Statistical Learning 2

RG chapter 3

Jakob Söhl

Delft University of Technology

Part I

The Bayesian approach to statistics

Major schools of thought

Main approaches towards statistics

- Frequentist statistics
- Bayesian statistics

Frequentist statistics

- Data have distributions
- Parameters do not
- Fixed true parameters
- Distinguish parameters and statistics
- Fixed population (repeated sampling scenario)

Bayesian thinking

Bayesian statistics

- Distinction data vs. parameters is irrelevant.
  Instead: observable vs. nonobservable variables.
- Information about any variable (quantity) is incorporated by a probability distribution.
- Think generatively: make hierarchical model that specifies the probabilistic structure of all variables.
- All inference is conditional on the observed variables (data).

Example

For a group of CS students we observe whether they get a positive or negative advice after their first year of studies. Define for the j-th student

$$y_j = \begin{cases} 1 & \text{if positive advice,} \\ 0 & \text{if negative advice.} \end{cases}$$

We get similar data for other first year TU Delft programmes.

Is there reason to believe that CS students do better or worse?

Naive approach

Let i index study programme, so $i \in \mathcal{I} = \{\mathsf{CS}, \mathsf{EE}, \mathsf{AM}, \ldots\}$.

The data are y_{ij}, $i \in \mathcal{I}$, $j \in \{1, \ldots, n_i\}$.

Naive solution: compare

$$\bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \qquad i \in \mathcal{I}.$$

Note that $\bar{y}_{i\cdot}$ is the MLE if we assume

- y_{ij} is a realisation of $Y_{ij} \sim Ber\,(\theta_i)$.
- θ_i is fixed; note that we observe the full population of students.
- *All* random variables Y_{ij} are (statistically) independent.

Bayesian approach

For CS students ($i = 1$)

$$y_{1,1}, \ldots, y_{1,n_1} \mid \theta_1 \overset{\text{ind}}{\sim} Ber\left(\theta_1\right)$$
$$\theta_1 \sim p(\theta_1)$$

For EE students ($i = 2$)

$$y_{2,1}, \ldots, y_{2,n_2} \mid \theta_2 \overset{\text{ind}}{\sim} Ber\left(\theta_2\right)$$
$$\theta_2 \sim p(\theta_2)$$

Etc.

Modelling one study programme

- We model different studies separately first (connecting them will follow...)
- So assume

$$
\begin{aligned}
y_1, \ldots, y_n \mid \theta & \overset{\text{ind}}{\sim} Ber\,(\theta) \\
\theta & \sim p(\theta)
\end{aligned}
$$

- For each study there is one parameter. The distribution on θ is called the **prior** distribution.
- The joint distribution of all variables factorises:

$$
p(\boldsymbol{y}, \theta) = \underbrace{p(\boldsymbol{y} \mid \theta)}_{\text{likelihood}} \quad \times \quad \underbrace{p(\theta)}_{\text{prior}},
$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$.

The posterior distribution

$$p(\boldsymbol{y}, \theta) = \underbrace{p(\boldsymbol{y} \mid \theta)}_{\text{likelihood}} \times \underbrace{p(\theta)}_{\text{prior}}$$

- The **posterior** distribution is the distribution of all unobserved variables conditioned on the observed variables.
- **Bayes** theorem:

$$p(\theta \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \theta)p(\theta)}{p(\boldsymbol{y})},$$

  where

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y} \mid \theta)p(\theta)\,\mathrm{d}\theta$$

  is the **marginal** density of $\boldsymbol{y}$.

Notes:

1. This is just following the rules of probability theory.
    1.1 Specify the joint distribution of all variables.
    1.2 Condition using Bayes theorem (hence the name Bayesian statistics).
2. $Y_1, \ldots, Y_n$ are **conditionally** independent, this is a much weaker assumption than independent.
3. Equivalent to exchangeable:

$$p(\boldsymbol{y}) = \int p(\theta) \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \, \mathrm{d}\theta.$$

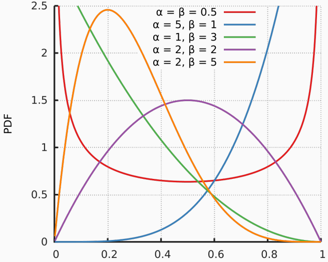
   Ordering is irrelevant.

