BML: exercise sheet

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

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

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

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

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

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

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

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 $\theta|y_{1:N} \sim \text{Dir}(\theta|\alpha)$ and that the marginal likelihood $p(y_{1:n}) = B(\alpha)/B(\alpha + c)$.

1.3 Estimating the mean of a Gaussian ($\star\star$)

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 $\hat{\mu}_{\mathrm{Bayes}}$? What is its (frequentist) risk $R(\hat{\mu}_{\mathrm{Bayes}})$? What is its Bayes risk $\mathbb{E}_\mu R(\hat{\mu}_{\mathrm{Bayes}})$?

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

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

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 Efron, 2012, Section 1.2 for a solution)* Show that for $N \geqslant 3$, for every $\mu \in \mathbb{R}^N$,

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

Frequentists say that $\hat{\mu}_{\mathrm{EB}}$ dominates μ_{MLE}, in the sense that whatever the value of μ, the risk of $\hat{\mu}_{\mathrm{EB}}$ 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 $\hat{\mu}_{\mathrm{EB}}$ to $\hat{\mu}_{\mathrm{MLE}}$. Many applied frequentists still use $\hat{\mu}_{\mathrm{MLE}}$, however; see (Efron, 2012, Section 1.3) 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 $\hat{\mu}_{\mathrm{EB}}$ shrinks $\hat{\mu}_{\mathrm{MLE}}$ towards zero, is now commonplace in large-dimensional regression. For more on frequentist guarantees for Bayesian estimators and shrinkage, see (Parmigiani and Inoue, 2009, Sections 7, 8, 9).

---

[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.
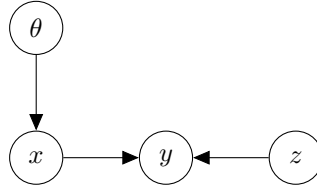
Figure 1: A DAG

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.

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

It is all in the title: show that the posterior is Gaussian, with mean the ridge regression estimator.

1.6 For more exercises on Bayesian derivations

- Exercises 5.1 to 5.4 of (Murphy, 2012).

- Go through Sections 4.4 to 4.6 of (Murphy, 2012) 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 (Marin and Robert, 2007). Solutions are here.

2 Lecture #2: MCMC

2.1 DAGs and dependence ($\star$)

Consider the DAG from Figure 1.

1. Write the corresponding factorization of $p(x, y, z, \theta)$.

2. Deduce from the factorization that $x \perp z$.

3. Deduce from the factorization that $x \perp z | \theta$.

4. Deduce from the factorization that $x \not\perp z | \theta, y$.

In particular, note how Item 3 is a case of *being independent from your non-descendents given your parents*, while Item 4 illustrates how conditioning on common children can induce dependence between parents. In more complicated DAGs, the so-called *Bayes ball* algorithm determines whether two sets of nodes are independent given a third one; see Murphy, 2012, Section 10.5.

2.2 Self-normalized importance sampling ($\star\star$)

Show a central limit theorem for the self-normalized importance sampling estimator. Hint: use the delta method.

2.3 The random scan Gibbs sampler always accepts

Consider the MH kernel with proposal

$$q(\theta'|\theta) = \frac{1}{d} \sum_{k=1}^{d} \pi(\theta_k|\theta_{\backslash k}), \quad \theta_{\backslash k} := (\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d).$$

Show that the MH acceptance probability $\alpha(\theta, \theta')$ is 1. When implementing a Gibbs sampler, it is thus enough to repeatedly draw from a conditional chosen uniformly at random.

2.4 Systematic scan Gibbs sampler ($\star\star$)

Show that the systematic scan Gibbs kernel, while not satisfying detailed balance, leaves π invariant.

2.5 Gibbs ($\star$) and collapsed ($\star\star$) Gibbs for LDA

Rederive all conditionals in the LDA and collapsed LDA model. *Hint: use (2); Check (Murphy, 2012, Section 27.3.4) for the solution.*

3 Lecture #3

3.1 VB 101 ($\star$)

Derive the coordinatewise updates for mean field VB to fit the mean and variance of a univariate Gaussian. *The solution is in (**Mu12**).*

3.2 A useful lemma for variational LDA ($\star$)

Let $\Psi(\cdot) := \Gamma'(\cdot)/\Gamma(\cdot)$ be the digamma function. Show that

$$\mathbb{E}_{\mathrm{Dir}(\theta|\alpha)} \log \theta_i = \Psi(\theta_i) - \Psi(\|\theta\|_1).$$

We used that lemma when deriving the coordinatewise updates for VB with mean field approximation.

3.3 VB for LDA with counts ($\star\star$)

Derive the coordinatewise updates for VB on the count version of LDA. The variational approximation should read

$$q(\pi_i, c_i, B) = \text{Dir}(\pi_i | \tilde{\pi}_i) \prod_v \text{Multinomial}(c_{iv\cdot} | n_{iv}, \tilde{c}_{iv\cdot}) \prod_k \text{Dir}(b_{\cdot k} | \tilde{b}_{\cdot k}).$$

Hint: See Murphy, 2012, Section 27.3.6.

4 Lecture #4

References

[1] B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction.* Vol. 1. Cambridge University Press, 2012.

[2] J.-M. Marin and C.P. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics.* New York: Springer-Verlag, 2007.

[3] K. Murphy. *Machine learning: a probabilistic perspective.* MIT Press, 2012.

[4] G. Parmigiani and L. Inoue. *Decision theory: principles and approaches.* Vol. 812. John Wiley & Sons, 2009.