SB2b HT 2017 - Problem Sheet 4

1. The receiver operating characteristic (ROC) curve plots the sensitivity against the specificity of a binary classifier as the threshold for discrimination is varied.

   Let the data space be $\mathbb{R}$, and denote the class-conditional densities with $g_0(x)$ and $g_1(x)$ for $x \in \mathbb{R}$ and for the two classes 0 and 1. Consider a classifier that classifies x as class 1 if $x \geq c$, where threshold c varies from $-\infty$ to $+\infty$.

   (a) Give expressions for the (population versions of) specificity and sensitivity of this classifier.

   (b) Show that the AUC corresponds to the probability that $X_1 > X_0$, where data items X_1 and X_0 are independent and come from classes 1 and 0 respectively.

2. (**1-NN risk in binary classification**) Let $\{(X_i, Y_i)\}_{i=1}^n$ be a training dataset where $X_i \in \mathbb{R}^p$ and $Y_i \in \{0, 1\}$. We denote by $g_k(x)$ the conditional density of X given $Y = k$ and assume that $g_k(x) > 0$ for all $x \in \mathbb{R}^p$, and the class probabilities as $\pi_k = \mathbb{P}(Y = k)$. We further denote $q(x) = \mathbb{P}(Y = 1|X = x)$.

   (a) Consider the Bayes classifier (minimizing risk w.r.t. 0/1 loss $\mathbf{1}\{f(X) \neq Y\}$):

   $$f_{\text{Bayes}}(x) = \underset{k \in \{0,1\}}{\arg\max}\, \pi_k g_k(x).$$

   Write the conditional expected loss $\mathbb{P}[f(X) \neq Y|X = x]$ at a given test point $X = x$ in terms of $q(x)$. [The resulting expression should depend *only* on $q(x)$].

   (b) The 1-nearest neighbour (1-NN) classifier assigns to a test data point x the label of the closest training point; i.e. $f_{1\text{NN}}(x) = y$ (class of nearest neighbour in the training set). Given some test point $X = x$ and its nearest neighbour $X' = x'$, what is the conditional expected loss $\mathbb{P}[f_{1\text{NN}}(X) \neq Y|X = x, X' = x']$ of the 1-NN classifier in terms of $q(x), q(x')$?

   (c) As the number of training examples goes to infinity, i.e. $n \to \infty$, assume that the training data fills the space such that $q(x') \to q(x)$, $\forall x$. Give the limit (as $n \to \infty$) of $\mathbb{P}[f_{1\text{NN}}(X) \neq Y|X = x]$. If we denote by $R_{\text{Bayes}} = \mathbb{P}\left[Y \neq f_{\text{Bayes}}(X)\right]$ and $R_{1\text{NN}} = \mathbb{P}[Y \neq f_{1\text{NN}}(X)]$, show that for sufficiently large n

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

3. Consider a binary classification problem with $\mathcal{Y} = \{1, 2\}$. We are at a node t in a decision tree and would like to split it based on Gini impurity. Consider a categorical attribute A with L levels, i.e., $x^{(A)} \in \{a_1, a_2, ..., a_L\}$. For a generic example (X_i, Y_i) reaching node t, denote:

   $$\begin{aligned} p_k &= \mathbb{P}(Y_i = k),\ k = 1, 2, \\ q_\ell &= \mathbb{P}\left(X_i^{(A)} = a_\ell\right),\ \ell = 1, \ldots, L, \\ p_{k|\ell} &= \mathbb{P}\left(Y_i = k|X_i^{(A)} = a_\ell\right),\ k = 1, 2, \text{ and } \ell = 1, \ldots, L. \end{aligned}$$

   Thus, the population Gini impurity is given by $2p_1(1 - p_1)$. Further, assume $N = n$ examples

$\{(X_i, Y_i)\}_{i=1}^n$ have reached the node t, and denote

$$\begin{aligned} N^k &= |\{i:\ Y_i = k\}|,\ k = 1, 2, \\ N_\ell &= \left|\left\{i:\ X_i^{(A)} = a_\ell\right\}\right|,\ \ell = 1, \ldots, L, \\ N_{k|\ell} &= \left|\left\{i:\ Y_i = k \text{ and } X_i^{(A)} = a_\ell\right\}\right|,\ k = 1, 2, \text{ and } \ell = 1, \ldots, L. \end{aligned}$$

(a) Assuming data vectors reaching node t are independent, explain why $N_\ell|N = n$, $N^k|N = n$ and $N_{k|\ell}|N_\ell = n_\ell$ have respectively multinomial, binomial and binomial distributions with parameters q_ℓ, p_k and $p_{k|\ell}$.

