BML: exercise sheet

Rémi Bardenet and Julyan Arbel

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 **2**
1.1 Conjugate priors 101: Gaussians ($\star$) 2
1.2 A conjugate prior on probability vectors ($\star$) 2
1.3 Empirical Bayes and the James-Stein effect ($\star\star$) 3
1.4 Classification with asymmetric loss ($\star$) 4
1.5 Linear regression with a Gaussian prior ($\star$) 5
1.6 For more exercises on Bayesian derivations 6

2 Lecture #2: MCMC **6**
2.1 DAGs and dependence ($\star$) 6
2.2 Self-normalized importance sampling ($\star\star$) 7
2.3 The random scan Gibbs sampler always accepts ($\star$) 8
2.4 Systematic scan Gibbs sampler ($\star\star$) 8
2.5 Gibbs ($\star$) and collapsed ($\star\star$) Gibbs for LDA 8

3 Lecture #3: Variational inference **8**
3.1 VB 101: fitting a univariate Gaussian ($\star$) 8
3.2 A useful lemma for variational LDA ($\star$) 8
3.3 VB for LDA with counts ($\star\star$) 9

4 Lecture #4: Bayesian nonparametrics **9**
4.1 Combinatorial properties of K_n for Dirichlet process ($\star$) 9
4.2 Combinatorial properties of K_n for Pitman–Yor process ($\star\star$) . . 9
4.3 For more exercises on Bayesian nonparametrics 10

5 Lecture #5: Foundations **10**
5.1 A simple application of the likelihood principle ($\star$) 10
5.2 The Blackwell-McQueen urn scheme and exchangeability ($\star\star$) . . 10
5.3 McAllester's PAC bound ($\star\star\star$) 11

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 $\mu = (\mu_1, \ldots, \mu_N) \in \mathbb{R}^N$, 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?

[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.

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 $\mu \in \mathbb{R}^N$,

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

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 (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 $\hat{\mu}_{\text{EB}}$ 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 (Parmigiani and Inoue, 2009, Sections 7, 8, 9).

Solution: The solution is basically Efron, 2012, Section 1.2. The book is also highly recommended, especially if you are into large-scale hypothesis tests. At least, read the prologue for statistical culture.

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

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

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

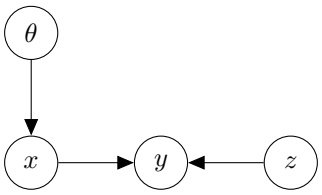

Figure 1: A DAG

to

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

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

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

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 (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 (⋆)

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.

Solution:

1. By definition, we write the product of the conditionals of each node given its parents, that is,

$$p(x, y, z, \theta) = p(y|z, x)p(x|\theta)p(\theta)p(z). \tag{4}$$

2. By (4),

$$p(x, z) = \int p(x, y, z, \theta)\mathrm{d}y\mathrm{d}\theta = p(z) \int p(x|\theta)p(\theta)\mathrm{d}\theta.$$

   In particular,

$$p(x) = \int p(x, z)\mathrm{d}z = \int p(x|\theta)p(\theta)\mathrm{d}\theta,$$

   so that $p(x, z) = p(x)p(z)$.

3. We use Bayes' theorem and (4),

$$\begin{aligned} p(x, z|\theta) &= \int p(x, y, z|\theta)\mathrm{d}y \\ &= \int \frac{p(x, y, z, \theta)}{p(\theta)}\mathrm{d}y \\ &= \int p(y|z, x)p(x|\theta)p(z)\mathrm{d}y \\ &= p(x|\theta)p(z). \end{aligned}$$

   In particular,

$$p(z|\theta) = \int p(x, z|\theta)\mathrm{d}x = p(z),$$

   so that $p(x, z|\theta) = p(x|\theta)p(z|\theta)$.

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 ($\star$)

Consider the MH kernel with proposal

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

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: Variational inference

3.1 VB 101: fitting a univariate Gaussian ($\star$)

Consider a univariate Gaussian model $y|\mu, \lambda \sim \mathcal{N}(\mu, \lambda^{-1})$, where $\lambda = 1/\sigma^2$ is called the precision parameter.

