MS1b: SDM - Problem Sheet 6

1. (k-Nearest Neighbours, Curse of Dimensionality) Consider using a k-NN classifier where the real-valued features are uniformly distributed in the p-dimensional unit cube. Suppose we are interested in estimating the distribution over class labels around a test point x by using neighbours within a hyper-cube centred at x.

   (a) Suppose we wish to use a fraction α of the training data to estimate the distribution over class labels at x. What should be the edge length of this hyper-cube to ensure that it includes on average $\alpha\%$ of the training data? If $p = 10$ and $\alpha = 1\%$, compute the edge length of this hyper-cube. In this scenario, is k-NN a "local" algorithm, i.e. using only local neighbours to x?

   (b) Assuming you have access to say $n = 500$ training data, does it appear reasonable to perform k-NN for large values of k (say $k > 10$)? Explain briefly why or why not.

2. (k-Nearest Neighbours, Risk) We will prove here that the asymptotic (in the number n of training data) error rate of the 1-nearest neighbour classifier is at most twice the Bayes-optimal error rate, for a 2 class classification problem.

   Let $(X_i, Y_i)_{i=1}^n$ be some training data where $X_i \in \mathbb{R}^p$ and $Y_i \in \{0, 1\}$. We denote by $f_k(x)$ the conditional density of X given $Y = k$ and assume that $f_k(x) > 0$ for any $x \in \mathbb{R}^p$. We also denote $\pi_k = P(Y = k)$.

   (a) Express $q(x) = P(Y = 1 | X = x)$ in terms of $f_0(x)$, $f_1(x)$ and π_1.

   (b) Consider the optimal Bayesian classifier minimizing the risk associated to the 0/1 loss function, equivalently maximizing the probability of correct classification; i.e.

   $$\widehat{y}_{\text{Bayes}}(x) = \arg\max_{k \in \{0,1\}} \pi_k f_k(x).$$

   Given some test point $X = x$, what is the expected probability of error (w.r.t. to the distribution of Y) of the optimal Bayesian classifier in terms of $q(x)$? [The resulting expression should depend *only* of $q(x)$].

   (c) The 1-nearest neighbour (1-nn) classifier assigns a test data point x the label of the closest training point; i.e. $\widehat{y}_{1\text{nn}}(x) = y$ (class of nearest neighbour in the training set). Given some test point $X = x$ with nearest neighbour x', what is the expected error of the 1-nn classifier (w.r.t. to the distribution of Y), in terms of $q(x)$, $q(x')$?

   (d) As the number of training data goes to infinity, i.e. $n \to \infty$, the training data fills the space in a dense fashion and the nearest neighbour x' of x is such that $q(x') \to q(x)$. By performing this substitution in the previous expression, give the asymptotic (in n) of the expected error of the 1-nn classifier given some test point $X = x$.

If we denote by $R_{\text{Bayes}} = \mathbb{E}\left[\mathbb{I}\left(Y \neq \widehat{y}_{\text{Bayes}}\left(X\right)\right)\right]$ and $R_{\text{1nn}} = \mathbb{E}\left[\mathbb{I}\left(Y \neq \widehat{y}_{\text{1nn}}\left(X\right)\right)\right]$, show that

$$R_{\text{Bayes}} \leq R_{\text{1nn}} \leq 2R_{\text{Bayes}}\left(1 - R_{\text{Bayes}}\right).$$

(e) Consider now the case where $Y_i \in \{0, 1, ..., K-1\}$ and show using the same reasoning that, as $n \to \infty$, we have

$$R_{\text{Bayes}} \leq R_{\text{1nn}} \leq R_{\text{Bayes}}\left(2 - \frac{K}{K-1}R_{\text{Bayes}}\right).$$

(Hint: Cauchy inequality yields $(K-1)\sum_{i\neq c} P^2\left(Y = i|\, x\right) \geq \left(\sum_{i\neq c} P\left(Y = i|\, x\right)\right)^2$).

3. Load the Vanveer gene expression data used in a previous practical and the previous problem sheet. Make use of the 20 ‘best’ genes (according to a marginal t-test) by using the following commands.

load(url(”http://www.stats.ox.ac.uk/%7Eteh/MS1b/PracticalObjects.RData”))

vanv <- vanveer.4000[,2:21]

prog <- vanveer.4000[,1]

Your X matrix is thus `vanv` and the response Y is `prog`. Split the data into a test and training set (of equal size).

Use k-nearest neighbour classification. Find an estimate of the test error rate as you vary k, the number of nearest neighbours. What seems to be a good choice of k, the number of nearest neighbours? What is the estimated misclassification error under an optimal choice of k? Is it possible to produce a ROC curve for k-nearest neighbour classification?