Machine learning@Data science, 2021–Exercises.

Set 2. Due February 12, 2021

Problem 5 Consider a binary classification problem in which the observation X is real valued, $\mathbf{P}\{Y=0\} = \mathbf{P}\{Y=1\} = 1/2$, and the class-conditional cumulative distribution functions are

$$\mathbf{P}\{X \le x | Y=0\} = \begin{cases} 0 & \text{if } x \le 0 \\ x/2 & \text{if } 0 < x \le 2 \\ 1 & \text{if } x > 2 \end{cases} \quad \text{and} \quad \mathbf{P}\{X \le x | Y=1\} = \begin{cases} 0 & \text{if } x \le 1 \\ (x-1)/3 & \text{if } 1 < x \le 4 \\ 1 & \text{if } x > 4 \end{cases}$$

Determine $\eta(x) = \mathbf{P}\{Y=1|X=x\}$. Compute the Bayes classifier and the Bayes risk R^*. Compute the asymptotic risks R_{1-NN} and R_{3-NN} of the 1-, and 3-nearest neighbor classifiers.

Problem 6 (NEAREST NEIGHBOR REGRESSION.) Let (X,Y) be a pair of random variables such that $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ is real valued and $\mathbf{E}Y^2 < \infty$. Denote the regression function by $m(x) = \mathbf{E}[Y|X=x]$. Suppose one wants to predict the value of Y upon observing X. If $f : \mathbb{R}^d \to \mathbb{R}$ is a predictor, define its risk by the mean squared error $R(f) = \mathbf{E}(f(X)-Y)^2$. Determine the optimal predictor f^* (i.e., the one minimizing the risk) and give an expression of $R(f^*)$.

Now let $(X_1,Y_1),\dots,(X_n,Y_n)$ be training data of n i.i.d. pairs, distributed as (X,Y). For a given $x \in \mathbb{R}^d$, order the X_i according to their distance to x so that $X^{(1)}(x)$ is the nearest neighbor of x, $X^{(2)}(x)$ is the second nearest neighbor of x, etc. Let $Y^{(1)}(x), Y^{(2)}(x), \dots$ denote the corresponding labels.

Define nearest neighbor regression by $f_n^{1NN}(x) = Y^{(1)}(x)$. What is the asymptotic risk

$$\lim_{n\to\infty} \mathbf{E}R(f_n^{1NN})?$$

(No formal proof is necessary.)

Determine the asymptotic risk $\lim_{n\to\infty} \mathbf{E}R(f_n^{k-NN})$ of the k-nearest neighbor regression estimator defined by $f_n^{k-NN}(x) = (Y^{(1)}(x) + \cdots + Y^{(k)}(x))/k$.

Interpret your findings.

Problem 7 Consider a binary classification problem with observation/label pair (X,Y) and assume that the corresponding Bayes risk (i.e., minimal probability of misclassification) is $R^* \in (0,1/2)$. Now suppose that $\boldsymbol{X} = (X_1,\dots,X_k)$ is such that, conditionally on any of the two values of Y, the components $X_1,\dots,X_k$ are i.i.d., with the same distribution as $X|Y$. Let R_k^* denote the Bayes risk corresponding to the classification problem $(\boldsymbol{X},Y)$. Prove that $R_k^* \le e^{-ck}$ for a positive constant c.

Problem 8 Coinsider a binary classification problem in which X takes values in $\mathbb{R}^d$ and $Y \in \{0,1\}$. The joint distribution is such that X is uniformly distributed in $[0,1]^d$ and $\mathbf{P}\{Y=1|X=x\} = (x^{(1)}+x^{(2)})/2$ (where $x^{(i)}$ is the i-th component of $x = (x^{(1)},\dots,x^{(d)})$).

Compute R^*,R^{1-NN}, and R^{3-NN} (i.e., the Bayes risk and the asymptotic risk of the 1-, and 3-nearest neighbor rules). How do these quantities depend on the dimension d?

Write a program that generates training data of n i.i.d. pairs $D_n = \{(X_1,Y_1),\dots,(X_n,Y_n)\}$ of random variables distributed as described above.

Classify X using the $1,3,5,7,9$-nearest neighbor rules. Re-draw (X,Y) many times so that you can estimate the risk of these rules. Try this for various values of n and d and plot the estimated risk. Explain what you observe.