TMA4300 Computer Intensive Statistical Methods
Interactive lecture 5, Spring 2018

Problem A: Classification, apparent error rate and bootstrap estimate of bias

Assume we have observations $x_1, \ldots, x_n \in \mathbf{R}^2$ and corresponding class labels $y_1, \ldots, y_n \in \{0, 1\}$. Now we have a new observation x_0 and we want to classify the new observation to be in class 0 or to be in class 1. The k-nearest neighbour classifier do this by finding the k values of $x_1, \ldots, x_n$ that is closest to x_0 (in the usual Eucledian norm say) and assign to x_0 the class that is most frequent among the k "neighbours" to x_0.

1. Use rnorm to generate (for example) 40 samples from the bivariate normal distribution $N_2(0, I)$ and let these samples have class labels 0. Correspondingly generate (for example) 40 samples from the bivariate normal distribution $N_2([1, 1]^T, I)$ and let these have class labels 1. Visualise the simulated data in a scatter plot. Use different colours to represent the class labels.

2. In R give the command "library("class")". Then you can use the function "knn" to find the k nearest neighbour classifier for each row in a matrix. Generate some samples from the $N_2([0, 5, 0.5]^T, I)$ distribution and see the result for $k = 1$ and for $k = 5$.

Let now

$$\widehat{y}(x_0; (x_1, y_1), \ldots, (x_n, y_n)) \tag{1}$$

be the k nearest classifier rule result for x_0 when using the training data $(x_1, y_1), \ldots, (x_n, y_n))$. The misclassification rate for this classifier is defined as

$$\text{err}(z, F) = P_F(y_0 \neq \widehat{y}(x_0; (x_1, y_1), \ldots, (x_n, y_n)), \tag{2}$$

where $z = ((x_1, y_1), \ldots, (x_n, y_n))$ is the training data and F is the distribution of the test data (x_0, y_0). The misclassification rate for this classification rule is then defined as

$$\theta = \text{E}_F[\text{err}(z, F)],$$

when we assume the training data and the test data to come from the same distribution F. A naïve estimate of the misclassification rate is the so called "apparent" misclassification rate,

$$\widehat{\theta} = \text{err}(z, \widehat{F}) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \widehat{y}(x_i; (x_1, y_1), \ldots, (x_n, y_n))), \tag{3}$$

where $\widehat{F}$ is the empirical distribution for the training data $(x_1, y_1), \ldots, (x_n, y_n)$.

3. Compute the apparent misclassification rate for your generated data for the k nearest neighbour classifier when $k = 1$ and when $k = 5$.

4. Using the notation introduced above, write up an expression for the bias of the apparent misclassification rate. Use this to define the ideal bootstrap estimator for the bias of the misclassification rate.

5. Use simulation (i.e. bootstrapping) to estimate the ideal bootstrap estimate of the bias of the apparent misclassification rate and compute the bias corrected estimate of the misclassification rate.

6. Since your data are simulated you can also estimate the misclassification rate by simulating from the true distribution. Do this, and compare the result by the bias corrected estimate obtained above.