Machine learning@Data science, 2019–Exercises.

Set 2. Due February 14, 2019

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

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

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

Problem 6 Let (X,Y) be a pair of random variables such that X takes its values in a set $\mathcal{X}$ and $Y \in \{0,1\}$ and let $\eta(x) = \mathbf{P}\{Y=1|X=x\}$.

Let (X',Y') have the same distribution as (X,Y) and assume that the two pairs are independent. Upon observing X and X', one wishes to rank them, that is, to decide whether $Y > Y'$ or $Y' > Y$. Thus, a *ranking rule* takes X, X' as an input and outputs a binary decision $p(X,X') \in \{0,1\}$ according to the guessed ranking.

Suppose that the loss equals one if $p(X,X') = 1$ and $Y < Y'$ or if $p(X,X') = 0$ and $Y > Y'$ and it equals zero in all other cases.

Determine the optimal ranking rule, that is, the one that minimizes the risk (i.e., expected loss).

Express the optimal risk in terms of the function $\eta(x)$.

Problem 7 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 $\mathbf{P}\{Y=1|X=x\} = x^{(1)}$ (where $x^{(1)}$ is the first 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.

Now consider the classification rule that uses an additional set of m independent data $D'_m = \{(X'_1,Y'_1), \dots, (X'_m,Y'_m)\}$ drawn from the same distribution to select the value of $k \in \{1,3,5,7,9\}$ in the k-nearest neighbor rule (trained on the data D_n) using empirical risk minimization based on D'_m. Estimate the probability of error of this rule (using independent test data) and compare it to the probability of error of the best of these five classification rules. How large does m have to be to make sure that the data/based selection is close to optimal?