Part 1 - Getting started.

In this part, we use Pima Indians Diabetes Database (file "diabetes.arff"), which can be found in the directory "$WEKAHOME/data". Its original owner is National Institute of Diabetes and Digestive and Kidney Diseases. The data set has 768 instances, which consist of nine attributes (eight + class). The attributes are:

1. Number of times pregnant

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test

3. Diastolic blood pressure (mm Hg)

4. Triceps skin fold thickness (mm)

5. 2-Hour serum insulin (mu U/ml)

6. Body mass index (weight in kg/(height in m)^2)

7. Diabetes pedigree function

8. Age (years)

9. Class variable (0 or 1). Class value 0 is interpreted as "tested negative for diabetes" and class value 1 as "tested positive for diabetes".

What to report?

Answer the following questions and report the answers as well as the required Weka outputs in your exercise report.

Questions

1. Cross-validation.

1-b)
Cross-validation, sometimes called rotation estimation,[1][2][3] is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset).[4] The goal of cross validation is to define a dataset to "test" the model in the training phase (i.e., the validation dataset), in order to limit problems like overfitting, give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem), etc.

(a) What does it mean to perform 10-fold cross-validation?

(b) For what purpose is cross-validation used?

2. Start Weka and choose Explorer as GUI. Open the file "diabetes.arff". Let us start with rule based classifier ZeroR. It assigns the most common class value in a training set to any new instances and it can be used as a baseline for evaluating other machine learning schemes.

(a) Run ZeroR classifier using "Use training set" as a test option. Report the number of correctly and incorrectly classified instances.

(b) Run ZeroR classifier using 10-fold cross-validation and report the number of correctly and incorrectly classified instances.

3. J48 tree classifier.

(a) Run J48 classifier using "diabetes.arff" as a training set. Report the number of instances correctly and incorrectly classified.

(b) Visualize the decision tree by right-clicking the result list. By maximizing the window and right-clicking on the background of the visualized tree, the tree can be fit to screen so that all the nodes can be seen better. Include the decision tree in your report.

(c) Study the learned model. Assume an instance having attribute values as follows:

Att1: preg = 6

Att2: plas = 148

Att3: pres = 72

Att4: skin = 35

Att5: insu = 0

Att6: mass = 33.6

Att7: pedi = 0.627

Att8: age = 50.

In which class would the learned model classify the instance?

(d) Run J48 classifier for the "diabetes.arff" using 10-fold cross-validation. Report the number of instances correctly and incorrectly classified.

4. Bagging J48 classifiers.

(a) What does bagging mean?

(b) Run Bagging meta classifier using "diabetes.arff" as a training set. Edit parameters of Bagging classifier so that it uses J48 as a base classifier. Report the number of instances correctly and incorrectly classified.

(c) Run Bagging classifier for the "diabetes.arff" using 10-fold cross-validation. Use J48 as a base classifier. Report the number of instances correctly and incorrectly classified.

5. Comparison of classifiers.

(a) Compare the results of the J48 classifier with the results of the simple ZeroR classifier. How would you describe the performance of the J48 classifier with the used data set?

(b) Compare the results of the J48 classifier with the results obtained by bagging J48 classifiers. How would you describe the performance of the Bagging meta classifier with the used data set?

Last modified: 31.8.2015 by Riitta Kerminen