1. Take as prior

   $$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\kappa_0\lambda)^{-1})\text{Gamma}(\lambda|\alpha_0, \beta_0).$$

   What is the posterior? *Hint: the prior is conjugate.*

2. Derive the updates for mean field VB in this model, i.e., with approximation

   $$q(\mu, \lambda) = q(\mu)q(\lambda).$$

3. Since we know the actual posterior, what can you say of the mean field solution in that case? Could you extend VB to nonconjugate priors?

The solution is in (Murphy, 2012, Section 21.5.1).

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

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

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

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

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

Hint: See Murphy, 2012, Section 27.3.6.

4 Lecture #4: Bayesian nonparametrics

4.1 Combinatorial properties of K_n for Dirichlet process ($\star$)

Let K_n be the number of clusters observed when drawing n observations from a Dirichlet process with concentration parameter $\alpha \in \mathbb{R}_+$.

1. Show that

$$\mathbb{E}[K_n] = \sum_{i=0}^{n-1} \frac{\alpha}{\alpha + i} \quad \text{and} \quad \mathrm{Var}(K_n) = \sum_{i=0}^{n-1} \frac{\alpha i}{(\alpha + i)^2}.$$

2. Show the following large n asymptotics for the expectation and variance of K_n:

$$\mathbb{E}[K_n] \sim \alpha \log n \quad \text{and} \quad \mathrm{Var}(K_n) \sim \alpha \log n.$$

4.2 Combinatorial properties of K_n for Pitman–Yor process ($\star\star$)

Let K_n be the number of clusters observed when drawing n observations from a Pitman–Yor process with discount parameter $\sigma \in (0, 1)$ and concentration parameter $\alpha \in \mathbb{R}_+$.

1. Show that

$$\mathbb{E}[K_{n+1}] = \frac{\alpha}{n + \alpha} + \frac{\sigma + \alpha + n}{n + \alpha} \mathbb{E}\left[K_n\right].$$

   Hint: use the PY predictive distribution and a conditional expectation to get this iterative formula from n to $n + 1$.

2. Deduce that

$$\mathbb{E}[K_n] = \sum_{i=0}^{n-1} \frac{(\alpha + \sigma)_i}{(\alpha + 1)_i},$$

   where $(x)_n = x(x + 1) \dots (x + n - 1)$.

3. Show the following large n asymptotics for the expectation of K_n:

$$\mathbb{E}[K_n] \sim \frac{\Gamma(\alpha+1)}{\sigma\Gamma(\alpha+\sigma)} n^\sigma.$$

4. Show that the following recursive formula holds for the variance of K_n:

$$\mathrm{Var}(K_{n+1}) = \mathrm{Var}(K_n)\frac{n+\alpha+2\sigma}{n+\alpha} + \frac{(\sigma\mathbb{E}[K_n]+\alpha)(n-\sigma\mathbb{E}[K_n])}{(n+\alpha)^2}. \quad (5)$$

Hint: use the law of total variance.

5. Derive again the simpler expression of expectation and variance of K_n in the Dirichlet process case (by setting $\sigma = 0$ in the above formulas).

4.3 For more exercises on Bayesian nonparametrics

- Exercises 4.4, 4.8, 4.9, 4.10, 4.12, 4.13, 4.18, 4.24, 4.25, 4.26, 4.32, 4.39 of Ghosal and Van der Vaart, 2017.

5 Lecture #5: Foundations

5.1 A simple application of the likelihood principle ($\star$)

Consider experiments E_1: tossing a coin 10 times, vs. E_2: tossing the same coin until obtaining 4 heads. Say we ran E_1 and E_2, and we obtained two samples of the same size $n = 10$.

1. Write down the binomial and negative binomial likelihoods corresponding to E_1 and E_2, respectively.
2. Build two credible intervals for the bias θ of the coin, one for each experiment. Are the two intervals the same?
3. Build two (frequentist) confidence intervals for the bias θ of the coin, one for each experiment. Are the two intervals the same?
4. Which answer bothers you the most?

5.2 The Blackwell-McQueen urn scheme and exchangeability ($\star\star$)

1. Show that the colors $X_1, \ldots$ drawn in the BMC urn scheme are exchangeable.
2. Prove that the corresponding measure on $\mathcal{P}(\mathcal{X})$ given by de Finetti's theorem is a Dirichlet process.

5.3 McAllester's PAC bound ($\star\star\star$)

Prove McAllester's PAC bound. Hint: Check out Chapter 31 of (Shalev-Shwartz and Ben-David, 2014).

References

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

[2] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press, 2017.

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

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

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

[6] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.