SDMML HT 2016 - MSc 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. $f_{\text{1NN}}(x) = y$ (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 $\mathbb{P}[f_{\text{1NN}}(X) \neq Y|X = x, X' = x']$ 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 $\mathbb{P}[f_{\text{1NN}}(X) \neq Y|X = x]$. If we denote by $R_{\text{Bayes}} = \mathbb{P}\left[Y \neq f_{\text{Bayes}}(X)\right]$ and $R_{\text{1NN}} = \mathbb{P}[Y \neq f_{\text{1NN}}(X)]$, show that for sufficiently large n

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

3. Recall the definition of a one-hidden layer neural network for binary classification in the lectures. The objective function is L_2-regularized log loss:

   $$J = -\sum_{i=1}^{n} y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i) + \frac{\lambda}{2}\left(\sum_{jl}(w_{jl}^h)^2 + \sum_{l}(w_l^o)^2\right)$$

   and the network definition is:

   $$\hat{y}_i = s\left(b^o + \sum_{l=1}^{m} w_l^o h_{il}\right), \qquad h_{il} = s\left(b_l^h + \sum_{j=1}^{p} w_{jl}^h x_{ij}\right),$$

   with transfer function $s(a) = \frac{1}{1+e^{-a}}$.

(a) Verify that the derivatives needed for gradient descent are:

$$\frac{\partial J}{\partial w_l^o} = \lambda w_l^o + \sum_{i=1}^{n} (\hat{y}_i - y_i) h_{il},$$
$$\frac{\partial J}{\partial w_{jl}^h} = \lambda w_{jl}^h + \sum_{i=1}^{n} (\hat{y}_i - y_i) w_l^o h_{il}(1 - h_{il}) x_{ij}.$$

(b) Suppose instead that you have a neural network for binary classification with L hidden layers, each hidden layer having m neurons with logistic transfer function. Give the parameterization for each layer, and derive the backpropagation algorithm to compute the derivatives of the objective with respect to the parameters. For simplicity, you can ignore bias terms.

4. In this question you will investigate fitting neural networks using the `nnet` library in R. We will train a neural network to classify handwritten digits 0-9. Download files `usps_trainx.data`, `usps_trainy.data`, `usps_testx.data`, `usps_testy.data` from `http://www.stats.ox.ac.uk/~sejdinov/sdmml/data/`.
Each handwritten digit is 16×16 in size, so that data vectors are $p = 256$ dimensional and each entry (pixel) takes integer values 0-255. There are 2000 digits (200 digits of each class) in each of the training set and test set. You can view the digits with

```
image(matrix(as.matrix(trainx[500,]),16,16),col=grey(seq(0,1,length=256)))
trainy[500,]
```

Download the R script `nnetusps.R` from the course webpage. The script trains a 1-hidden layer neural network with $S = 10$ hidden units for $T = 10$ iterations, reports the training and test errors, runs it for another 10 iterations, and reports the new training and test errors. To make computations quicker, the script down-samples the training set to 200 cases, by using only one out of every 10 training cases. You will find the documentation for the `nnet` library useful: `http://cran.r-project.org/web/packages/nnet/nnet.pdf`.

(a) Edit the script to report the training and test error after every iteration of training the network. Use networks of size $S = 10$ and up to $T = 100$ iterations. Plot the training and test errors as functions of the number of iterations. Discuss the results and the figure.

(b) Edit the script to vary the size of the network, reporting the training and test errors for network sizes $S = 1, 2, 3, 4, 5, 10, 20, 40$. Use $T = 25$ iterations. Plot these as a function of the network size. Discuss the results and the figure.

5. 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

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

(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.

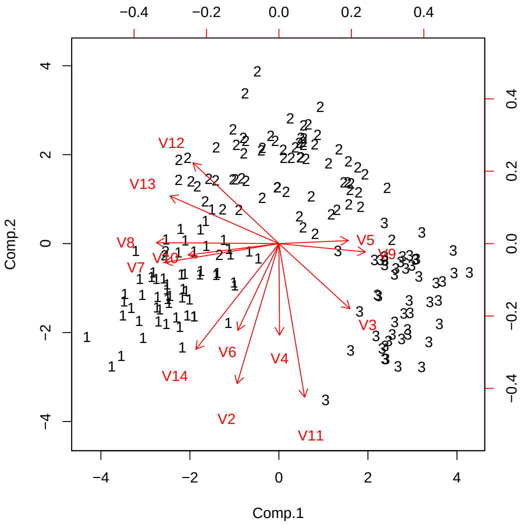
(c) Suppose the attribute-levels are actually uninformative about the class label, so that $p_{k|\ell} = p_k$. Show that, conditioned on $N = n$, the expected estimated Gini impurity change is then equal

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

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

6. 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.

Optional

7. A **mixture of experts** is an ensemble model in which a number of experts "compete" to predict a label.

Consider a regression problem with dataset $\{(x_i, y_i)\}_{i=1}^n$ and $y_i \in \mathbb{R}$. We have E experts, each associated with a parametrized regression function $f_j(x; \theta_j)$, for $j = 1, \ldots, E$ (for example, each expert could be a neural network).

(a) A simple mixture of experts model uses as objective function

$$J(\pi, \sigma^2, (\theta_j)_{j=1}^E) = \sum_{i=1}^n \log \sum_{j=1}^E \pi_j e^{-\frac{1}{2\sigma^2}\|f_j(x_i;\theta_j)-y_i)\|^2}$$

where $\pi = (\pi_1, \ldots, \pi_E)$ are mixing proportions and σ^2 is a parameter.

Relate the objective function to the log-likelihood of a mixture model where each component is a conditional distribution of Y given $X = x$.

(b) Differentiate the objective function with respect to θ_j. Introduce a latent variable z_i, indicating which expert is responsible for predicting y_i, and interpret $\frac{\partial J}{\partial \theta_j}$ in the context of the corresponding EM algorithm. In this context, one needs to use the generalized EM algorithm, where in the M-step gradient descent is used to update the expert parameters θ_j.

(c) A mixture of experts allows each expert to specialize in predicting the response in a certain part of the data space, with the overall model having better predictions than any one of the experts.

However to encourage this specialization, it is useful also for the mixing proportions to depend on the data vectors x, i.e. to model $\pi_j(x;\phi)$ as a function of x with parameters ϕ. The idea is that this **gating network** controls where each expert specializes. To ensure $\sum_{j=1}^{E} \pi_j(x;\phi) = 1$, we can use the softmax nonlinearity:

$$\pi_j(x;\phi) = \frac{\exp(h_j(x;\phi_j))}{\sum_{\ell=1}^{E} \exp(h_\ell(x;\phi_\ell))}$$

where $h_j(x;\phi_j)$ are parameterized functions for the gating network.

The previous generalized EM algorithm extends to this scenario easily. Describe what changes have to be made, and derive a gradient descent learning update for ϕ_j.