STAT 535, Homework 4

Due date: Nov 21 Thursday 23:59:59. Submit the homework through Canvas in a PDF file. If the questions involved programming, please include your codes.

1. In binary classification with $0-1$ loss, we see that we should classify the label based on the label with a higher probability. This will not be true when using other loss function.

   Consider the following loss function

   $$L(c(x), y) = \begin{cases} 0, & \text{if } c(x) = y \\ 1, & \text{if } c(x) = 0 \text{ and } y = 1 \\ 2, & \text{if } c(x) = 1 \text{ and } y = 0 \end{cases}.$$

   Namely, we will loss more when we misclassify a label 0 to a label 1.

   (a) **(5 pts)** In this new loss function, the Bayes classifier $c^*(x)$ (the classifier that minimizes the risk) will be

   $$c^*(x) = \begin{cases} 0, & \text{if } P(0|x) \geq \tau_0 \cdot P(1|x), \\ 1, & \text{if } P(1|x) > \frac{1}{\tau_0} \cdot P(0|x) \end{cases}$$

   for some constant τ_0. Find out what is τ_0.

   (b) **(5 pts)** In this case, the classifier from a regression estimator also needs to be modified. Let $m(x) = \mathbb{E}(Y|X = x)$ be the regression function. Show that the Bayes classifier is equivalent to

   $$c^*(x) = \begin{cases} 0, & \text{if } m(x) \leq \rho_0, \\ 1, & \text{if } m(x) > \rho_0 \end{cases}$$

   for some ρ_0.

2. The `iris` dataset is a famous dataset in Statistics. You can find this dataset in `R` by typing `iris`. It is a dataset consists of $n = 150$ observations with 4 continuous variables and 1 categorical variable (`species`). To simplify the problem, we focus on `species` equals to `versicolor` and `virginica` so that it becomes a binary classification problem. We will use the idea of k-NN classification to perform our analysis with the `species` variable being the class label and the other 4 variables are the features.

   However, a challenge of using the k-NN in this case is that the 4 variables have different ranges so naively using the k-NN may not work well. So we redefine the distance as follows. Let $\widehat{\Sigma}$ be the sample covariance matrix (of `versicolor` and `virginica` species). For any two feature vectors $X_1, X_2 \in \mathbb{R}^4$, their distance is given by

   $$d_\Sigma(X_1, X_n) = \sqrt{(X_1 - X_2)^T \widehat{\Sigma}^{-1} (X_1 - X_2)}.$$

   This distance is also known as the Mahalanobis distance. We use $0-1$ loss in our classification.

   (a) **(5 pts)** With d_Σ, we may compute the performance of k-NN. Apply 5-fold cross-validation to $k = 3, 4, 5, \cdots, 10$. Show the cross-validation error versus k. If there is a tie, we randomly assign

it to a class in the tie with an equal probability.

Note: you may use packages for finding the k-NN observations or distance to the k-th NN observation of any given point. But you cannot use the package to directly do the classification part.

Under the optimal k, what is the performance under the 5-fold cross-validation error?

(b) **(5 pts)** In regression or classification, sometimes we are interested in the *variable importance*– some variable may not contribute to the classification performance much. To investigate this, we may remove one variable and re-compute the classification error under 5-fold cross-validation with the same optimal k chosen in the previous question (we do not search for a new optimal k). The change in classification error will be a measure of variable importance. Since we have 4 variables, each time we remove one variable and use the other 3 variables to do classification. Show the change in the classification error when removing each variable. Which variable is the least important one (the error decrease the least)? and which one is the most important one (the error decreases the most)?

3. *Mean-shift algorithm.* We will derive some interesting properties of the mean-shift algorithm and mode clustering. Let $X_1, \cdots, X_n \in \mathbb{R}$ be IID from an unknown PDF p and let

$$\widehat{p}_h(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

where $K(x)$ is the Gaussian kernel. Given an initial point $x_{(0)}$, the the mean-shift algorithm updates it via

$$x_{(t+1)} = \frac{\sum_{i=1}^{n} X_i K\left(\frac{X_i - x_{(t)}}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x_{(t)}}{h}\right)}. \tag{1}$$

Alternatively, you may implement the gradient ascent algorithm, which updates $x_{(t)}$ via

$$x_{(t+1)} = x_{(t)} + \gamma \nabla \widehat{p}_h(x_{(t)}), \tag{2}$$

where $\gamma > 0$ is the step size.

(a) **(5 pts)** Show that the mean-shift algorithm is the gradient descent algorithm with a particular choice of γ. What is the step size γ for the mean-shift algorithm?

