1 Asymptotic Expected Risk of 1NN

In this problem you will investigate the asymptotic expected risk of the 1NN classifier and show that under certain assumptions, it will be upper-bounded by a constant factor of the Bayes risk. Consider a classification problem of K classes, and let

$$\theta_k(\mathbf{x}) := \text{Prob.}(\mathbf{x} \text{ is in class } k),\ k \in \{1, 2, \ldots, K\}$$

denote the true class probabilities given some feature vector $\mathbf{x}$. Given an i.i.d. sample of training pairs $D_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ drawn from some joint distribution $P(\mathbf{x}, y)$, where $\mathbf{x}_i$ denote the ith training vector and $y_i \in \{1, 2, \ldots, K\}$ denote the corresponding class label, the goal of supervised learning is to construct a classification rule $\hat{f}_n(\cdot)$ from D_n such that the expected risk over the training sample D_n and an unseen test example $(\mathbf{x}^*, y^*) \sim P(\mathbf{x}, y)$

$$\mathbb{E}_{D_n}[\mathbb{E}_{\mathbf{x}^*, y^*}[\mathbf{1}\{\hat{f}_n(\mathbf{x}^*) \neq y^*\}]]$$

is small. The Bayes rule, denoted as $f^{Bayes}(\cdot)$, is defined by the following property:

$$\mathbb{E}_{\mathbf{x}^*, y^*}[\mathbf{1}\{f^{Bayes}(\mathbf{x}^*) \neq y^*\}] \quad \leq \quad \mathbb{E}_{\mathbf{x}^*, y^*}[\mathbf{1}\{f(\mathbf{x}^*) \neq y^*\}] \quad \text{for any classification rule } f,$$

and the risk of the Bayes rule is called the Bayes risk.

1. What is the Bayes rule for the K-class classification problem? And what is the Bayes risk? Several notes:

   (a) For the Bayes rule, your answer should be in terms of $\theta_k(\mathbf{x})$.

   (b) For the Bayes risk, there should be an expectation over $\mathbf{x}$ in your answer.

   (c) Think about the conditional expectation of the loss given some $\mathbf{x}$.

   Ans. We first consider the conditional expectation of the loss given some $\mathbf{x}^*$ incurred by some classification rule f, which can be either deterministic or probabilistic. If f is probabilistic, then

$$\begin{aligned}
\mathbb{E}_{y^*}[\mathbf{1}\{f(\mathbf{x}^*) \neq y^*\}] &= \sum_{k=1}^{K} \theta_k(\mathbf{x}^*)(1 - P(\mathbf{1}\{f(\mathbf{x}^*) = k\})) \\
&= \sum_{k=1}^{K} (1 - \theta_k(\mathbf{x}^*)) P(\mathbf{1}\{f(\mathbf{x}^*) = k\}) \\
&\geq 1 - \max_k \theta_k(\mathbf{x}^*). \qquad (1)
\end{aligned}$$

   If f is deterministic, a similar argument shows that this inequality still holds. It is easy to see that the following rule

$$\arg\max_k \theta_k(\mathbf{x}^*)$$

   achieves the lower bound (1) on the conditional expected loss, and therefore is the Bayes rule. The Bayes risk then is

$$\mathbb{E}_{\mathbf{x}^*}[1 - \max_k \theta_k(\mathbf{x}^*)].$$

2. Let $\hat{f}_n^{1NN}(\cdot)$ denote the 1NN classification rule constructed from D_n. Since this rule depends on the random training sample D_n, the prediction $\hat{f}_n^{1NN}(\mathbf{x}^*)$ it makes on some test vector $\mathbf{x}^*$ is also random, and we can think about

$$P(\mathbf{1}\{\hat{f}_n^{1NN}(\mathbf{x}^*) = k\}),\ k \in \{1, 2, \ldots, K\}.$$

