Machine learning@Data science, 2016–Exercises.

Set 1. Due January 29, 2016

Problem 1 Consider the binary classification problem with a priori probabilities $\mathbf{P}\{Y=1\} = \mathbf{P}\{Y=0\} = 1/2$ and class-conditional densities $f_0(x) = f(x|Y=0)$ and $f_1(x) = f(x|Y=1)$ on $\mathcal{X} = \mathbb{R}^d$. Prove that the Bayes risk equals

$$R^* = \frac{1}{2} - \frac{1}{4}\int |f_0(x) - f_1(x)|dx \ .$$

Problem 2 Consider a binary classification problem in which both class-conditional densities are multivariate normal of the form

$$f_i(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma_i}} e^{-\frac{1}{2}(x-m_i)^T \Sigma_i^{-1}(x-m_i)}, \qquad i = 0, 1,$$

where $m_i = \mathbf{E}[X|Y=i]$ and Σ_i is the covariance matrix for class i. Let $q_0 = \mathbf{P}\{Y=0\}$ and $q_1 = \mathbf{P}\{Y=1\}$ be the a priori probabilities.

Determine the Bayes classifier. Characterize the cases when the Bayes decision is linear (i.e., it is obtained by thresholding a linear function of x).

Problem 3 Let (X, Y) be a pair of random variables taking values in $\mathcal{X} \times \mathbb{R}$ and consider a prediction problem in which one desires to guess the value of Y upon observing X. Suppose that the loss function is $\ell(y, y') = (y - y')^2$. Determine the predictor function $f : \mathcal{X} \to \mathbb{R}$ that minimizes the expected loss $\mathbf{E}\ell(f(X), Y)$.

Problem 4 Repeat the previous problem but with $\ell(y, y') = |y - y'|$. You may assume that for each $x \in \mathcal{X}$, the conditional distribution of Y, given $X = x$, has a density $\phi(y|x)$.

Problem 5 Let $X, X_1, \ldots, X_n$ be i.i.d. random vectors, uniformly distributed on $[0,1]^d$. Let k be a fixed positive integer and let $X_{(k)}$ denote the k-th nearest neighbor of X among $X_1, \ldots, X_n$. (We assume $n \geq k$.) Prove that

$$\lim_{n\to\infty} \|X_{(k)} - X\| = 0 \quad \text{in probability.}$$

Problem 6 Show that for any sample size n there exists a distribution of (X, Y) such that $R^* = 0$ but the expected risk of the 1-nearest neighbor classifier is greater than $1/4$.