(b) **(5 pts)** We know that taking a monotone transform will not change the location of the modes. Consider the log-PDF $\log \widehat{p}_h(x)$ and its gradient $\nabla \log \widehat{p}_h(x)$. Show that the mean-shift algorithm is performing a gradient ascent of $\nabla \log \widehat{p}_h(x)$ with a specific step size $\eta > 0$. The gradient ascent of $\nabla \log \widehat{p}_h(x)$ is

$$x_{(t+1)} = x_{(t)} + \eta \nabla \log \widehat{p}_h(x_{(t)}).$$

What is η in this case?

(c) **(5 pts)** During the mean-shift process, there is a very interesting phenomenon that the observations are being moved toward local modes. In fact, this can be viewed as creating a sequence of push-forward measures.

Here is how a push-forward measure can be understood. Suppose that we have a random variable $Z \sim P_Z$, where P_Z is the probability distribution/measure of Z. Define $Y = f(Z)$ to be another

random variable. Y itself has a distribution/measure P_Y. Because Y is created by a function transform $Y = f(Z)$, the probability measure between P_Z and P_Y are related. We call P_Y a push-forward measure of P_Z and denote it as $P_Y = f\#P_Z$.

Suppose that we apply a one-step gradient descent to the observed data, we create a new set of observations

$$X_{1,(1)}, \cdots, X_{n,(1)},$$

where $X_{i,(1)} = X_{i,(0)} + \gamma \nabla \widehat{p}_h(X_{i,(0)})$ and $X_{i,(0)} = X_i$. The original observations form an empirical measure $\widehat{P}_n$ and the new set of observations form another empirical measure $\widehat{P}_n^{(1)}$. Using the notation of the push-forward measure, we can associate them via

$$\widehat{P}_n^{(1)} = \widehat{f}\#\widehat{P}_n.$$

Find out the expression of $\widehat{f}$.

4. *k-means clustering.* In this question, we will study the k-means problem from the quantization point of view.

   (a) **(5 pts)** Suppose that we have a distribution of interest P that we can sample from but we do not know how to represent this distribution function (it may not have a simple closed-form). Suppose that we want to represent this distribution using k points. We know that k points is not sufficient to represent the entire distribution function so we define the risk of representation using points $\mathbf{c} = \{c_1, \cdots, c_k\}$ as

   $$R(\mathbf{c}) = \mathbb{E} \min_j \|X - c_j\|^2 = \int \min_j \|x - c_j\|^2 dP(x),$$

   where X is a random variable with distribution P. In practice, we do not have a simple form of the distribution function P so we approximate the above expectation by a Monte Carlo approximation that we generates $X_1, \cdots, X_n \sim P$, which leads to the *empirical risk:*

   $$\widehat{R}_n(\mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_j \|X_i - c_j\|^2.$$

   Show that the minimization of $\widehat{R}_n(\mathbf{c})$ is the same as the k-means minimization problem.

   (b) **(5 pts)** Following the previous question (2-(a)). Let $\mathbf{c}^*$ be the optimal population quantizer (k points) and $\widehat{\mathbf{c}}^*$ be the optimal sample quantizer. Namely,

   $$\mathbf{c}^* = \mathsf{argmax}_{\mathbf{c}} R(\mathbf{c}), \quad \widehat{\mathbf{c}}^* = \mathsf{argmax}_{\mathbf{c}} \widehat{R}_n(\mathbf{c}).$$

   Since the empirical risk is a 'sample-mean' process with its expectation being the population mean risk of a given $\mathbf{c}$, we would expect the difference to be small. Suppose that we have a bound:

   $$\sup_{\mathbf{c}} |\widehat{R}_n(\mathbf{c}) - R(\mathbf{c})| \leq \epsilon.$$

   Show that this implies

   $$R(\widehat{\mathbf{c}}^*) - R(\mathbf{c}^*) \leq 2\epsilon.$$

(c) **(5 pts)** Following the first question (2-(a)). Suppose that the distribution of interest has a PDF p that we can evaluate everywhere. However, we are unable to sample from p. In this case, we can still construct the optimal k points to represent P via the importance sampling. Suppose that we know how to sample from another density q such that $q(x) = 0 \Rightarrow p(x) = 0$ for any x (you can think of q as the Gaussian). We can then generate $Z_1, \cdots, Z_n \sim q$.

Show that we can approximate the risk function $R(\mathbf{c})$ using

$$\tilde{R}_q(\mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} W_i \min_j \|Z_i - c_j\|^2,$$

where W_i depends on Z_i, q, p. Give the closed-form of W_i.