Moreover, if we increase the sample size n while holding fixed the dimension of feature vectors, we would expect, under reasonable assumptions on the joint distribution $P(\mathbf{x}, y)$, that $\mathbf{x}^*$ and its nearest neighbor become closer and closer to each other, and so do their class conditional probabilities. For simplicity, here we assume that as $n \to \infty$,

$$\mathbb{E}_{D_n}[\mathbf{1}\{\hat{f}_n^{1NN}(\mathbf{x}^*) = k\}] \ = \ P(\mathbf{1}\{\hat{f}_n^{1NN}(\mathbf{x}^*) = k\}) \ \to \ \theta_k(\mathbf{x}^*)$$

for all $k \in \{1, 2, \ldots, K\}$ and uniformly over all $\mathbf{x}^*$. Show that

$$\lim_{n\to\infty} \mathbb{E}_{D_n}[\mathbb{E}_{\mathbf{x}^*,y^*}[\mathbf{1}\{\hat{f}_n^{1NN}(\mathbf{x}^*) \neq y^*\}]] \ \leq \ 2\mathbb{E}_{\mathbf{x}^*,y^*}[\mathbf{1}\{f^{Bayes}(\mathbf{x}^*) \neq y^*\}],$$

i.e., the asymptotic expected risk of 1NN is no more than two times the Bayes risk. Some notes:

(a) You may find this fact useful: the solution to the following optimization problem:

$$\min \sum_{i=1}^{K} \theta_i^2 \quad \text{s.t.} \ \sum_{i=1}^{K} \theta_i = C$$

is $\theta_i^* = C/K$ for $i \in \{1, 2, \ldots, K\}$.

(b) Assume it is fine to exchange the limit with the expectation and vice versa.

Ans. Again, we begin by considering the conditional expected risk given some $\mathbf{x}^*$:

$$\begin{aligned}
\mathbb{E}_{y^*}\mathbb{E}_{D_n}[\mathbf{1}\{\hat{f}_n^{1NN}(\mathbf{x}^*) \neq y^*\}] &= \sum_{k=1}^{K} \mathbb{E}_{D_n}[\mathbf{1}\{\hat{f}_n^{1NN}(\mathbf{x}^*) \neq k\}]\theta_k(\mathbf{x}^*) \\
&= \sum_{k=1}^{K} P(\mathbf{1}\{\hat{f}_n^{1NN}(\mathbf{x}^*) \neq k\})\theta_k(\mathbf{x}^*) \\
&= \sum_{k=1}^{K} (1 - P(\mathbf{1}\{\hat{f}_n^{1NN}(\mathbf{x}^*) = k\}))\theta_k(\mathbf{x}^*).
\end{aligned}$$

Let $k^* := \arg\max_k \theta_k(\mathbf{x}^*)$, the class label given by the Bayes rule. Then we have

$$
\begin{aligned}
& \lim_{n\to\infty} \mathbb{E}_{y^*}\mathbb{E}_{D_n}[\mathbf{1}\{\hat{f}_n^{1NN}(\mathbf{x}^*) \neq y^*\}] \\
= & \sum_{k=1}^{K}(1-\theta_k(\mathbf{x}^*))\theta_k(\mathbf{x}^*) && \text{by our assumption} \\
= & \ 1-\theta_{k^*}(\mathbf{x}^*)^2 - \sum_{k\neq k^*}\theta_k(\mathbf{x}^*)^2 \\
\leq & \ 1-\theta_{k^*}(\mathbf{x}^*)^2 - \frac{(1-\theta_{k^*}(\mathbf{x}^*))^2}{K-1} && \text{by the property in (2a)} \\
= & \ (1-\theta_{k^*}(\mathbf{x}^*))\left(\frac{K(1+\theta_{k^*}(\mathbf{x}^*))-2}{K-1}\right) \\
\leq & \ 2(1-\theta_{k^*}(\mathbf{x}^*)),
\end{aligned}
$$

which implies the desired result.[1]

[1]This is true only when we can move the limit from outside of the expectation over $\mathbf{x}^*$ to inside the expectation. We avoid this technical difficulty by simply assuming we can do so, which is usually true if the joint distribution $P(\mathbf{x}, y)$ satisfies some regularity conditions.