Homework 1

Q1. **Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\beta \sim N(0, \tau^2 \mathbf{I})$, and Gaussian sampling model $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I})$. Find the relationship between the regularization parameter λ in the ridge formula, and the variances τ^2 and σ^2.**

S1. We assume that our input data is centered which allows us to ignore the intercept term β_0. The posterior distribution is given by

$$\Pr(\beta|\mathbf{y}, \mathbf{X}) = \frac{1}{K} \Pr(\mathbf{y}|\beta, \mathbf{X}) \Pr(\beta)$$

where $K = K(\mathbf{y}, \mathbf{X}) = \int \Pr(\mathbf{y}|\beta, \mathbf{X}) \Pr(\beta)\, d\beta$, is clearly independent of β. Using our assumptions, we see that

$$\Pr(\beta|\mathbf{y}, \mathbf{X}) = \frac{1}{Z} \frac{1}{(2\pi)^{p/2}\sigma^p} \exp\left(-\frac{(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2}\right) \frac{1}{(2\pi)^{p/2}\tau^p} \exp\left(-\frac{\beta^T\beta}{2\tau^2}\right). \tag{1}$$

Then,

$$\log\left(\Pr(\beta|\mathbf{y}, \mathbf{X})\right) = -C - \frac{(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2} - \frac{\beta^T\beta}{2\tau^2},$$

where C collects the terms without β dependence. It is then not difficult to see that this expression is maximized for

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}\right)^{-1} \mathbf{X}^T\mathbf{y}.$$

Letting $\lambda = \frac{\sigma^2}{\tau^2}$, we see the equivalence of the above approach and ridge regression.

It is clear that $\Pr(\beta|\mathbf{y}, \mathbf{X})$ is Gaussian and its mean and mode coincide. We will now show that its mean $m = \hat{\beta}$. To this end, note that (1) implies that its covariance Σ satisfies

$$\Sigma^{-1} = \frac{1}{\sigma^2}\left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}\right).$$

This gives $\hat{\beta} = \frac{1}{\sigma^2}\Sigma\mathbf{X}^T\mathbf{y}$ and equating the relevant terms in (1), we see that this must be the mean.

Q2. **Show that the ridge regression estimates can be obtained by ordinary least squares regression on an augmented data set. We augment the centered matrix $\mathbf{X}$ with p additional rows $\sqrt{\lambda}\mathbf{I}$ and augment $\mathbf{y}$ with p zeroes. By introducing artificial data having response value zero, the fitting procedure is forced to shrink the coefficients toward zero.**

S2. Denote by $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ the augmented data sets, i.e.,

$$\tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I}_{p\times p} \end{pmatrix}, \qquad \tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y} \\ 0_{p\times 1} \end{pmatrix}.$$

By (3.6) in [HTF09] an ordinary least squares regression yields the estimate

$$\hat{\beta} = \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}\tilde{\mathbf{y}}.$$

Using the definition of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$, it is not difficult to see

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} \quad \text{and} \quad \tilde{\mathbf{X}}^T\tilde{\mathbf{y}} = \mathbf{X}^T\mathbf{y}.$$

So, $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1} \mathbf{X}^T\mathbf{y}$.

Q3. **Consider a mixture model density in p-dimensional feature space,**

$$g(x) = \sum_{k=1}^{K} \pi_k g_k(x),$$

where $g_k = \mathcal{N}(\mu_k, \sigma^2 \mathbf{I})$ and $\pi_k \geq 0$ for all k with $\sum_k \pi_k = 1$. Here $\{\mu_k, \pi_k\}$, $k = 1, ..., K$ and σ^2 are unknown parameters. Suppose we have data $x_1, ..., x_N \sim g(x)$ and we wish to fit the mixture model.

- **a. Write down the log-likelihood of the data.**
- **b. Derive an EM algorithm for computing the maximum likelihood estimates.**
- **c. Show that if σ has a known value in the mixture model and we take $\sigma \to 0$, then in a sense, this EM algorithm coincides with K-means clustering.**

S3.a) The log-likelihood function for $\{x_i\}_{i=1}^N$ is given by

$$l(\theta, \mathbf{Z}) = \log\left(\prod_{i=1}^{N} g(x_i)\right) = \log\left(\prod_{i=1}^{N}\left(\sum_{k=1}^{K} \pi_k g_k(x_i)\right)\right) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k g_k(x_i)\right), \tag{2}$$

where $\theta = (\sigma^2, \theta_1, ..., \theta_K) = (\sigma^2, \pi_1, \mu_1, ..., \pi_K, \mu_K)$.

