{2} \sum_{i=1}^{n} (x_i - \mu)^2\right] \\
&\propto \tau^{n/2} \exp\left[\frac{-\tau}{2} \sum_{i=1}^{n} (x_i - \bar{x} + \bar{x} - \mu)^2\right] \\
&\propto \tau^{n/2} \exp\left[\frac{-\tau}{2} \sum_{i=1}^{n} \left((x_i - \bar{x})^2 + (\bar{x} - \mu)^2\right)\right] \\
&\propto \tau^{n/2} \exp\left[\frac{-\tau}{2} \left((n-1)s^2 + n(\bar{x} - \mu)^2\right)\right],
\end{aligned}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

So, the posterior distribution of the parameters is proportional to the prior times the likelihood as given below.

$$
\begin{aligned}
&P(\tau, \mu \mid X) \\
&\propto \mathcal{L}(X \mid \tau, \mu)\ p(\tau, \mu) \\
&\propto \tau^{n/2} \exp\left[\frac{-\tau}{2}\left((n-1)s^2 + n(\bar{x}-\mu)^2\right)\right] \\
&\qquad \times \tau^{\alpha_0 - \frac{1}{2}}\ \exp[-\beta_0 \tau]\ \exp\left[-\frac{\lambda_0 \tau (\mu - \mu_0)^2}{2}\right] \\
&\propto \tau^{\frac{n}{2} + \alpha_0 - \frac{1}{2}} \exp\left[-\tau\left(\frac{1}{2}(n-1)s^2 + \beta_0\right)\right] \\
&\qquad \times \exp\left[-\frac{\tau}{2}\left(n(\bar{x}-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right)\right]
\end{aligned}
$$

The last exponential term is simplified as follows:

$$
\begin{aligned}
& n(\bar{x}-\mu)^2 + \lambda_0(\mu-\mu_0)^2 \\
& = (n+\lambda_0)\mu^2 - 2(n\bar{x}+\lambda_0\mu_0)\mu + n\bar{x}^2 + \lambda_0\mu_0^2 \\
& = (n+\lambda_0)(\mu^2 - 2\frac{n\bar{x}+\lambda_0\mu_0}{n+\lambda_0}\mu) + n\bar{x}^2 + \lambda_0\mu_0^2 \\
& = (n+\lambda_0)\left(\mu - \frac{n\bar{x}+\lambda_0\mu_0}{n+\lambda_0}\right)^2 + n\bar{x}^2 + \lambda_0\mu_0^2 - \frac{(n\bar{x}+\lambda_0\mu_0)^2}{n+\lambda_0} \\
& = (n+\lambda_0)\left(\mu - \frac{n\bar{x}+\lambda_0\mu_0}{n+\lambda_0}\right)^2 + \frac{\lambda_0 n(\bar{x}-\mu_0)^2}{n+\lambda_0}
\end{aligned}
$$

Combining all the results yields

$$
\begin{aligned}
& P(\tau, \mu \mid X) \\
& \propto \tau^{\frac{n}{2}+\alpha_0-\frac{1}{2}} \exp\left[-\tau\left(\frac{1}{2}(n-1)s^2+\beta_0\right)\right] \\
& \qquad \times \exp\left[-\frac{\tau}{2}\left((n+\lambda_0)\left(\mu-\frac{n\bar{x}+\lambda_0\mu_0}{n+\lambda_0}\right)^2+\frac{\lambda_0 n(\bar{x}-\mu_0)^2}{n+\lambda_0}\right)\right] \\
& \propto \tau^{\frac{n}{2}+\alpha_0-\frac{1}{2}} \exp\left[-\tau\left(\beta_0+\frac{1}{2}\left((n-1)s^2+\frac{\lambda_0 n(\bar{x}-\mu_0)^2}{n+\lambda_0}\right)\right)\right] \\
& \qquad \times \exp\left[-\frac{\tau}{2}(n+\lambda_0)\left(\mu-\frac{n\bar{x}+\lambda_0\mu_0}{n+\lambda_0}\right)^2\right]
\end{aligned}
$$

That is, the posterior is exactly in the same form as a Normal-Gamma distribution, i.e.,

$$P(\tau, \mu \mid X) = \mathsf{NormalGamma}(\mu_1, \lambda_1, \alpha_1, \beta_1)$$

where

$$\begin{aligned}
\mu_1 &= \frac{n\bar{x} + \lambda_0\mu_0}{n + \lambda_0}, \\
\lambda_1 &= n + \lambda_0, \\
\alpha_1 &= \alpha_0 + \frac{n}{2}, \\
\beta_1 &= \beta_0 + \frac{1}{2}\left((n-1)s^2 + \frac{\lambda_0 n(\bar{x} - \mu_0)^2}{n + \lambda_0}\right)
\end{aligned}$$

Wishart Distribution

Suppose $\mathbf{X}$ is a $p \times n$ matrix, each column of which is independently drawn from a p-variate normal distribution with zero mean

$$\mathbf{x}_i=(x_i^1, \cdots, x_i^p)^\top \sim \mathcal{N}_p(0, \mathbf{\Sigma}), 1 \le i \le n.$$

Then, the Wishart distribution is the probability distribution of the $p \times p$ random matrix

$$\mathbf{S} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top, \qquad \mathbf{S} \sim \mathsf{W}_p(\mathbf{\Sigma}, n).$$

The positive integer n is the degrees of freedom. For $n \ge p$ the matrix $\mathbf{S}$ is invertible with probability 1 if $\mathbf{\Sigma}$ is invertible.
If $p = \mathbf{\Sigma} = 1$, then this distribution is a chi-squared distribution with n degrees of freedom.

