Mini-Project I
Visualizing Random Forest

Alexander Reinthal 880726-4851 reinthal@student.chalmers.se
Magnus Lindström 910511-1752 guslinmagc@student.gu.se

April 2017

1 Dataset

We chose to work with the dataset `DoctorsAUS` from the Ecdat package. This dataset contains patient information gathered from doctor visits in Australia. It contains 5190 data points with 15 variables ($n = 5190$ and $p = 15$). The features are found in Tab. 1.

Table of Features		
Feature Name	Representation	Description
sex	$\{0, 1\}$	gender of the person
age	$\mathbb{R}$	age of the person
income	$\mathbb{R}$ mult. 10^4	income of person
insurance	categorical	The insurance of the person
actdays	$\mathbb{N}$	number of days of reduced activity in past 2 weeks due to illness or injury
hscore	$\{1, ..., 12\}$	health score
chcond	categorical	has chronic condition?
doctorco	$\mathbb{N}$	consultations with a doctor or specialist in the past 2 weeks
nondocco	$\mathbb{N}$	consultations with a non-doctor health professionals in the past 2 weeks
hospadmi	$\mathbb{N}$	admissions to hospital in the past 2 weeks
hospdays	$\mathbb{N}$	days in hospital past year
medecine	$\mathbb{N}$	medicines used the past 2 days
prescrib	$\mathbb{N}$	prescribed medicines used the past 2 days
nonpresc	$\mathbb{N}$	non-prescribed medicines used the past 2 days

We chose to make a random forest binary classifier of the feature **sex** using all the other features of the dataset.

2 Visualisations of Random Forests

There are many different ways of visualising the performance of a random forest. Below, a few important measures are listed.

Partial Plots

Partial plots are a way of visualising the importance of each variable when it comes to classifying the response variable. A typical plot is shown in Fig. 1. The y-axis shows a measure of the votes in favor of predicting a class with values ranging from -1 to 1. 1 means that for 100% of the data points, the classifier voted in favor of the class we're trying to predict, for the value of the predictor variable specified on the x-axis. Conversely, -1 means that every data point was classified as another class than the one we're trying to predict. Thus, 0 is the middle value when 50% of votes are cast in favor of either decision. For the particluar feature shown in Fig. 1, the percentage of votes in favor of one class drops as the income decreases.

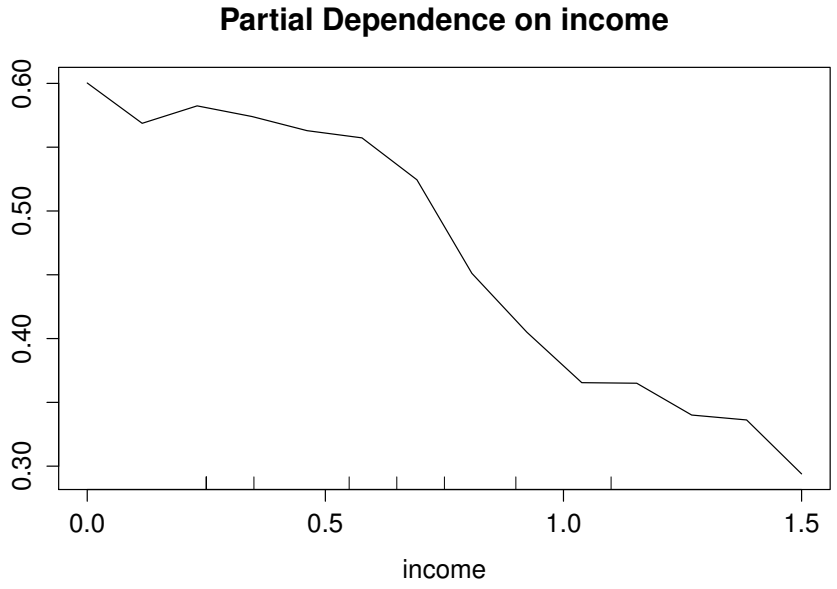

Figure 1: The partial dependency of the `actdays` feature.

