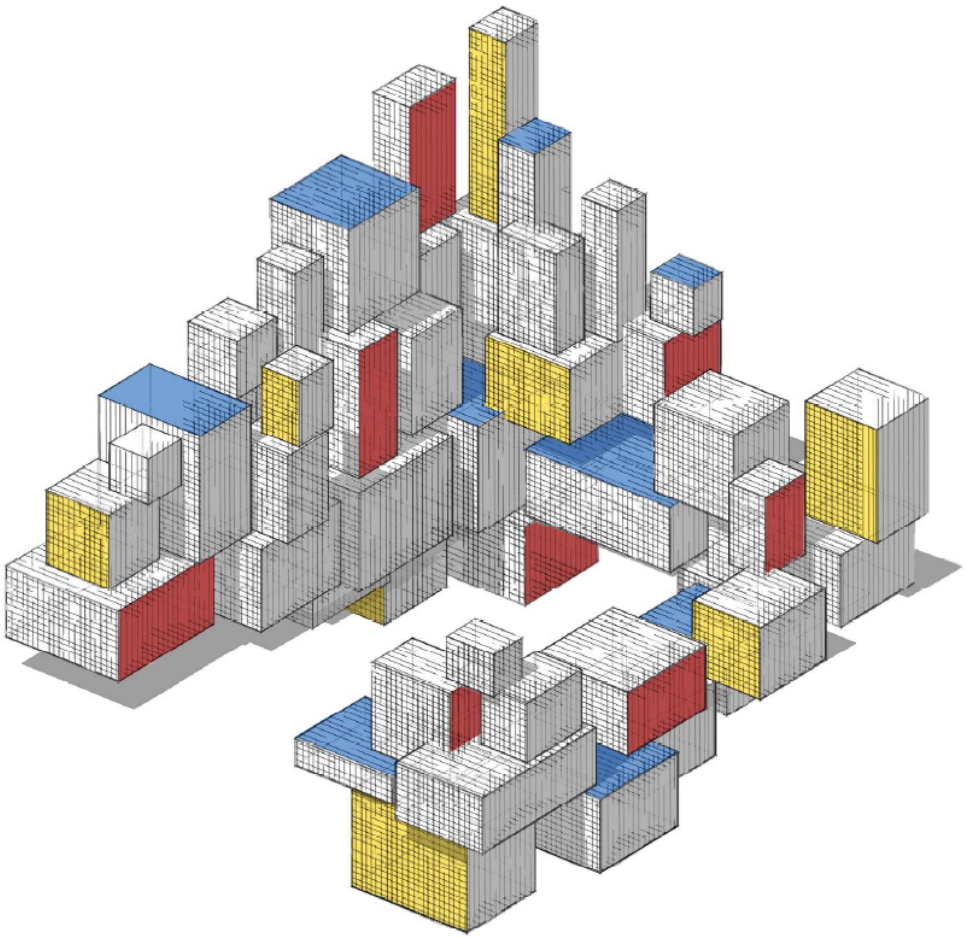
Bayesian Learning



Purpose

In this lecture we discuss:

- Bayesian unsupervised learning
- Prior, likelihood, posterior pdfs
- Bayesian normal model
- MAP estimates and posterior mean

Bayesian Unsupervised Learning

In Bayesian unsupervised learning, we seek to approximate the unknown joint density $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ of the training data $\mathcal{T}_n = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ via a joint pdf of the form

$$\int \left(\prod_{i=1}^{n} g(\boldsymbol{x}_i \,|\, \boldsymbol{\theta}) \right) w(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}, \tag{1}$$

where

- $g(\cdot \,|\, \boldsymbol{\theta}) \in \mathcal{G}_p := \{g(\cdot \,|\, \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p\}$,
- $w(\boldsymbol{\theta}) \in \mathcal{W}_p$.

Note that this includes the case of exchangeable training data.

Notation

Following standard practice in a Bayesian context, instead of writing $f_X(x)$ and $f_{X\,|\,Y}(x \mid y)$ for the pdf of X and the conditional pdf of X given Y, one simply writes $f(x)$ and $f(x \mid y)$. If Y is a different random variable, its pdf (at y) is thus denoted by $f(y)$.

Thus, we will use the same symbol

- g for (conditional) approximating probability densities, and
- f for (conditional) true and unknown probability densities.