S3.b) We generalize the ideas in [HTF09, p.272]. Introduce the random vector $\Delta = (\Delta_1, ..., \Delta_K)$ satisfying $\Delta_k \in \{0, 1\}$, $\sum_{k=1}^{K} \Delta_k = 1$ and $\Pr(\Delta_k = 1) = \pi_k$. Note $\Pr(\Delta) = \prod_{k=1}^{K} \pi_k^{\Delta_k}$ and $\Pr(x|\Delta_k = 1) = g_k(x)$. We set

$$\gamma_{kn}(\theta) := \Pr(\Delta_k = 1|\theta, \mathbf{Z} = x_n) = \frac{\pi_k g_k(x_n)}{\sum_{j=1}^{K} \pi_j g_j(x_n)}. \tag{3}$$

In (2) we calculate the derivatives $\frac{dl}{d\mu_k}$, $\frac{dl}{d\sigma^2}$ and $\frac{dl}{d\pi_k}$, we determine their zeros and find the extreme points

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma_{nk} x_n}{\sum_{n=1}^{N} \gamma_{nk}}, \quad \sigma^2 = \frac{\sum_{k=1}^{K} \sum_{n=1}^{N} \gamma_{nk} (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{k=1}^{K} \sum_{n=1}^{N} \gamma_{nk}}, \quad \pi_k = \frac{\sum_{n=1}^{N} \gamma_{nk}}{N} \tag{4}$$

Now, guess $\mu_k^0, \sigma^0, \pi_k^0$ and calculate γ_{kn}^0 using (3). With γ_{kn}^0 at hand, we can use (4) to update our parameters to $\mu_k^1, \sigma^1, \pi_k^1$. Repeating this procedure, we improve our estimates. To see why, assume we have determined $\mu_k^i, \sigma^i, \pi_k^i$. Note that $\sum_k \gamma_{nk}^i = 1$ and $\gamma_{nk}^i \geq 0$. Applying Jensen's inequality to (2), we get

$$l(\theta, \mathbf{Z}) \geq \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk}^i \log\left(\frac{\pi_k g_k(x_n)}{\gamma_{nk}^i}\right) = \sum_{n,k} \gamma_{nk}^i \log(\pi_k g_k(x_n)) - \sum_{n,k} \gamma_{nk}^i \log(\gamma_{nk}^i) = B_i(\theta).$$

The extreme points for $B_i(\theta)$ turn out to be $\mu_k^{i+1}, \sigma^{i+1}, \pi_k^{i+1}$ when calculated in (4) using γ_{nk}^i.

In conclusion, the EM algorithm is given by:

1. Take initial guesses for the parameters $\sigma^2, \hat{\mu}_i, \hat{\pi}_i$ for $i = 1, ..., K$.

2. Expectation step: Compute the responsibilities

$$\hat{\gamma}_{nk} = \frac{\pi_k g_k(x_n)}{\sum_{j=1}^{K} \hat{\pi}_j g_j(x_n)}, \qquad i = 1, ..., N, \quad k = 1, ..., K$$

3. Maximization step: Compute the weighted means and variances

$$\hat{\mu}_k = \frac{\sum_{n=1}^{N} \hat{\gamma}_{nk} x_n}{\sum_{n=1}^{N} \hat{\gamma}_{nk}}, \quad \hat{\sigma}^2 = \frac{\sum_{k=1}^{K} \sum_{n=1}^{N} \hat{\gamma}_{nk} (x_n - \hat{\mu}_k)(x_n - \hat{\mu}_k)^T}{\sum_{k=1}^{K} \sum_{n=1}^{N} \hat{\gamma}_{nk}}, \quad \hat{\pi}_k = \frac{\sum_{n=1}^{N} \hat{\gamma}_{nk}}{N}$$

4. Iterate steps 2 and 3 until convergence.

S3.c) For each n choose j such that $(x_n - \mu_j)^T(x_n - \mu_j) \leq (x_n - \mu_k)^T(x_n - \mu_k)$ for all k and provided $\pi_k \neq 0$. Note from (3), that for $k \neq j$ $\gamma_{nk} \to 0$ as $\sigma \to 0$ and $\gamma_{nj} \to 1$. Hence, we can write

$$\gamma_{nk} \to r_{nk} := \begin{cases} 1 \text{ if } k = \arg\min_j (x_n - \mu_j)^T(x_n - \mu_j) \\ 0 \text{ otherwise} \end{cases},$$

which assigns each data point to the cluster having the closest mean.

Q4. **Derive equation (6.8) in [HTF09, p. 195] for multidimensional x.**

S4. We want to determine

$$\left(\hat{\beta}_0, ..., \hat{\beta}_p\right) = \arg\min_{\beta_0,...,\beta_p} \sum_{j=1}^{N} K_\lambda(x_0, x_j) \left(y_j - \beta_0(x_0) - \sum_{i=1}^{p} \beta_i(x_0) x_{i,j}\right)^2.$$

Define $b(x)^T = (1, x)$, let $\mathbf{B}$ be the regression matrix with ith row $b(x_i)^T$, and $\mathbf{W}(x_0)$ the matrix with ith diagonal element $K_\lambda(x_0, x_i)$. Then,

$$\hat{\beta} = \left(\hat{\beta}_0, ..., \hat{\beta}_p\right) = \arg\min_{\beta=(\beta_0,...,\beta_p)} (\mathbf{B}\beta - \mathbf{y})^T \mathbf{W}(x_0) (\mathbf{B}\beta - \mathbf{y})$$

It is then not difficult to reduce the problem to ordinary least squares, which yields

$$\beta = (\mathbf{B}\mathbf{W}(x_0)\mathbf{B})^{-1} \mathbf{B}^T \mathbf{W}(x_0)\mathbf{y}.$$

Consequently,

$$\hat{f}(x_0) = b(x_0)^T (\mathbf{B}\mathbf{W}(x_0)\mathbf{B})^{-1} \mathbf{B}^T \mathbf{W}(x_0)\mathbf{y}.$$

References

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning. 2(1), 2009.