Prior specification

Computationally convenient choice: Beta distribution.

$$p(\theta) = \frac{1}{\mathsf{B}(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\mathbf{1}_{[0,1]}(\theta),$$

where $\alpha, \beta > 0$. [1]



[1] Here, $B(\alpha, \beta) = \int \theta^{\alpha-1}(1-\theta)^{\beta-1}\mathbf{1}_{[0,1]}(\theta)\,\mathrm{d}\theta$.

Posterior computation

Let $s = \sum_{i=1}^{n} y_i$.

$$
\begin{aligned}
p(\theta \mid \boldsymbol{y}) \propto p(\boldsymbol{y}, \theta) \quad & \propto \quad \theta^{s}(1-\theta)^{n-s} \times \frac{1}{\mathsf{B}(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \mathbf{1}_{[0,1]}(\theta) \\
& \propto \quad \theta^{s+\alpha-1}(1-\theta)^{n-s+\beta-1} \frac{1}{\mathsf{B}(\alpha,\beta)} \mathbf{1}_{[0,1]}(\theta) \\
& \propto \quad \theta^{s+\alpha-1}(1-\theta)^{n-s+\beta-1} \mathbf{1}_{[0,1]}(\theta).
\end{aligned}
$$

As $p(\theta \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y},\theta)}{\int p(\boldsymbol{y},\theta)\,\mathrm{d}\theta}$ we obtain a Beta distribution

$$
\theta \mid \boldsymbol{y} \sim Be\left(s+\alpha, n-s+\beta\right).
$$

Combining data from multiple study programmes

For CS students ($i = 1$)

$$y_{1,1}, \ldots, y_{1,n_1} \mid \theta_1 \overset{\text{ind}}{\sim} Ber\,(\theta_1)$$
$$\theta_1 \sim p(\theta_1)$$

For EE students ($i = 2$)

$$y_{2,1}, \ldots, y_{2,n_2} \mid \theta_2 \overset{\text{ind}}{\sim} Ber\,(\theta_2)$$
$$\theta_2 \sim p(\theta_2)$$

etc.

Replace with

$$y_{ij} \mid \theta_i \overset{\text{ind}}{\sim} Ber\,(\theta_i) \quad 1 \leq i \leq |\mathcal{I}|,\, 1 \leq j \leq n_i$$
$$\theta_1, \theta_2, \ldots, \theta_{|\mathcal{I}|} \mid (\alpha, \beta) \overset{\text{iid}}{\sim} Be\,(\alpha, \beta)$$
$$\alpha, \beta \overset{\text{iid}}{\sim} Exp\,(1/2)$$

Including covariates

Before claiming a particular i to do bad teaching, we may wish to include information that takes variation of students into account.

For student (i, j), let x_{ij} be her/his score on math at high-school.

Logistic regression idea: model

$$
\begin{aligned}
y_{ij} \mid \theta_{ij} &\overset{\text{ind}}{\sim} Ber\left(\theta_{ij}\right) \quad 1 \le i \le |\mathcal{I}|,\ 1 \le j \le n_i \\
\log\left(\frac{\theta_{ij}}{1-\theta_{ij}}\right) &= \alpha_i + \beta_i x_{ij} \\
\begin{bmatrix}\alpha_1 \\ \beta_1\end{bmatrix}, \ldots, \begin{bmatrix}\alpha_{|\mathcal{I}|} \\ \beta_{|\mathcal{I}|}\end{bmatrix} &\overset{\text{iid}}{\sim} N(0, \Sigma)
\end{aligned}
$$

Lot's of parameters! **Shrinkage** is key.

