BML: exercise sheet

Rémi Bardenet

Stars indicate the difficulty level, from 1 to 3. One star means that everyone should be able to do it without too much effort.

Contents

1 Lecture #1: Bayesics

1.1 Conjugate priors 101: Gaussians ($\star$)

Let $y|\mu \sim \mathcal{N}(\mu, I_N)$ and $\mu \sim \mathcal{N}(0, aI_N)$, for some $a > 0$. Show that

$$\mu|y \sim \mathcal{N}(by, bI_N), \text{ where } b = a/(a+1). \tag{1}$$

Solution: We apply Bayes' theorem and keep track of only the terms that will not end up in the normalization constant of the posterior. This gives

$$\begin{aligned}
\log p(\mu|y) &\propto \log p(y|\mu) + \log p(\mu) \\
&\propto -\frac{\|y-\mu\|^2}{2} - \frac{\|\mu\|^2}{2a} \\
&\propto -\frac{1}{2}\|\mu\|^2\left(1+\frac{1}{a}\right) + y^T\mu \\
&\propto -\frac{\|\mu - by\|^2}{2b}.
\end{aligned}$$

1.2 A conjugate prior on probability vectors ($\star$)

Let

$$\Delta_d = \{\theta \in [0,1]^d \text{ such that } \sum_{k=1}^d \theta_d = 1\}.$$

Let further $\alpha \in (\mathbb{R}_+)^d$. The Dirichlet pdf is defined by

$$\mathrm{Dir}(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^d \theta_k^{\alpha_k - 1} 1_{\theta \in \Delta_d},$$

where $B(\alpha) = \prod_{k=1}^d \Gamma(\alpha_k)/\Gamma(\sum_{k=1}^d \alpha_k)$ is the so-called beta function.

Now put a prior $\mathrm{Dir}(\theta|\alpha)$ on θ, and consider drawing $y_{1:N}$ from the multinomial distribution with parameter $\theta \in \Delta_d$. Show that

$$p(\theta, y_{1:N}) = \frac{B(\alpha + c)}{B(\alpha)} \mathrm{Dir}(\theta|\alpha + c), \tag{2}$$

where $c = (\sum_{i=1}^N 1_{y_i = k})_{1 \leqslant k \leqslant d}$ is the vector of counts. Note that (2) implies that $\theta | y_{1:N} \sim \mathrm{Dir}(\theta|\alpha)$ and that the marginal likelihood $p(y_{1:n}) = B(\alpha)/B(\alpha + c)$.

Solution: Once you express the multinomial pdf, the Dirichlet distribution becomes the obvious conjugate prior. This time, we keep track of the normalizing constant, because the script requires it. This gives

$$\begin{aligned} p(\theta, y_{1:N}) &= p(y_{1:N}|\theta) p(\theta) \\ &= \prod_{i=1}^N \prod_{k=1}^d \theta_k^{1\{y_i = k\}} \times \frac{1}{B(\alpha)} \prod_{k=1}^d \theta_k^{\alpha_k - 1} 1_{\theta \in \Delta_d} \\ &= \frac{1}{B(\alpha)} \prod_{k=1}^d \theta_k^{\alpha_k + c_k - 1} 1_{\theta \in \Delta_d} \\ &= \frac{B(\alpha + c)}{B(\alpha)} \mathrm{Dir}(\theta|\alpha + c). \end{aligned}$$

1.3 Empirical Bayes and the James-Stein effect ($\star\star$)

Let $\mu = (\mu_1, \dots, \mu_N) \in \mathbb{R}^N$, and consider N i.i.d. real variables $y_i | \mu \sim \mathcal{N}(\mu_i, 1)$. We wish to infer μ.

1. What is the maximum likelihood estimator $\hat{\mu}_{\mathrm{MLE}}$?

2. Henceforth, we judge estimators by the square loss. The frequentist risk of an estimator $\hat{\mu}$ is
$$R(\hat{\mu}) = \mathbb{E}_{y|\mu} \|\mu - \hat{\mu}\|^2.$$
show that $R(\hat{\mu}_{\mathrm{MLE}}) = N$.