(b) If we split using attribute A (and are not using dummy variables) we will have an L-way split and the resulting impurity change will be

$$\Delta_{\text{Gini}} = 2p_1(1 - p_1) - 2\sum_{\ell=1}^{L} q_\ell p_{1|\ell}(1 - p_{1|\ell})$$

The parameters p_k, q_ℓ and $p_{k|\ell}$ are unknown, however. The Gini impurity estimate $\hat{\Delta}_{\text{Gini}}$ is thus computed using the plug-in estimates $\hat{p}_k = N^k/N$, $\hat{q}_\ell = N_\ell/N$ and $\hat{p}_{k|\ell} = N_{k|\ell}/N_\ell$ respectively. Calculate the expected estimated impurity change $\mathbb{E}[\hat{\Delta}_{\text{Gini}}|N = n]$ between node t and its L child-nodes, conditioned on $N = n$ data vectors reaching node t.

(c) Suppose the attribute-levels are actually uninformative about the class label, so that $p_{k|\ell} = p_k$. Show that, conditioned on $N = n$, the expected estimated Gini impurity change is then equal

$$2p_1(1 - p_1)(L - 1)/n.$$

(d) Is this attribute selection criterion biased in favor of attributes with more levels?

4. Download the wine dataset from
`https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data`
and load it using `read.table("wine.data",sep=",")`. Description of the dataset is given at `https://archive.ics.uci.edu/ml/datasets/Wine`.

(a) Train a classification tree using `rpart`. Give the plots of the tree as well as of the cross-validation results in `rpart` object using `plotcp`.

(b) Now produce a Random Forest fit, calculating the out-of-bag estimation error and compare with the tree analysis. You could start like:

```
library(randomForest)
rf <- randomForest(td[,2:14],td[,1],importance=TRUE)
print(rf)
```

Use `tuneRF` to find an optimal value of `mtry`, the number of attribute candidates at each split. Use `varImpPlot` to determine what are the most important variables.

5. In this question we will take a Bayesian approach to learning decision trees. Assume that we have a binary classification problem, with classes $\{0, 1\}$ and we have n data items $(x_i, y_i)_{i=1}^n$.

   Recall the greedy tree growing heuristics for decision trees: we start with the root, for each leaf node of the tree we find an optimal feature j and split point v, according to some criteria, to split the node, and recurse on both sides.

   (a) Consider a greedy model selection procedure for determining the structure of T:

      i. We start with a trivial tree with a single node.

      ii. At each iteration we consider expanding a leaf node m of the tree by creating a split at feature j, value v. This produces a tree T' with two more nodes than T, both children of node m.

      iii. We compute the marginal probability of $\mathbf{Y}$ under T', for each j and v, and find the split producing the highest marginal probability.

      iv. If the marginal probability of the resulting T' is larger than T, we split the node, otherwise we consider expanding other nodes.

      v. We stop once all leaf nodes of the current tree T have been considered for expansion, but all lead to trees T' with lower marginal probability than T.

      Calculate the marginal probability $p(\mathbf{Y}|\mathbf{X}, T')$ of the responses given the data vectors under tree T'.

   (b) Explain how the ratio of marginal probabilities under T' and T (the so-called **Bayes factor**) simplifies to a function which depends only on the data items under region $\mathcal{R}_m$.

   (c) For each j, explain why the marginal probability under T' is a piecewise constant function of v.

   (d) Describe an algorithm that can determine the optimal split of node m, with computational cost $p \times N_m$, where N_m is the number of data items in region m, and p the number of features.

Optional

6. A factor analysis model is described as follows: each latent variable $y_i \in \mathbb{R}^d$ and modelled with a standard multivariate normal, and:

$$y_i \sim \mathcal{N}(0, I)$$
$$x_i|y_i \sim \mathcal{N}(Ay_i, D)$$

   where the parameters are D, a $p \times p$ diagonal matrix with positive entries on the diagonal, and $A \in \mathbb{R}^{p \times d}$.

   We will derive the EM algorithm for the model.

   (a) Derive the E-step, i.e., show that the posterior of y_i given x_i for fixed A and D is

$$\mathcal{N}(A^\top (D + AA^\top)^{-1} x_i, I - A^\top (D + AA^\top)^{-1} A).$$

   You may find the Woodbury matrix inversion lemma useful:

$$(S + UTV)^{-1} = S^{-1} - S^{-1}U(T^{-1} + VS^{-1}U)^{-1}VS^{-1}$$

(b) Derive the M step for D.

(c) Derive the M step for A.