"Missing data"

Strange, yet common, term.

Refers to the situation where covariates (x_{ij}), or response (y_{ij}) is not registered.

This is simply an unobserved variable.

Intermediate summary

- Generative thinking.
- Follow the rules of probability theory (Bayes theorem).
- Distinction between "fixed parameters"-"data with a distribution" not relevant.
- Very flexible.
- Can easily include huge number of parameters (shrinkage does the job).
- Computational problem can be daunting.
  1. Conjugate priors can be handled easy.
  2. Otherwise: MCMC, SMC, etc.

Part II

Empirical Bayes

Empirical Bayes

- A bit hidden in the book is the idea of **empirical Bayes** (misleading name).
- This is a way to determine hyperparameters based on the data. (So this is not a Bayesian procedure!)
- Consider the model

$$\begin{aligned} X \mid \Theta = \theta &\sim f_{X\mid\Theta}(\cdot \mid \theta) \\ \Theta &\sim f_\Theta(\theta;\eta), \end{aligned}$$

  where η is the hyperparameter.
- Empirical Bayes: estimate η from f_X.

- Common method for estimating η: "type II Maximum Likelihood"

$$\hat{\eta} = \operatorname*{argmax}_{\eta} f_X(x; \eta). \tag{1}$$

- The "posterior" obtained by the empirical Bayes method is the "ordinary" posterior, with $\hat{\eta}$ substituted for η.

Empirical Bayes: exercise

Assume $X_1, \ldots, X_p$ are independent conditionals on $\Theta_1, \ldots, \Theta_p$. Suppose $X_i \mid \Theta_i = \theta_i \sim Unif\,(0, \theta_i)$. Consider estimation of the parameters $\Theta_1, \ldots, \Theta_p$ based on data $X_1, \ldots, X_p$.

(a) Model $\Theta_1, \ldots, \Theta_p$ as independent with common density

$$f_\Theta(\theta) = \theta\lambda^2 e^{-\lambda\theta}\mathbf{1}_{[0,\infty)}(\theta).$$

Find the posterior mean for Θ_i $(1 \le i \le p)$.

(b) Determine λ by marginal maximum likelihood.

Part III

Bayesian analysis of the linear model

Linear model in the Bayesian setup

For simplicity, assume σ^2 (measurement variance) is known.

$$
\begin{aligned}
y \mid \theta &\sim N_n(X\theta, \sigma^2 I) \\
\theta &\sim N_p(\mu_0, \Sigma_0)
\end{aligned}
$$

The prior induces conjugacy.

$$
\begin{aligned}
p(\theta \mid y, X) \quad &\propto \quad p(y, \theta \mid X) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\|y - X\theta\|^2\right) \times \\
&\qquad (2\pi)^{-p/2} |\det \Sigma_0|^{-1/2} \exp\left(-\frac{1}{2}(\theta - \mu_0)^T \Sigma_0^{-1} (\theta - \mu_0)\right) \\
&\propto \quad \exp\left\{-\frac{1}{2}\theta^T \left(\frac{X^T X}{\sigma^2} + \Sigma_0^{-1}\right)\theta + \theta^T \left(\frac{X^T y}{\sigma^2} + \Sigma_0^{-1}\mu_0\right)\right\}
\end{aligned}
$$

This implies

$$\theta \mid y, X \sim N_p^{\mathrm{can}}\left(\frac{X^T y}{\sigma^2} + \Sigma_0^{-1}\mu_0, \frac{X^T X}{\sigma^2} + \Sigma_0^{-1}\right).$$

In other words, the posterior precision equals

$$P_{\mathrm{post}} = \frac{X^T X}{\sigma^2} + \Sigma_0^{-1}$$

and the posterior mean equals

$$\theta_{\mathrm{post}} = P_{\mathrm{post}}^{-1}\left(\frac{X^T y}{\sigma^2} + \Sigma_0^{-1}\mu_0\right).$$