The probability density function of $\mathbf{S} \sim \mathsf{W}_p(\mathbf{\Sigma}, n)$ is given by

$$f(\mathbf{S}) = \frac{|\mathbf{S}|^{\frac{n-p-1}{2}}}{|\mathbf{\Sigma}|^{\frac{n}{2}} 2^{\frac{np}{2}} \mathbf{\Gamma}_p\left(\frac{n}{2}\right)} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S})\right)$$

where $\mathbf{\Gamma}_p(x) = \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \mathbf{\Gamma}(x + \frac{1-j}{2})$.

Inverse-Wishart Distribution

The Inverse-Wishart distribution is a probability distribution on positive definite matrices.

We say that $\mathbf{T}$ follows the Inverse-Wishart distribution with $\mathbf{\Psi}$ and m, denoted by $\mathbf{T} \sim \mathsf{InvW}_p(\mathbf{\Psi}, m)$ if its inverse $\mathbf{T}^{-1}$ follows $\mathsf{W}_p(\mathbf{\Psi}^{-1}, m)$.

The probability density function of $\mathbf{T} \sim \mathsf{InvW}_p(\mathbf{\Psi}, m)$ is given by

$$f(\mathbf{T}) = \frac{|\mathbf{\Psi}|^{\frac{m}{2}}}{|\mathbf{T}|^{\frac{m+p+1}{2}} 2^{\frac{mp}{2}} \mathbf{\Gamma}_p\left(\frac{m}{2}\right)} \exp\left(-\frac{1}{2}\mathsf{tr}(\mathbf{\Psi}\mathbf{T}^{-1})\right)$$

where $\mathbf{\Gamma}_p(x) = \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \mathbf{\Gamma}(x + \frac{1-j}{2})$.

We now show that the Inverse-Wishart distribution is conjugate prior on the covaraince matrix parameter of a multivariate normal distribution. Consider

$$\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n), \mathbf{x}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} \sim \mathsf{InvW}_p(\boldsymbol{\Psi}, m)$$

Then the posterior distribution of $\boldsymbol{\Sigma}$ satisfies

$$\begin{aligned}
f(\boldsymbol{\Sigma}|\mathbf{X}) &\propto f(\mathbf{X}|\boldsymbol{\Sigma}) g(\boldsymbol{\Sigma}|\boldsymbol{\Psi}, m) \\
&\propto (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}_i\right) \\
&\quad \times \frac{|\boldsymbol{\Psi}|^{\frac{m}{2}}}{|\boldsymbol{\Sigma}|^{\frac{m+p+1}{2}} 2^{\frac{mp}{2}} \boldsymbol{\Gamma}_p\left(\frac{m}{2}\right)} \exp\left(-\frac{1}{2}\mathsf{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})\right) \\
&\propto |\boldsymbol{\Sigma}|^{-\frac{n}{2}-\frac{m+p+1}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}_i - \frac{1}{2}\mathsf{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})\right)
\end{aligned}$$

Noting that

$$
\begin{aligned}
\sum_{i=1}^{n} \mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}_i = \mathrm{tr}\left(\sum_{i=1}^{n} \mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}_i\right) &= \sum_{i=1}^{n} \mathrm{tr}(\mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}_i) \\
&= \sum_{i=1}^{n} \mathrm{tr}(\mathbf{x}_i \mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1}) = \mathrm{tr}(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1}) \\
&= \mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}), \text{ where } \mathbf{S} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top,
\end{aligned}
$$

we obtain

$$
\begin{aligned}
f(\boldsymbol{\Sigma}|\mathbf{X}) &\propto |\boldsymbol{\Sigma}|^{-\frac{n+m+p+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})\right) \\
&\propto |\boldsymbol{\Sigma}|^{-\frac{n+m+p+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}((\mathbf{S} + \boldsymbol{\Psi})\boldsymbol{\Sigma}^{-1})\right).
\end{aligned}
$$

Bayesian Decision Theory

The mode of a posterior distribuiton, e.g., the MAP estimator, is often a very poor choice as a summary because the mode is usually quite untypical of the distribution, unlike the mean and the median.

In this case we use Bayesian decision theory where a loss function $L(\theta, \hat{\theta})$ is considered. Here, $L(\theta, \hat{\theta})$ is the loss we have if the truth is θ and our estimate is $\hat{\theta}$.

With a loss function and a posterior distribution we are trying to minimize

$$\underset{\hat{\theta}}{\operatorname{argmin}}\, E[L(\theta, \hat{\theta})|\mathcal{D}].$$

By choosing different loss functions we have different estimators other than the MAP estimator.

Bayesian Decision Theory

- $L(\theta, \hat{\theta}) = I_{\{\hat{\theta} \neq \theta\}}$: The MAP estimator

$$E[L(\theta, \hat{\theta})|\mathcal{D}] = p(\hat{\theta} \neq \theta|\mathcal{D}) = 1 - p(\theta|\mathcal{D})$$

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

$$E[L(\theta, \hat{\theta})|\mathcal{D}] = E[(\hat{\theta} - \theta)^2|\mathcal{D}] = E[(\hat{\theta} - E[\hat{\theta}])^2|\mathcal{D}] + (E[\hat{\theta}] - \theta)^2$$

- $L(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$: the median

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

References

- K.P. Murphy, Machine Learning, The MIT Press, 2012.