SDMML HT 2015 - Part C Problem Sheet 6

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. We are deciding how to split a node in a binary classification tree. The node contains 300 data vectors of class 1 and 100 of class 2. Write this as $(300, 100)$. We can split the node either as (a) $(200, 100)$, $(100, 0)$, or (b) $(150, 50)$, $(150, 50)$. Compute the change in the impurity measure when using the misclassification rate and when using the Gini impurity for two splits. Which split is preferred?

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:

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

   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

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

   (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.

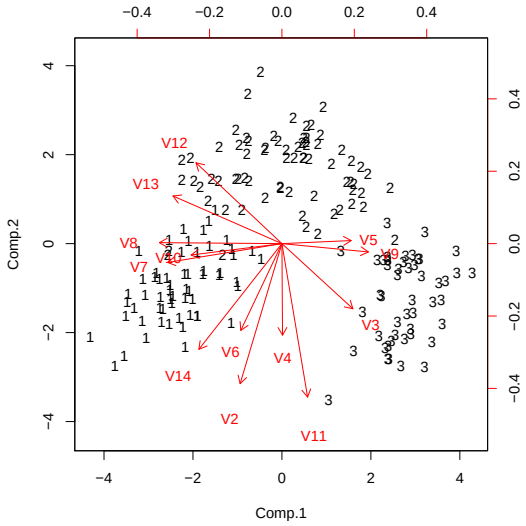
(c) Suppose the attribute-levels are actually uninformative about the class label, so that $p_{k|\ell} = p_k$. Show that, conditioned on $N = n$, the expected estimated Gini impurity change is then equal

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

(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) Make a biplot using the `scale=0` option, and then use the `xlabs=as.numeric(td$Type)` option in `biplot` to label points by their `$Type`. The output should look like:



(b) Now train a classification tree using `rpart`, and relate the decision rule discovered there to the projections of the original variable axes displayed in the biplot. Give the plots of the tree as well as of the cross-validation results in `rpart` object using `plotcp`.

(c) 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. Suppose we have a model $p(x, y|\theta)$ where X is an observed variable and Y unobserved. We would like to take a Bayesian approach to learning, treating the parameter Θ to be random as well, with prior $p(\theta)$.

   (a) Suppose that $Q(y, \theta)$ is a distribution over both Y and Θ. Explain why the following is a lower bound on $p(x)$:

   $$\mathcal{F}(Q) = \mathbb{E}_Q[\log p(x, y, \theta) - \log Q(y, \theta)]$$

   (b) Show that the optimal $Q(y, \theta)$ is simply the posterior $p(y, \theta|x)$.

   (c) Typically the posterior is intractable. Consider a factorized distribution $Q(y, \theta) = Q_Y(y)Q_\Theta(\theta)$. In other words we assume that Y and Θ are independent. Derive the optimal Q_Y given a Q_Θ, and hence describe an algorithm to optimize $\mathcal{F}(Q)$ subject to assumption of independence between Y and Q.