CS 189: Introduction to Machine Learning - Discussion 3

1. Matrix calculus

   Let A be a $n \times n$ matrix, and let $x \in \mathbb{R}^n$. Find $\nabla_x(x^\top Ax)$.

Solution: We can write

$$x^\top Ax = \sum_i \sum_j a_{ij}x_ix_j$$

Let's differentiate this with respect to a single element x_k:

$$\frac{\partial}{\partial x_k}(x^\top Ax) = \frac{\partial}{\partial x_k}\left(\sum_i \sum_j a_{ij}x_ix_j\right)$$

We can drop all terms that don't contain x_k:

$$= \frac{\partial}{\partial x_k}\left[\left(\sum_i a_{ik}x_ix_k\right) + \left(\sum_j a_{kj}x_kx_j\right) - a_{kk}x_k^2\right]$$

Isolating the x_k^2 terms gives

$$= \frac{\partial}{\partial x_k}\left[\left(\sum_{i\neq k} a_{ik}x_ix_k\right) + \left(\sum_{j\neq k} a_{kj}x_kx_j\right) + a_{kk}x_k^2\right]$$

$$= \frac{\partial}{\partial x_k}\left[\left(\sum_{i\neq k} a_{ik}x_i\right)x_k + \left(\sum_{j\neq k} a_{kj}x_j\right)x_k + a_{kk}x_k^2\right]$$

Now we can differentiate with respect to x_k

$$= \left(\sum_{i\neq k} a_{ik}x_i\right) + \left(\sum_{j\neq k} a_{kj}x_j\right) + 2a_{kk}x_k$$

$$= \left(\sum_{i} a_{ik}x_i\right) + \left(\sum_{j} a_{kj}x_j\right)$$

$$= (k^{\text{th}} \text{ column of A})^\top x + (k^{\text{th}} \text{ row of A})^\top x$$

Placing all partial derivatives into a single vector, we get

$$\nabla_x(x^\top Ax) = (A^\top + A)x$$

Notice that if A is symmetric, this reduces to

$$\nabla_x(x^\top Ax) = 2Ax$$

2. Logistic Posterior with different variances

We have seen in class that Guassian class conditionals with the same variance lead to a logistic posterior. Now we will consider the case when the class conditionals are Guassian, but have different variance, i.e.

$$\begin{aligned} X|Y = i &\sim \mathcal{N}(\mu_i, \sigma_i^2), \quad \text{where } i \in \{0, 1\} \\ Y &\sim \text{Bernoulli}(\pi) \end{aligned}$$

Show that the posterior distribution of the class label given X is also a logistic function, however with a quadratic argument in X. Assuming 0-1 loss, what will the decision boundary look like (i.e., describe what the posterior probability plot looks like)?

Solution:

We are solving for $\mathbb{P}(Y = 1|x)$. By Bayes Rule, we have

$$\begin{aligned} \mathbb{P}(Y = 1|x) &= \frac{\mathbb{P}(x|Y = 1)\mathbb{P}(Y = 1)}{\mathbb{P}(x|Y = 1)\mathbb{P}(Y = 1) + \mathbb{P}(x|Y = 0)\mathbb{P}(Y = 0)} \\ &= \frac{1}{1 + \frac{\mathbb{P}(Y=0)\mathbb{P}(x|Y=0)}{\mathbb{P}(Y=1)\mathbb{P}(x|Y=1)}} \\ &= \frac{1}{1 + \frac{\sigma_1}{\sigma_0}\frac{1-\pi}{\pi}\exp\left(\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(x-\mu_0)^2}{2\sigma_0^2}\right)} \end{aligned}$$

Looking at the bottom right equation, we have

$$\frac{\sigma_1}{\sigma_0}\frac{1-\pi}{\pi}\exp\left(\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(x-\mu_0)^2}{2\sigma_0^2}\right)$$

$$= \exp\left[\left(\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_0^2}\right)x^2 + \left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right)x + \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_0^2}{\sigma_0^2} + \ln\left(\frac{\sigma_1}{\sigma_0}\frac{1-\pi}{\pi}\right)\right)\right]$$

Now we see that we have a logistic function $\frac{1}{1+\exp(-h(x))}$, where $h(x) = ax^2 + bx + c$, for appropriate values of a, b, c, is a quadratic function. Note that the special case examined in class of $\sigma_1 = \sigma_0$ gives a linear function in x.