Using Bayesian notation, we can write $g(\tau \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} g(\boldsymbol{x}_i \mid \boldsymbol{\theta})$ and thus the approximating joint pdf (1) can then be written as $\int g(\tau \mid \boldsymbol{\theta})\, w(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$ and the true unknown joint pdf as $f(\tau) = f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$.

Kullback–Leibler Risk

We can use the Kullback–Leibler risk to measure the discrepancy between the proposed approximation (1) and the true $f(\tau)$:

$$\ell(g) = \mathbb{E} \ln \frac{f(\mathcal{T})}{\int g(\mathcal{T} \mid \boldsymbol{\theta})\, w(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}} = \int f(\tau) \ln \frac{f(\tau)}{\int g(\tau \mid \boldsymbol{\theta})\, w(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}}\, \mathrm{d}\tau.$$

Because the training data is not necessarily iid, the expectation must be with respect to the *joint* density of $\mathcal{T}$, not with respect to the *marginal* $f(\boldsymbol{x})$ (as in the iid case).

Training Loss

As the true KL risk is not known, we minimize the corresponding training loss, which is equivalent to maximizing the likelihood of the training data τ; that is, solving the optimization problem

$$\max_{w \in \mathcal{W}_p} \int g(\tau \mid \boldsymbol{\theta})\, w(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta},$$

where the maximization is over the class of densities $\mathcal{W}_p$.

Suppose that we have a rough guess, denoted $w_0(\boldsymbol{\theta})$, for the best $w \in \mathcal{W}_p$ that minimizes the Kullback–Leibler risk.

We can always increase the likelihood $L_0 := \int g(\tau \mid \boldsymbol{\theta})\, w_0(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$ by instead using the density $w_1(\boldsymbol{\theta}) := w_0(\boldsymbol{\theta})\, g(\tau \mid \boldsymbol{\theta}) / L_0$, giving a likelihood $L_1 := \int g(\tau \mid \boldsymbol{\theta})\, w_1(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$.

Increasing the Likelihood

To see this, write L_0 and L_1 as expectations with respect to w_0. In particular, we can write

$$L_0 = \mathbb{E}_{w_0}\, g(\tau \mid \boldsymbol{\theta}) \quad \text{and} \quad L_1 = \mathbb{E}_{w_1}\, g(\tau \mid \boldsymbol{\theta}) = \mathbb{E}_{w_0} g^2(\tau \mid \boldsymbol{\theta})/L_0.$$

It follows that

$$L_1 - L_0 = \frac{1}{L_0}\mathbb{E}_{w_0}\left[g^2(\tau \mid \boldsymbol{\theta}) - L_0^2\right] = \frac{1}{L_0}\mathbb{V}\mathrm{ar}_{w_0}\left[g(\tau \mid \boldsymbol{\theta})\right] \geqslant 0. \quad (2)$$

We may thus expect to obtain better predictions using w_1 instead of w_0, because w_1 has taken into account the observed data τ and increased the likelihood of the model.

If we iterate this process, we obtain a sequence of densities $w_1, w_2, \ldots$ that degenerates to the point mass at the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$.

Bayesian Learner

In many situations, the maximum likelihood estimate $g(\tau \mid \widehat{\boldsymbol{\theta}})$ is either not an appropriate approximation to $f(\tau)$ or simply fails to exist.

In such cases, given an initial non-degenerate guess $w_0(\boldsymbol{\theta}) = g(\boldsymbol{\theta})$, one can obtain a more appropriate and non-degenerate approximation to $f(\tau)$ by taking $w(\boldsymbol{\theta}) = w_1(\boldsymbol{\theta}) = g(\tau \mid \boldsymbol{\theta})\, g(\boldsymbol{\theta}) / g(\tau) = g(\theta \mid \tau)$ in (1).

This gives the following Bayesian learner of $f(\boldsymbol{x})$:

$$g_\tau(\boldsymbol{x}) := \int g(\boldsymbol{x} \mid \boldsymbol{\theta}) \frac{g(\tau \mid \boldsymbol{\theta})\, g(\boldsymbol{\theta})}{g(\tau)} \, \mathrm{d}\boldsymbol{\theta}. \tag{3}$$

Prior, Likelihood, Posterior

With this notation, we have the following definitions.