3. Suppose we have prior belief that μ lies near 0, and we choose to represent it by $\mu \sim \mathcal{N}(0, aI_N)$, $a > 0$. What is the Bayes estimator $\hat{\mu}_{\text{Bayes}}$? What is its (frequentist) risk $R(\hat{\mu}_{\text{Bayes}})$? What is its Bayes risk $\mathbb{E}_\mu R(\hat{\mu}_{\text{Bayes}})$?

4. Since we actually have no idea what a should be, we propose to estimate it from data[1] Show that the marginal of y is

$$\int p(y, \mu)\mathrm{d}\mu = \mathcal{N}(y|0, (a+1)I_N).$$

   In particular, what is the law of $S = \|y\|^2$? Deduce from it that $(N-2)/S$ is an unbiased estimator of $a+1$, and consider the empirical Bayes estimator

$$\hat{\mu}_{\text{EB}} = \left(1 - \frac{N-2}{S}\right) y.$$

   What is its Bayes risk?

5. *Note: This particular item is* $(\star\star\star)$ *because it is longer to solve, but all individual arguments are elementary; do this only if you have solved all the preceding exercises, though. Also, see* **Efr10** *for a solution)* Show that for $N \geqslant 3$, for every $\mu \in \mathbb{R}^N$,

$$R(\hat{\mu}_{\text{EB}}) < R(\hat{\mu}_{\text{MLE}}). \tag{3}$$

   Frequentists say that $\hat{\mu}_{\text{EB}}$ dominates μ_{MLE}, in the sense that whatever the value of μ, the risk of $\hat{\mu}_{\text{EB}}$ is the smallest of the two. This happens even when μ is far from zero, in which case one might have thought that our $\mathcal{N}(0, aI_N)$ prior would have been a poor choice. Finally, if you are a strict Waldian, you should thus prefer $\hat{\mu}_{\text{EB}}$ to $\hat{\mu}_{\text{MLE}}$. Many applied frequentists still use $\hat{\mu}_{\text{MLE}}$, however; see (**Efr10**) for a tentative answer.

   Equation 3 is called the James-Stein effect, and is a standard example of why following Bayesian guidelines can end up giving good frequentist estimators. Shrinkage, like $\hat{\mu}_{\text{EB}}$ shrinks $\hat{\mu}_{\text{MLE}}$ towards zero, is now commonplace in large-dimensional regression. For more on frequentist guarantees for Bayesian estimators and shrinkage, see (**PaIn09**).

Solution: The solution is basically **Efr10**. The book is also highly recommended, especially if you are into large-scale hypothesis tests. At least, read the prologue for statistical culture.

[1] This procedure of using data to tune the prior is called *empirical Bayes* (EB). The expected utility principle allows it, but statisticians who like to interpret their prior as encoding their belief before the data is collected are uncomfortable with EB. At the other extreme, Bayesians who insist on using estimators with good frequentist properties are happy using the data or the likelihood to design their prior.

1.4 Classification with asymmetric loss ($\star$)

Consider the classification problem, but with loss

$$L(a_g, s) = \alpha 1_{y \neq g(x; x_{1:n}, y_{1:n})} 1_{y=0} + \beta 1_{y \neq g(x; x_{1:n}, y_{1:n})} 1_{y=1},$$

for some $\alpha, \beta > 0$. Show that the Bayes decision rule is

$$g^\star(x; x_{1:n}, y_{1:n}) = 1_{p(y|x, x_{1:n}, y_{1:n}) \geqslant \frac{\alpha}{\alpha+\beta}}.$$

In particular, if $\alpha \ll \beta$, one will often decide for predicting 1, because the cost for misclassifying a 0 is low.

