Weka – k-Nearest Neighbours

Lab4 (in-class): 24 Nov 2016, 10:00-12:00, Turing

Acknowledgements: information, examples and tasks in this lab come from several web sources.

Learning objectives

In this assignment you are going to:

- IBk, ie Weka implentation of k-Nearest Neighbours
- Roc curve, AUC
- Cost/gain analysis
- Paired t-test

Datasets

The **glass** dataset comes from the U.S. Forensic Science Service contains data on six types of glass. Glass is described by its refractive index and the chemical elements that it contains; the the aim is to classify different types of glass based on these features. This dataset is taken from the UCI datasets, which have been collected by the University of California at Irvine and are freely available on the Web. They are often used as a benchmark for comparing data mining algorithms.

The **ACE (**Angeotensin-Converting Enzyme**)** dataset contains protein-ligand binding data. Ligands exhibiting strong binding affinity towards a certain protein being considered as "active" with respect to it. If it is not known about the binding affinity of a ligand towards the protein, such ligand is conventionally considered as "nonactive. The goal of classification models is to be able to predict whether a new ligand will exhibit strong binding activity toward certain protein biotargets. We are interested in it because this dataset is highly unbalanced.

We are already familiar with the spam/junk dataset.

Task 1 – Warming up: Decision Trees [max time: 10min]

Download the glass dataset
< http://stp.lingfil.uu.se/~santinim/ml/2016/Datasets/glass.arff >
Load it into the Explorer interface.

Q1: How many attributes are there in the dataset? What are their names? What is the class attribute?

First of all, run J48, default parameters. Write down accuracy, kappa statistic, P, R, F-score. Annotate also AUC values. Create a small table to tabulate the results (for ex like the one below).

Classifier	Acc	k-stat	Avg. P	Avg. R	Avg. F	Avg. AUC	…	…
J48, default								
IBk, k=1								
IBk, k=5								

Q2: Are happy with the performance of this classifier? What's the matter with "**vehic wind non-float**"? What do the error measure indicate (have you read the Appendix at the end of the previous lab)?

Task 2 –IBk [max time: 10min]

Now, in the classifier frame, click **Choose**, then select the **IBk** method from the **lazy** submenu. The lazy submenu contains a group of methods, in which the training phase is almost omitted – it actually amounts to memorizing all instances from the training set. Instead of it, all main calculations are delayed to the test phase. That is why such methods are sometimes called lazy, instance-based and memory-based. The price for this "laziness" is however rather high – computations at the test phase are very intensive, and that is why such methods work very slowly during prediction, especially for big training sets. So, the abbreviation IBk means that this is an Instance-Based method based on k-neighbours. The default value of k is 1.

Run the classification algorithm lazy→IBk (weka.classifiers.lazy.IBk). Use cross-validation to test its performance, leaving the number of folds at the default value of 10. Recall that you can examine the classifier options in the Generic Object Editor window that pops up when you click the text beside the Choose button. As we said, the default value of the kNN field is 1: This sets the number of neighboring instances to use when classifying.

Q3: What is the performance of IBk? Fill in the fields in the table you created in the previous step. Are you happy with the performance of this classifier? which one is performing better according to the metrics of your table?

Task 3 – Change the k-value [max time: 5min]

Run IBk again, but increase the number of neighboring instances to k = 5 by entering this value in the KNN field. Run the classifier and create the model. Write down the matrics in your table.

Q4: What is the performance of IBk with five neighboring instances (k = 5)? Why? look at value of the AUC: why? Any comments about the performance of the 3 models?

Task 4 – *distanceWeighting* [max time: 5min]

Click with the left mouse button on the word IBk in the Classifier frame. The window for setting options for the k-NN method pops up. Change the option distanceWeighting to **Weight to 1-distance**. The option called "distanceWeighting" gets the distance

weighting method used. Weighting allows neighbors closer to the data point to have more influence on the predicted value. **Weight to 1-distance** corresponds to the parameter –F. This parameter weight neighbours by 1 - their distance (to be used when k > 1). Read here:
< http://weka.sourceforge.net/doc.dev/weka/classifiers/lazy/IBk.html >

Run the classifier and create the model. Write down the results in your table.