Definition: Prior, Likelihood, and Posterior

Let τ and $\mathcal{G}_p := \{g(\cdot \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be the training set and family of approximating functions.

- A pdf $g(\boldsymbol{\theta})$ that reflects our *a priori* beliefs about $\boldsymbol{\theta}$ is called the prior pdf.
- The conditional pdf $g(\tau \mid \boldsymbol{\theta})$ is called the likelihood.
- Inference about $\boldsymbol{\theta}$ is given by the posterior pdf $g(\boldsymbol{\theta} \mid \tau)$, which is proportional to the product of the prior and the likelihood:

$$g(\boldsymbol{\theta} \mid \tau) \propto g(\tau \mid \boldsymbol{\theta})\, g(\boldsymbol{\theta}).$$

Prior and Posterior Predictive Density

The initial guess $g(\boldsymbol{\theta})$ conveys the *a priori* (prior to training the Bayesian learner) information about the optimal density in $\mathcal{W}_p$ that minimizes the KL risk. Using this prior $g(\boldsymbol{\theta})$, the Bayesian approximation to $f(\boldsymbol{x})$ is the prior predictive density:

$$g(\boldsymbol{x}) = \int g(\boldsymbol{x} \mid \boldsymbol{\theta})\, g(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}.$$

The posterior pdf conveys improved knowledge about this optimal density in $\mathcal{W}_p$ after training with τ. Using the posterior $g(\boldsymbol{\theta} \mid \tau)$, the Bayesian learner of $f(\boldsymbol{x})$ is the posterior predictive density:

$$g_\tau(\boldsymbol{x}) = g(\boldsymbol{x} \mid \tau) = \int g(\boldsymbol{x} \mid \boldsymbol{\theta})\, g(\boldsymbol{\theta} \mid \tau)\, \mathrm{d}\boldsymbol{\theta},$$

where we have assumed that $g(\boldsymbol{x} \mid \boldsymbol{\theta}, \tau) = g(\boldsymbol{x} \mid \boldsymbol{\theta})$; that is, the likelihood depends on τ only through the parameter $\boldsymbol{\theta}$.

Choice of the Prior

The choice of the prior is typically governed by two considerations:

1. The prior should be simple enough to facilitate the computation or simulation of the posterior pdf.
2. The prior should be general enough to model ignorance of the parameter of interest.

Priors that do not convey much knowledge of the parameter are said to be uninformative. The uniform or flat prior is frequently used.

> For the purpose of analytical and numerical computations, we can view $\boldsymbol{\theta}$ as a random vector with prior density $g(\boldsymbol{\theta})$, which after training is updated to the posterior density $g(\boldsymbol{\theta} \mid \tau)$.

The above thinking allows us to write
$g(\boldsymbol{x} \mid \tau) \propto \int g(\boldsymbol{x} \mid \boldsymbol{\theta})\, g(\tau \mid \boldsymbol{\theta})\, g(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$, for example, thus ignoring any constants that do not depend on the argument of the densities.

Example: Normal Model

Suppose that the training data $\mathcal{T} = \{X_1, \ldots, X_n\}$ is modeled using the likelihood $g(x \mid \boldsymbol{\theta})$ that is the pdf of

$$X \mid \boldsymbol{\theta} \sim \mathcal{N}(\mu, \sigma^2),$$

where $\boldsymbol{\theta} := [\mu, \sigma^2]^\top$. For the prior distribution of $\boldsymbol{\theta}$ we can specify prior distributions for μ and σ^2 separately and then assume independence. A possible prior distribution for μ is

$$\mu \sim \mathcal{N}(\nu, \phi^2).$$

A common prior distribution for $1/\sigma^2$ is

$$\frac{1}{\sigma^2} \sim \mathsf{Gamma}(\alpha, \beta).$$

The smaller α and β are, the less informative is the prior.

Posterior Pdf

Under this prior, σ^2 is said to have an inverse gamma distribution, with pdf proportional to $\exp(-\beta/z)/z^{\alpha+1}$.

