

AUC and ensemble methods with R

Objective: The objective of today's exercise is to understand (i) how the ensemble methods bagging and boosting is used to improve the performance of classifiers, (ii) how the receiver operating characteristic (ROC) is used to evaluate the performance of two-class classification problems, and (iii) how artificial neural network and logistic regression can be generalized to multi-class classification.

Piazza discussion forum: You can get help by asking questions on Piazza: piazza.com/dtu.dk/fall2019/october2019

Software installation: Extract the R toolbox from the Dropbox folder . Start R and go to the `<base-dir>/02450Toolbox_R/` directory by setting the working directory through `setwd(<base-dir>/02450Toolbox_R/)` and run `source('setup.R')`. Remember the purpose of the exercises is not to re-write the code from scratch but to work with the scripts provided in the directory `<base-dir>/02450Toolbox_R/Scripts/`

Representation of data in R:

	R var.	Type	Size	Description
	X	Numeric	$N \times M$	Data matrix: The rows correspond to N data objects, each of which contains M attributes.
	attributeNames	Cell array	$M \times 1$	Attribute names: Name (string) for each of the M attributes.
	N	Numeric	Scalar	Number of data objects.
	M	Numeric	Scalar	Number of attributes.
Regression	y	Numeric	$N \times 1$	Dependent variable (output): For each data object, y contains an output value that we wish to predict.
Classification	y	Numeric	$N \times 1$	Class index: For each data object, y contains a class index, $y_n \in \{0, 1, \dots, C - 1\}$, where C is the total number of classes.
	classNames	Cell array	$C \times 1$	Class names: Name (string) for each of the C classes.
	C	Numeric	Scalar	Number of classes.
Cross-validation				All variables mentioned above appended with <code>_train</code> or <code>_test</code> represent the corresponding variable for the training or test set.
	*_train	—	—	Training data.
	*_test	—	—	Test data.

9.1 The receiver operating characteristic (ROC)

The receiver operating characteristic (ROC) is commonly used to evaluate and compare the performance of two-class classifiers, where one class is denoted “positive” and the other is denoted “negative.” The ROC is a graphical approach for displaying the tradeoff between the true positive rate (y -axis) and the false positive rate (x -axis). For classifiers such as logistic regression and artificial neural networks that estimate the class labels by thresholding a continuous output variable, an ROC curve can be plotted by varying the threshold value.

- 9.1.1 Inspect and run the script `ex9_1_1.R`. The script loads the wine data (`Data/wine2`) into R by the command `dat <- readMat('Data/wine2.mat')`, after the library `R.matlab` has been loaded. Notice how the data is divided into 50% for training and 50% for test. Fit a logistic regression model to the training set to classify the wines as red or white. Consider the red wines as “positive” and white wines as “negative.” Notice how the script makes a plot of the ROC curve showing that the AUC is around 0.99.

Script details:

- *You have fitted a logistic regression model to the wine data before in exercise 5.2.6.*
- *As before, use `cvFolds` to partition the data into training and test sets.*
- *As usual, use the functions `glm` and `predict` to fit and evaluate the logistic regression model.*
- *There is a function in the course toolbox called `rocplot` that can make the ROC plot and compute the AUC score for you. Open `rocplot` in the folder `Tools` in the toolbox and read the help in the comments at the beginning of the function to learn how to use it.*

- 9.1.2 Consult the script `ex9_1_2.R`. The experiment in exercise 9.1.1 is repeated, but this time it is examined how well the type of wine can be classified using only the “Alcohol” attribute. Explain how it can be seen that the alcohol contents of the wine is not very useful for classifying wine as red or white.

Discussion:

- ◇ You have showed that using only the single attribute “Alcohol” to classify the wine as red or white performs worse than using all attributes. When using logistic regression, is it always best to use as many attributes as possible for the classification?

9.2 Ensemble methods

In this part of the exercise we will consider ensemble methods to improve the classification performance. In particular we will consider bagging and boosting methods.

In bagging we randomly sample with replacement the same number of samples as the size of the training data. This is also denoted bootstrapping and it can be

shown that on average each bootstrap sample will contain approximately 63% of the samples in the data.

In boosting we adaptively change the distribution that we sample the training examples from so that the classifiers will focus on examples that are hard to classify. A particularly well known boosting method is “*AdaBoost*” which is described in section 15.4 of the lecture notes *Introduction to Machine Learning and Data Mining*.

In this part of the exercise we will work with a synthetic data set which has two classes and two attributes, x_1 and x_2 . This data set cannot be classified correctly with any linear classifier, such as logistic regression or an artificial neural network with one hidden unit, as long as only x_1 and x_2 are used as features.

- 9.2.1 Load the artificial dataset (Data/synth5 file) by first loading the package `R.matlab` and then typing `dat = read.mat('Data/synth5.mat')`. Inspect the data by making a scatterplot. Why can this data set not be classified correctly using logistic regression?