Solution: For brevity, we drop the dependence of g in the training set and write $g(x)$ for $g(x; x_{1:n}, y_{1:n})$. Following the posterior expected loss rationale, we pick action

$$\begin{aligned} a^\star = a_{g^\star} &\in \arg\min \int L(a_g, s) p(s_u | s_o) \mathrm{d}s_u \\ &= \arg\min \int \left[\alpha 1_{y \neq g(x)} 1_{y=0} + \beta 1_{y \neq g(x)} 1_{y=1}\right] p(y | x_{1:N}, y_{1:N}, x) \mathrm{d}y \\ &= \arg\min \alpha 1_{0 \neq g(x)} p(y = 0 | x_{1:N}, y_{1:N}, x) \\ &\qquad\qquad + \beta 1_{1 \neq g(x)} p(y = 1 | x_{1:N}, y_{1:N}, x). \end{aligned}$$

This is equivalent to setting $g^\star(x) = 1$ if and only if

$$\alpha p(y = 0 | x_{1:N}, y_{1:N}, x) \leqslant \beta p(y = 1 | x_{1:N}, y_{1:N}, x).$$

Letting $q = p(y = 1 | x_{1:N}, y_{1:N}, x)$, this becomes

$$\alpha(1 - q) \leqslant \beta q,$$

or, equivalently,

$$q \geqslant \alpha / (\alpha + \beta).$$

1.5 Linear regression with a Gaussian prior ($\star$)

Consider $y_i | x_i, \theta \sim \mathcal{N}(x_i^T \theta, \sigma^2)$ i.i.d., $i = 1, \ldots, N$. Take a Gaussian prior $\theta \sim \mathcal{N}(0, \sigma_0^2)$. Show that the posterior $\theta | x_{1:N}, y_{1:N}$ is Gaussian, with mean the ridge regression estimator.

Solution: We write Bayes' theorem and keep track only of the terms that won't end up in the normalization constant. This gives

$$\begin{aligned}
\log p(\theta|y_{1:N}, x_{1:N}) &\propto \log p(y_{1:N}|x_{1:N}, \theta) + \log p(\theta) \\
&\propto -\sum_{i=1}^{N} \frac{(y_i - x_i^T\theta)^2}{2\sigma^2} + \frac{1}{2\sigma_0^2}\|\theta\|^2 \\
&= -\frac{1}{2\sigma^2}\|y - X\theta\|^2 + \frac{1}{2\sigma_0^2}\|\theta\|^2 \\
&\propto -\frac{1}{2\sigma^2}\left[\theta^T\left(X^TX + \frac{\sigma^2}{\sigma_0^2}I_d\right)\theta - 2y^TX\theta\right] \\
&= -\frac{1}{2}\left[\theta^T\Lambda\theta - \frac{2}{\sigma^2}y^TX\theta\right],
\end{aligned}$$

where $\Lambda := \frac{1}{\sigma^2}X^TX + \frac{1}{\sigma_0^2}I_d$ is symmetric and positive definite. This leads to

$$\log p(\theta|y_{1:N}, x_{1:N}) \propto -\frac{1}{2}\left(\theta - \frac{1}{\sigma^2}\Lambda^{-1}X^Ty\right)^T \Lambda \left(\theta - \frac{1}{\sigma^2}\Lambda^{-1}X^Ty\right),$$

so that $\theta|y_{1:N}, x_{1:N}$ is indeed Gaussian, with mean the ridge regression estimator

$$\frac{1}{\sigma^2}\Lambda^{-1}X^Ty = \left(X^TX + \frac{\sigma^2}{\sigma_0^2}I_d\right)^{-1} X^Ty$$

and variance Λ^{-1}. Note how the ratio σ/σ_0 is playing the role of the regularization parameter in ridge regression.

1.6 For more exercises on Bayesian derivations

- Exercises 5.1 to 5.4 of (**Mur12**).
- Go through Sections 4.4 to 4.6 of (**Mur12**) with pen and paper. Linear Gaussian models appear all the time.
- Exercises 2.6, 2.9, 2.10, 2.13, 2.14, and 2.15 of (**MaRo07**). Solutions are here.