Special case

Suppose $\Sigma_0 = \sigma_0^2 I$ and $\mu_0 = 0$. Then

$$\theta_{\mathrm{post}} = \left(X^T X + \frac{\sigma^2}{\sigma_0^2} I\right)^{-1} X^T y.$$

1. Viewed as frequentist estimator, θ_{post} is not unbiased for θ.
2. Well defined even if X does not have full column rank.
3. Example of **regularisation**.
4. **Shrinkage** towards the prior mean, depending on $\lambda = \sigma^2/\sigma_0^2$.
5. In frequentist statistics known as **ridge regression**.

Bayesian prediction in the linear model

Use the rules of probability theory!

$$
\begin{aligned}
p(y_{\text{new}} \mid y, X, x_{\text{new}}) &= \int p(y_{\text{new}}, \theta \mid y, X, x_{\text{new}}) \, \mathrm{d}\theta \\
&= \int p(y_{\text{new}} \mid \theta, x_{\text{new}}) p(\theta \mid y, X) \, \mathrm{d}\theta
\end{aligned}
$$

So we average over the posterior uncertainty.

Homework:

$$
y_{\text{new}} \mid y, X, x_{\text{new}} \sim N\left(x_{\text{new}}^T \theta_{\text{post}}, x_{\text{new}}^T P_{\text{post}}^{-1} x_{\text{new}} + \sigma^2\right)
$$

(Cf. RG section 3.8.)

Predictive covariance: Frequentist and Bayes compared

- **Frequentist.** Recall that the covariance of $\hat{\theta}^T x_{\text{new}}$ (with $\hat{\theta}$ the MLE) is given by
$$\sigma^2 x_{\text{new}}^T (X^T X)^{-1} x_{\text{new}}$$
So $y_{\text{new}} = \hat{\theta}^T x_{\text{new}} + \varepsilon_{\text{new}}$ has covariance matrix
$$V_{\text{freq}} = \sigma^2 x_{\text{new}}^T (X^T X)^{-1} x_{\text{new}} + \sigma^2$$

- **Bayesian.**
$$V_{\text{Bayes}} = x_{\text{new}}^T P_{\text{post}}^{-1} x_{\text{new}} + \sigma^2$$
with
$$P_{\text{post}} = \sigma^{-2} X^T X + \Sigma_0^{-1}$$
Assume $\Sigma_0 = \tau I$ and suppose $\tau \to \infty$. Then
$$V_{\text{Bayes}} \to \sigma^2 x_{\text{new}}^T (X^T X)^{-1} x_{\text{new}} + \sigma^2 = V_{\text{freq}}.$$

Assignment 1: Bayesian updating in linear models and Kalman filtering

To make the notation easier, any dependence on x is dropped in the following derivation.

Bayesian updating: let $\boldsymbol{y}_n = (y_1, \ldots, y_n)$. Then

$$p(\theta \mid \boldsymbol{y}_n, y_{n+1}) \propto p(y_{n+1} \mid \theta) p(\theta \mid \boldsymbol{y}_n),$$

provided that $p(y_{n+1} \mid \theta, \boldsymbol{y}_n) = p(y_{n+1} \mid \theta)$.

This partly explains the huge popularity of the Bayesian approach in signal processing.

Rethinking: what about the marginal distribution of X?

In regression we model the conditional distribution of y_i (conditional on x_i). Why no distribution on x?

- Suppose

$$p(x, y \mid \theta, \psi) = p(y \mid x, \theta)p(x \mid \psi).$$

- Then

$$p(\theta, \psi \mid x, y) \propto p(y \mid x, \theta)p(x \mid \psi)p(\theta, \psi).$$

- **Key point**: if we assume $p(\theta, \psi) = p(\theta)p(\psi)$, then

$$p(\theta \mid x, y) \propto p(\theta)p(y \mid x, \theta).$$

For inferring θ it suffices to model the conditional distribution!