Out of Bag-Error Plot

The individual trees of random forests are build with data from a bootstrap sample of observations $\mathbf{Z}^* = (x_i, y_i), i = 1, \ldots, N$ from the training data. When

the forest is fully grown one lets each observation (x_j, y_j) be classified by a majority vote of the trees that did not use this specific observation in the creation of the tree. The out of bag-error rate is then constructed from these classifications. The number of trees generated for a forest will impact the performance and speed of classification/regression. The more trees generated, the better classification and regression results we expect, but there is probably a point where the OOB error rate stops decreasing and settles around a value. It is thus interesting to see the OOB error progression as a function of increasing forest size, as shown in Fig. 2. From this figure, it is clear that there is no point in increasing the forest size after around 200 trees.

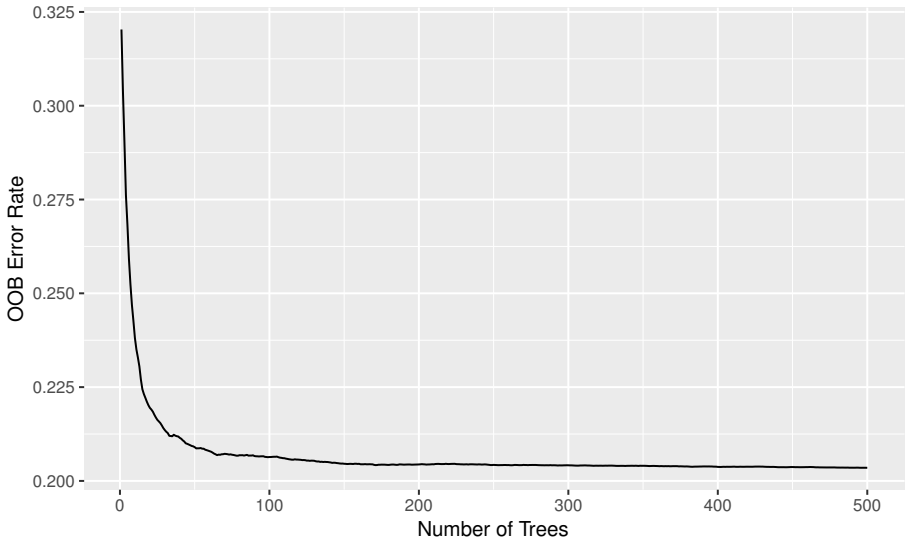

Figure 2: The OOB error rate as a function of forest size.

Proximity Plot

When constructing a random forest and classifying OOB observations, one can obtain a measure of proximity for the training set. Every pair x_i, x_j of data points are classified by the trees for which they are OOB, and whenever two data points share a terminal node in a tree, they are deemed to lie "close" to one another and their so-called "proximity value" is increased by one. After having created a proximity matrix for the data points, the points are then represented in two dimensions using multidimensional scaling. This is called a *proximity plot*. In the proximity plot, the proportion of the Euclidean distances between the points stay roughly the same as the proportion of the distances between the data points, as measured by the random forest. The proximity matrix for the data set chosen in this report is shown in Fig. 3.

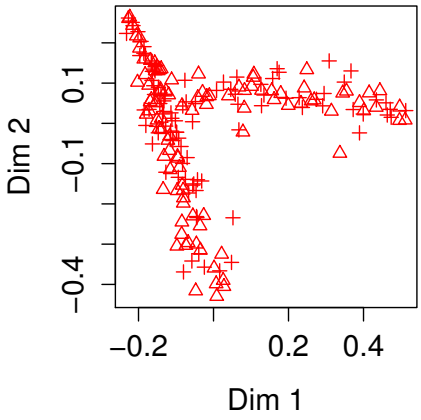

Figure 3: The figure above shows a proximity plot on the classification RF on the feature **sex** on 200 random samples of the dataset DoctorsAUS. The algorithm used was R's randomForest. Samples that are deemed close by the classifier are clustered together. A higher separation of triangles and crosses would imply a lower OOBerror.

ROC Curve

The ROC curve illustrates how good the classifier is as distinguishing between the two classes. The x-axis measures false positive rate and the y-axis the true positive rate. Each point on the curve is for a different setting of the classifier, in this case the number of "tree votes" in favor on one class for the prediction to be that outcome. The bigger the area under the curve is the better our classifier is. Since the maximum value of the AUC – area under curve – is 1 and the area in Fig. 4 is ≈ 0.78, the classifier performs okay, not perfect.