Since we are assuming 0-1 loss, we use the optimal classifier $f^*(x) = 1$ when $\mathbb{P}(Y = 1|x) > \mathbb{P}(Y = 0|x)$. Thus, the decision boundary can be found when $\mathbb{P}(Y = 1|x) = \mathbb{P}(Y = 0|x) = \frac{1}{2}$. This happens when $h(x) = 0$. Solving for the roots of $h(x)$ results in 2 values where this equality holds. One can convince herself/himself that in the plot of posterior probability graph, the horizontal (x) axis will be split into three regions: we classify the two outer regions as one class, and the middle one as another class. The choice of which class to classify in the outer regions depends on the values of σ_1 and σ_2.

3. MLE of the Laplace Distribution

Let X have a Laplace distribution with density

$$f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

Suppose that n samples $x_1, \ldots, x_n$ are drawn independently according to $f(x|\mu, b)$.

(a) Find the maximum likelihood estimate of μ.

Solution:

$$\mathcal{L}(\mu, b|D) = \prod_{i=1}^{n} \frac{1}{2b} exp\left(-\frac{|x_i - \mu|}{b}\right) = \left(\frac{1}{2b}\right)^n exp\left(-\frac{1}{b}\sum_{i=1}^{n}|x_i - \mu|\right)$$

$$l(\mu, b|D) = n\log(\frac{1}{2b}) - \frac{1}{b}\sum_{i=1}^{n}|x_i - \mu|$$

We take the argmax with respect to μ, and arrive here:

$$\underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} |x_i - \mu|$$

where we see that the best value for μ_{MLE} is the sample median.

(b) Find the maximum likelihood estimate of b.

Solution: Taking the derivative of the log likelihood, we have:

$$\frac{\partial}{\partial b} l(\mu, b|D) = -\frac{n}{b} + \frac{\sum_{i=1}^{n}|x_i - \mu|}{b^2} = 0$$

$$b = \frac{1}{n}\sum_{i=1}^{n}|x_i - \mu|$$

which is the average absolute deviation from the mean.

(c) Assume that μ is given. Show that b_{MLE} is an unbiased estimator (to show that the estimator is unbiased, show that $\text{E}[b_{\text{MLE}} - b] = 0$).

Solution:

$$E[b_{MLE} - b] = E[b_{MLE}] - b = 0$$

To solve for $E[b_{MLE}]$, we use linearity of expectation:

$$E[b_{MLE}] = \frac{1}{n}\sum_{i=1}^{n} E[|x_i - \mu|] = E[|x_i - \mu|]$$

where the last equality holds because each random variable has the same expectation. Now, consider what is in the absolute value signs. We have a random variable minus a constant. A good exercise is to prove that a random variable plus a constant shifts the entire distribution by that constant (this is very intuitive to see). I define $z = x_i - \mu$. Thus, z is a Laplacian RV with parameters $0, b$.

We have boiled down $E[b_{MLE}]$ to $E[|z|]$. What is the PDF of $|z|$? Well, since z is symmetric about the origin, we can simply double the probability density function (or fold it over) from 0 to infinity and define $z' = |z|$ to take on real values in the range $[0, \infty)$. The density of z' is then

$$f(z) = \frac{1}{b} exp(-\frac{z}{b})$$

This is the density of an exponential random variable with parameter $\frac{1}{b}$, therefore, $E[z'] = b$, showing that (given μ), b_{MLE} is unbiased.