Inspect and run the script `ex9_2_1.R`. The script fits an ensemble of logistic regression models to the data, using the bootstrap aggregation (bagging) algorithm. Use $L = 100$ bootstrap samples. This requires creating 100 bootstrapped training sets, fit a logistic regression model to each, and combine the results of their outputs to make the final classification. Explain how the error rate is computed (on the training set) and how the decision boundary is plotted.

Script details:

- To generate the bootstrap sample, you need to draw N data objects with replacement from the data set.
- You can use the function `discreternd` to generate random numbers from a discrete distribution. You can write something like `i=discreternd(rep(1, times=N),N); X_train=X[i,]; y_train=y[i];` to make a bootstrap sample.
- The function `lapply` can be used to evaluate the whole ensemble of fitted models in one go. If you collect the outputs of `glm` in a list `w_est`, the command `lapply(w_est, FUN=function(model, newd, typ) predict(model, newdata=newd, type=typ), newd=Xdf, typ='response')` will return a vector of length $N \cdot L$ probabilities. Make this into an $N \times L$ matrix of probabilities (for each data object times rounds of bagging) using `p <- matrix(unlist(plist), nrow=N, byrow=FALSE)`.
- To make the final classification, take a majority vote amongst the classifiers in the ensemble.
- To plot the decision boundary, you can use the function `dbplot` in the toolbox. It requires as an input a function that evaluates the classifier, in this case the ensemble of classifiers. You can call it e.g. like this:

```
predictionFunction <- function(Xgriddf, model, typ){ plist <- lapply(w_est, FUN=funct
```

Bagging is known to be most effective for non-linear classifiers. Show that bagging only leads to a limited performance improvement for the logistic regression classifier.

Try also the data set in `Data/synth6` which is the one used as an example in the lecture.

- 9.2.2 In the script `ex9_2_2.R`, the script `ex9_2_1.R` has been modified so that boosting is used instead of bagging. The script fits an ensemble of logistic regression models to the data, using the AdaBoost algorithm. Notice the script uses $L = 100$ rounds of boosting. This requires creating a randomly chosen training set, fit a logistic regression model to it, evaluate its performance and update the weights accordingly, and compute a classifier importance. This process is repeated $L = 100$ times, and ultimately the trained ensemble of classifiers is combined to make a final classification. Compute the error rate (on the training set) and make a plot of the decision boundary.

Script details:

- *Read the hints to the previous exercise.*
- *Note that you will need two sets of weights in the algorithm. One is a weight for each of the N data objects, that is adapted when the boosting algorithm proceeds (we could call these **weights**). The other is the importance weights for the L trained classifiers (we could call these **alpha**).*
- *You can use the function `discreternd` to generate random numbers from a discrete distribution. You can write something like*
`i=discreternd(weights,N); X_train=X[i,]; y_train=y[i];` *to make a sample according to the probabilities in **weights**.*
- *To make the final classification, take a weighted majority vote amongst the classifiers in the ensemble, e.g., by computing*
`y_est <- rowSums(t(apply(p>0.5, 1, function(x, weights) x*weights, weights=alpha)))>0.`
*where **p** is computed as in the hint to the previous exercise and **alpha** is normalized such that it sums to one.*

Show that, if you use enough rounds of boosting, the data set can be perfectly classified.

Try also the data set in `Data/synth6` which is the one used as an example in the lecture.

Let us return to the bagging algorithm, which we found to be of little use for logistic regression. Now, we will try the algorithm on a non-linear classifier: the decision tree. Bagging applied to decision trees is often called “random forests”. R has implemented bagging for random trees, see the function `randomForest` in the package `randomForest`.

- 9.2.3 The data set `Data/synth5` we have used so far can trivially be fitted using a decision tree. Can you explain why? We will consider a different data set, which you can load from the file `Data/synth7`. Inspect and run the script `ex9_2_3.R`. Explore the data set and explain in what sense this is more challenging for a decision tree.

Notice the script fits a random forest (an ensemble of bagged decision trees) to the data using $L = 100$ rounds of bagging. Compute the error rate (on the training set) and make a plot of the decision boundary.

Script details:

- Type `?randomForest` to learn about the function in R that fits a random forest. Note, that this function will do all the work for you including making the L bootstrap samples and fitting the models. If you e.g. write `B=randomForest(fmla, data=Xdf, ntree=L)` R will fit an ensemble of L decision trees according to the formula `fmla`.
- The function `predict` can be used to make predictions. Type `?predict.randomForest` to learn how to use it. Note that the function returns the predicted class labels in a vector.

For comparison, you can try also to fit a regular decision tree (without bagging or pruning). Does bagging appear to improve on the classification, and if so, in what sense?

9.3 Tasks for the report

Continue working on the tasks for the reports as described in the previous exercise note.

References