Probabilistic Machine Learning

Exercise Sheet #9

Regression on Multinomial Data with the Laplace Bridge

This exercise sheet introduces a few concepts that will help you implement the large-scale example application we will go through over the coming weeks. Questions 2 and 3, together, develop a custom Laplace approximation for Dirichlet distributions that will be used in the final exercise sheet to build a regression model on multivariate count observations.

1. **Exam-Type Question** Recall that the pdf of the multivariate Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and symmetric positive definite covariance $\Sigma \in \mathbb{R}^{d \times d}$ is given by

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\intercal \Sigma^{-1}(x-\mu)\right). \tag{1}$$

   (a) Show that Eq. (1) can be phrased as an exponential family $\tilde{\mathcal{N}}(x; P, M)$ with the natural parameters $(P, M) := (0.5\Sigma^{-1}, -\Sigma^{-1}\mu)$ (sometimes called *precision* and *precision-adjusted mean*, respectively). *Hint:* find an alternative representation of $\operatorname{tr}(AB)$ for matrices A, B in terms of vectorised representation of A and B.

   (b) Give an explicit expression (in terms of the natural parameters) for the product and quotient

$$\tilde{\mathcal{N}}(x; P_1, M_1) \cdot \tilde{\mathcal{N}}(x; P_2, M_2) \qquad \tilde{\mathcal{N}}(x; P_1, M_1)/\tilde{\mathcal{N}}(x; P_2, M_2), \text{ respectively.}$$

2. **Theory Question** The Dirichlet distribution on multinomial distributions $\pi \in [0,1]^K$ with concentration parameter $\alpha \in \mathbb{R}^K$, $K \in \mathbb{N}$ has pdf (using a Dirac distribution $\delta(\mathbf{1}^\intercal\pi - 1)$ to explicitly encode that the entries of the probability π have to sum to 1) given by

$$p_\pi(\pi|\alpha) \propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \delta(\mathbf{1}^\top \pi - 1). \tag{2}$$

We will use $\delta(x) = \lim_{\epsilon \to \infty} \epsilon \exp(-\epsilon x^2)$. In this exercise, we will develop a Gaussian approximation to the Dirichlet as a Laplace approximations. Since π is a probability (i.e. its entries lie in $[0,1]$), and the Gaussian is a distribution on $\mathbb{R}^K$, we introduce a change of basis[1] $\pi = \sigma^{-1}(y)$, where σ is the softmax,

$$\sigma(y)_i = \frac{\exp(y_i)}{\sum_{j=1}^{K} \exp(y_j)}. \tag{3}$$

Since the Jacobian of this transformation is proportional to $\prod_k \sigma(y)$, we may write the distribution p_π of (2) in terms of y as

$$p_y(y|\alpha) \propto \prod_{k=1}^{K} \sigma(y)_k^{\alpha_k} g_\epsilon(\mathbf{1}^\top y), \quad g_\epsilon(x) = \epsilon \exp(-\epsilon x^2). \tag{4}$$

Note that the factor $\alpha_k - 1$ in the power of $\sigma(y)$ turned into α_k. You may take this pdf for granted.

[1] David JC MacKay. *Choice of Basis for Laplace Approximation.* Machine Learning **33**/1, 1998, 77–86

(a) Show that the Hessian of the negative logarithm $h = -\log p$ of the distribution in (4) is given by $H = (H_{k\ell})_{k\ell}$, with

$$H_{k\ell}(y) = \frac{\partial^2 h}{\partial y_k \partial y_\ell} = |\alpha|(\delta_{k\ell}\sigma(y)_k - \sigma(y)_k\sigma(y)_\ell) + 2\epsilon(\mathbf{1}\mathbf{1}^\top)_{k\ell}. \tag{5}$$

Here, we denoted $|\alpha| = \sum_k \alpha_k$.

(b) Show that the mode of p (as in Eq. (4)) is given by $y = \mu_k$,

$$\mu_k = \log \alpha_k - \frac{1}{K}\sum_{\ell=1}^{K} \log \alpha_\ell. \tag{6}$$

(c) Use the matrix inversion lemma (valid for matrices A, U, C, V, assuming all involved inverses exist)

$$(A + UCV^\intercal)^{-1} = A^{-1} - A^{-1}U(C^{-1} + V^\intercal A^{-1}U)^{-1}VA^{-1}$$

to show that the inverse Hessian is given by

$$H^{-1}_{k\ell} = \delta_{k\ell}\frac{1}{\alpha_k} - \frac{1}{K}\left[\frac{1}{\alpha_k} + \frac{1}{\alpha_\ell} - \frac{1}{K}\left(\frac{1}{\epsilon} + \sum_{m=1}^{K}\frac{1}{\alpha_m}\right)\right]. \tag{7}$$

(d) Since equation (7) is well-defined for all $\epsilon > 0$, we can take the limit $\epsilon \to \infty$. This, together with $K \gg 1$, allows us to approximate

$$\Sigma_{kk} = \frac{1}{\alpha_k}(1 - \frac{2}{K}) + \frac{1}{K^2}\sum_{m=1}^{K}\frac{1}{\alpha_m}. \tag{8}$$

Together, Eqs. (6) and (8) yield an analytic map $\alpha \mapsto (\mu, \Sigma)$. Show that, where this map can be inverted, its inverse is

$$\alpha_k = \frac{1}{\Sigma_{kk}}\left(1 - \frac{2}{K} + \frac{e^{\mu_k}}{K^2}\sum_\ell e^{-\mu_\ell}\right). \tag{9}$$

We will call these two approximations *the Laplace bridge* between Gaussians and Dirichlets. Note that it amounts to an *analytic* (in the sense of closed form, not involving optimization or integration) approximation between the two exponential families.

3. **Practical Question** In `Exercise_91.ipynb` we will use the Laplace bridge of Exercise 9.2 as a tool to perform approximate Gaussian process regression on time series involving count data. More information can be found in the Jupyter notebook.