The Bayesian posterior is then:

$$g(\mu, \sigma^2 \mid \tau) \propto g(\mu) \times g(\sigma^2) \times g(\tau \mid \mu, \sigma^2)$$

$$\propto \exp\left\{-\frac{(\mu-\nu)^2}{2\phi^2}\right\} \times \frac{\exp\left\{-\beta/\sigma^2\right\}}{(\sigma^2)^{\alpha+1}} \times \frac{\exp\left\{-\sum_i (x_i-\mu)^2/(2\sigma^2)\right\}}{(\sigma^2)^{n/2}}$$

$$\propto (\sigma^2)^{-n/2-\alpha-1} \exp\left\{-\frac{(\mu-\nu)^2}{2\phi^2} - \frac{\beta}{\sigma^2} - \frac{(\mu-\bar{x}_n)^2 + S_n^2}{2\sigma^2/n}\right\},$$

where $S_n^2 := \frac{1}{n}\sum_i x_i^2 - \bar{x}_n^2 = \frac{1}{n}\sum_i (x_i - \bar{x}_n)^2$ is the (scaled) sample variance.

Posterior Mean

The conditional pdf of μ given σ^2 and τ is

$$g(\mu \mid \sigma^2, \tau) \propto \exp\left\{-\frac{(\mu-\nu)^2}{2\phi^2} - \frac{(\mu-\bar{x}_n)^2}{2\sigma^2/n}\right\},$$

which can be recognized as the pdf of

$$(\mu \mid \sigma^2, \tau) \sim \mathcal{N}\left(\gamma_n \bar{x}_n + (1-\gamma_n)\nu,\ \gamma_n\, \sigma^2/n\right), \tag{4}$$

where $\gamma_n := \frac{n}{\sigma^2} \Big/ \left(\frac{1}{\phi^2} + \frac{n}{\sigma^2}\right)$.

The posterior mean $\mathbb{E}[\mu \mid \sigma^2, \tau] = \gamma_n \bar{x}_n + (1-\gamma_n)\nu$ is a weighted linear combination of the prior mean ν and the sample average $\bar{x}_n$.

As $n \to \infty$, the weight $\gamma_n \to 1$ and thus the posterior mean approaches the maximum likelihood estimate $\bar{x}_n$.

Normal Model (cont.)

An example of an improper prior is obtained when we take prior $g(\mu) \propto 1$.

Nevertheless, the posterior is a proper density, and in particular the conditional posterior of $(\mu \,|\, \sigma^2, \tau)$ simplifies to

$$(\mu \,|\, \sigma^2, \tau) \sim \mathcal{N}\left(\overline{x}_n, \sigma^2/n\right).$$

In addition,

$$g(\sigma^2 \,|\, \tau) = \int g(\mu, \sigma^2 \,|\, \tau)\, \mathrm{d}\mu \propto (\sigma^2)^{-(n-1)/2-\alpha-1} \exp\left\{-\frac{\beta + nS_n^2/2}{\sigma^2}\right\},$$

which we recognize as the density corresponding to

$$\frac{1}{\sigma^2} \,\Big|\, \tau \sim \mathsf{Gamma}\left(\alpha + \frac{n-1}{2},\ \beta + \frac{n}{2} S_n^2\right).$$

Normal Model (cont.)

In addition to $g(\mu) \propto 1$, we can also use an improper prior $g(\sigma^2) \propto 1/\sigma^2$ for σ^2.

In this case, the posterior marginal density for σ^2 implies that:

$$\frac{nS_n^2}{\sigma^2} \Big| \tau \sim \chi^2_{n-1}$$

and the posterior marginal density for μ implies that:

$$\frac{\mu - \bar{x}_n}{S_n/\sqrt{n-1}} \Big| \tau \sim \mathrm{t}_{n-1}. \tag{5}$$

In general, deriving a simple formula for the posterior density of $\boldsymbol{\theta}$ is either impossible or too tedious. Instead, Monte Carlo methods are used to simulate (approximately) from the posterior for the purposes of inference and prediction.

Credible Interval

One way in which a distributional result such as (5) can be useful is in the construction of a 95% credible interval $\mathcal{I}$ for the parameter μ; that is, an interval $\mathcal{I}$ such that the probability $\mathbb{P}[\mu \in \mathcal{I} \mid \tau]$ is equal to 0.95.

For example, the symmetric 95% credible interval is

$$\mathcal{I} = \left[\overline{x}_n - \frac{S_n}{\sqrt{n-1}}\gamma,\ \overline{x}_n + \frac{S_n}{\sqrt{n-1}}\gamma \right],$$

with γ the 0.975-quantile of the t_{n-1} distribution (conditional on τ).