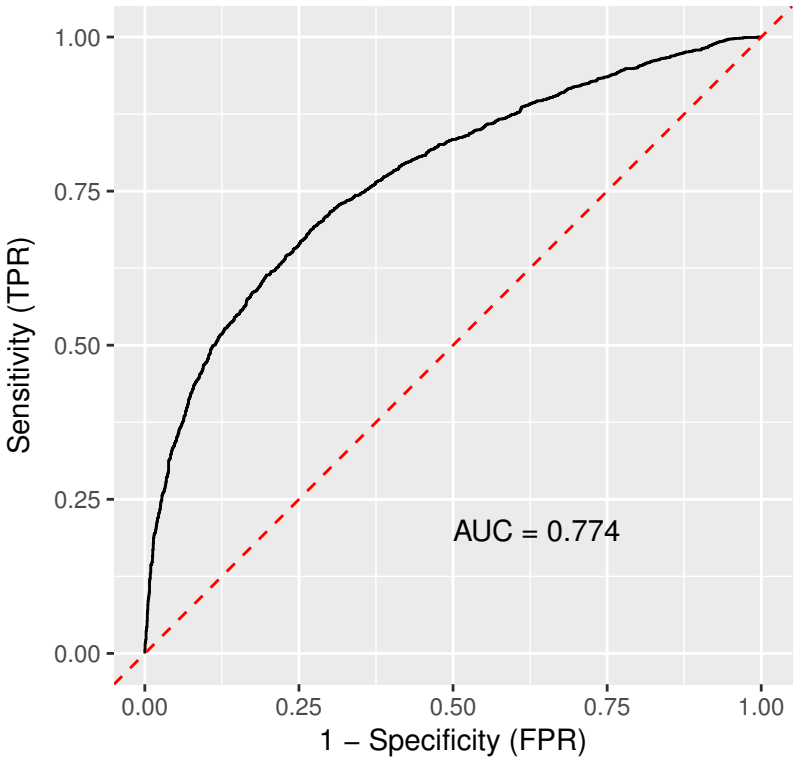

Figure 4: This is a Roccurve.

Variable Importance

In order to gain an understanding of the relative importance of the variables when it comes to classification, one can quantify the *variable importance*. For random forests, it is common to do the following: After calculating the OOB error rate the first time, variable j of the OOB samples are randomly permuted, which scrambles the data somewhat, and the points are once more classified. The increase in OOB error rate should correspond to the importance of variable j. This is performed for each of the variables, which creates a measure of the relative importance of the variables. The results of this procedure for our data is shown in Fig. 5. Apparently, the income, age and the number of prescribed medicines used the past two days are important in distinguishing between male and female patients.

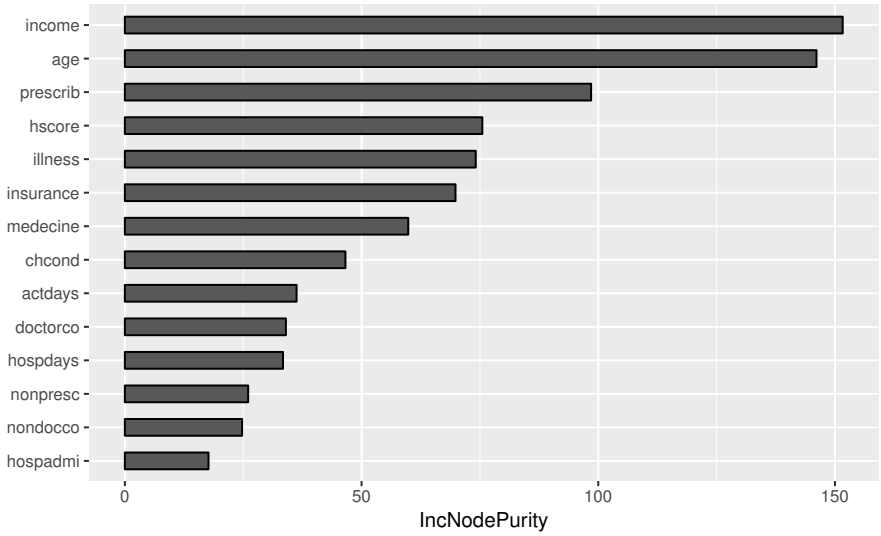

Figure 5: Caption

3 Analysis of results

The proximity plot and the ROC plot confirms each others results - that the classifier had a hard time separating the two classes (e.g men and women). As for variable importance it interesting to see that income is the most important feature to distinguish between male and female patients. Age however seems a bit odd - maybe the samples were not balanced in the age brackets.