CS 189: Introduction to Machine Learning - Discussion 7

1. Performance evaluation

   Suppose that we use some learning method to make a prediction y for a particular data point $\mathbf{x}$. Assume we are given a training set and validation set.

   1. Describe how we might estimate the standard deviation of our validation risk using bootstrap samples.
   2. Describe how we might estimate the standard deviation of our error rate by assuming errors are drawn from a binomial distribution.

Solution:

1. The bootstrap approach works by sampling (with replacement) a set of n observations from the training data set. This is done b times, for some large value of b, each time fitting a new model. We obtain the sample standard deviation of the validation risk over all b models.
2. This is a binomial distribution. We are looking for the standard deviation of the error rate:

$$\sigma = \sqrt{\frac{e(1-e)}{n}}, \text{ where } e \text{ is the validation error rate}$$

2. Logistic Posterior with different variances

We have seen that Gaussian class conditionals with the same variance lead to a linear decision boundary. Now we will consider the case where class conditionals are Gaussian but have different variances, 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 themself 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. Linear Regression with Laplace prior

   We saw in discussion 4 that there is a probabilistic interpretation of linear regression: $P(y|\mathbf{x}, \sigma^2) \sim \mathcal{N}(\mathbf{w^T x}, \sigma^2)$. We extend this by assuming some prior distribution on parameters $\mathbf{w}$. Let us assume the prior is a Laplace distribution, so we have:

$$w_j \sim Laplace(0, t), \text{ i.e. } P(w_j) = \frac{1}{2t} e^{-|w_j|/t} \text{ and } P(\mathbf{w}) = \prod_{j=1}^{D} P(w_j) = (\frac{1}{2t})^D \cdot e^{-\frac{\sum |w_j|}{t}}$$

   Show it is equivalent to minimizing the following risk function, and find the value of the constant λ:

$$R(\mathbf{w}) = \sum_{i=1}^{n} (y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 + \lambda \|\mathbf{w}\|_1, \text{ where } \|\mathbf{w}\|_1 = \sum_{j=1}^{D} |w_j|$$

Solution: Note that $\mathbf{X_i} = \mathbf{x}^{(i)}, Y_i = y^{(i)}$. We have to solve the MAP for parameter $\mathbf{w}$ and the posterior of $\mathbf{w}$ is,

$$P(w|\mathbf{X_i}, Y_i) \propto (\prod_{i=1}^{n} \mathcal{N}(Y_i|\mathbf{w^T X_i}, \sigma^2)) \cdot P(\mathbf{w}) = (\prod_{i=1}^{n} \mathcal{N}(Y_i|\mathbf{w^T X_i}, \sigma^2)) \cdot \prod_{j=1}^{D} P(w_j)$$

Taking log and we want to maximize

$$\begin{aligned} l(\mathbf{w}) &= \sum_{i=1}^{n} log \mathcal{N}(Y_i|\mathbf{w^T X_i}, \sigma^2) + \sum_{j=1}^{D} log P(w_j) \\ &= \sum_{i=1}^{n} log(\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(Y_i - \mathbf{w^T X_i})^2}{2\sigma^2})) + \sum_{j=1}^{D} log(\frac{1}{2t} exp(\frac{-|w_j|}{t})) \\ &= -\sum_{i=1}^{n} \frac{(Y_i - \mathbf{w^T X_i})^2}{2\sigma^2} + \frac{-\sum_{j=1}^{D} |w_j|}{t} + nlog(\frac{1}{\sqrt{2\pi}\sigma}) + Dlog(\frac{1}{2t}) \end{aligned}$$

So it is equivalent to minimizing the following function:

$$R(\mathbf{w}) = \sum_{i=1}^{n} (Y_i - \mathbf{w^T X_i})^2 + \frac{2\sigma^2}{t} \sum_{j=1}^{D} |w_j| = \sum_{i=1}^{n} (Y_i - \mathbf{w^T X_i})^2 + \lambda \|\mathbf{w}\|_1$$

where $\lambda = \frac{2\sigma^2}{t}$.