Note that the parameter μ is interpreted as a random variable with a distribution.

This is unlike the case of classical confidence intervals, where the parameter is nonrandom, but the interval is a random object.

Example

Consider analysing the number of deaths during birth in a maternity ward. Suppose that the data is given by $\tau = \{x_1, \ldots, x_n\}$, where $x_i = 1$ if the i-th baby has died during birth and $x_i = 0$ otherwise, for $i = 1, \ldots, n$.

A possible Bayesian model for the data is:

$$\begin{cases} \theta \sim \mathcal{U}(0,1) & \text{(uniform prior)}, \\ (X_1, \ldots, X_n \,|\, \theta) \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta). \end{cases}$$

The likelihood is therefore

$$g(\tau \,|\, \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} = \theta^s \, (1-\theta)^{n-s},$$

where $s = x_1 + \cdots + x_n$ is the total number of deaths.

Posterior Pdf

Since $g(\theta) = 1$, the posterior pdf is

$$g(\theta \mid \tau) \propto \theta^s \, (1-\theta)^{n-s}, \quad \theta \in [0,1],$$

which is the pdf of the $\mathsf{Beta}(s+1, n-s+1)$ distribution.

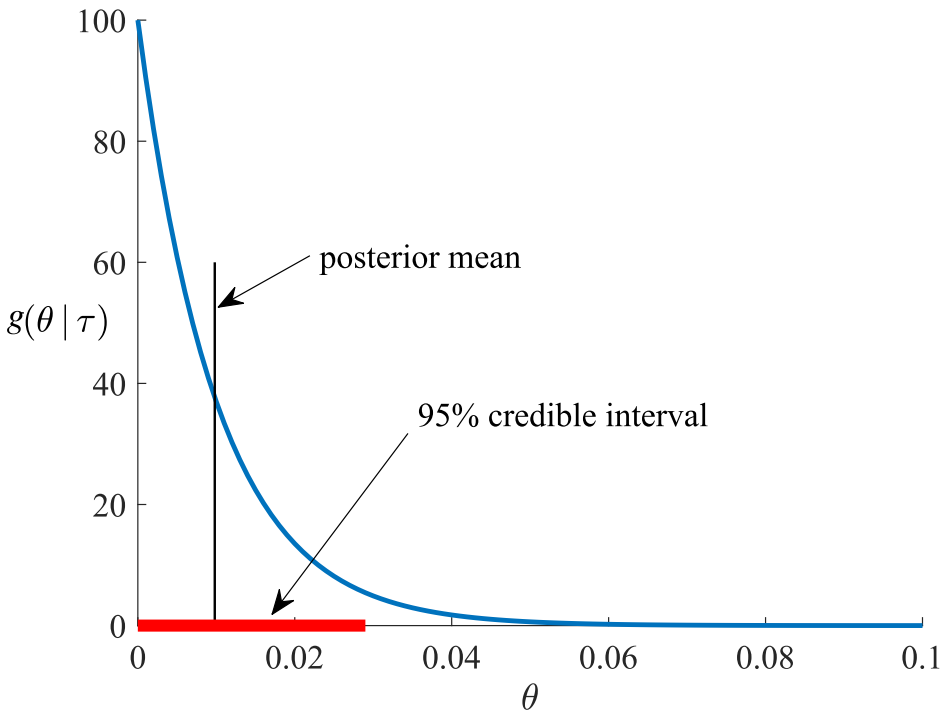


Figure: Posterior pdf for θ, with $n = 100$ and $s = 0$.

MAP Estimate and Posterior Mean

The maximum a posteriori (MAP) estimate of θ (the mode or maximizer of the posterior density) is

$$\underset{\theta}{\operatorname{argmax}}\, g(\theta \mid \tau) = \frac{s}{n},$$

which agrees with the maximum likelihood estimate.

When $(s, n) = (0, 100)$, the maximum likelihood estimate $\widehat{\theta} = 0$ infers (wrongly) that deaths at birth are not possible (same for the MAP).

In contrast the posterior mean $\mathbb{E}[\theta \mid \tau] = (s + 1)/(n + 2)$ is 0.0098 for this case and provides the more reasonable point estimate.