Q5: What is the performance of IBk with this option? Why? Any comments about the performance of the 4 models?

Task 6 – Cost/Benefit Analysis [max time: 30min]

Download the ace dataset
< http://stp.lingfil.uu.se/~santinim/ml/2016/Datasets/ace.arff >
Load it into the Explorer interface.

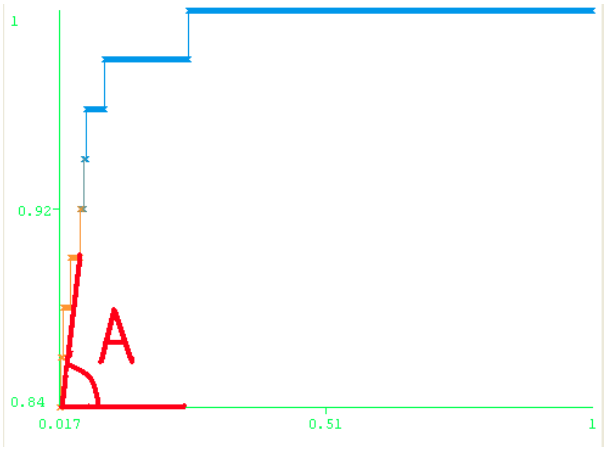
Q6: How many attributes are there in the dataset? What are their names? What is the class attribute?

We know by now that accuracy is not an reliable metric to measure the performance of the classification for unbalanced dataset. The ace dataset is highly unbalanced, since the number of non-active compounds is much larger the number of active compounds.

Run J48 and check accuracy, k and Roc Area. You can see that accuracy is very high (98.8), but the k value is much lower (0.76) and the averaged Roc area is 0.83.

Now run IBk with default parameters. You can see that 1-nearest neighbour model is more roubust that the J48 model: accuracy is 99.2 (which is not so indicative since contains predictions based on chance), but k statistic has increased to 0.87 and the Roc area is 0.95.

Visualize the Roc Curve. As we said, the ROC curve is shown in the Plot frame of the window. The axis X in it corresponds to the false positive rate, whereas its axis Y corresponds to the true positive rate. The color depicts the value of the threshold. The "colder" (closer to the blue) color corresponds to the lower threshold value. All compounds with probability of being "active" exceeding the current threshold are predicted as "active". If such prediction made for a current compound is correct, then the corresponding compound is true positive, otherwise it is false positive. If for some values of the threshold the true positive rate greatly exceeds the false positive rate (which is indicated by the angle A close to 90 degrees, see Fig. below), then the classification model with such threshold can be used to extract selectively "active" compounds from its mixture with the big number of "nonactive" ones in the course of virtual screening. In order to find the optimal value of the threshold (or the optimal part of compounds to be selected in the course of virtual screening), one can perform the cost/benefit analysis.



Q7: What is the ideal shape of the Roc curve? What does the AUC value make you understand about the performance?

Close the Roc Curve. Click the right mouse button on the model type lazy.IBk (k-1) in the Result list frame and select the menu item Cost/Benefit analysis / active. Click on the **Minimize Cost/Benefit** button at the right bottom corner of the window.

Consider attentively the window for the Cost/Benefit Analysis. It consists of several panels. The left part of the window contains the Plot: ThresholdCurve frame with the Threshold Curve (called also the Lift curve). The Threshold curve looks very similar to the ROC curve. In both of them the axis Y corresponds to the true positive rate. However, in contrast to the ROC curve, the axis X in the Threshold curve corresponds to the part of selected instances (the "Sample Size"). *In other words, the Threshold curve depicts the dependence of the part of "active" compounds retrieved in the course of virtual screening upon the part of compounds selected from the whole dataset used for screening.* Remember that only those compounds are selected in the course of virtual screening, for which the estimated probability of being "active" exceeds the chosen threshold. The value of the threshold can be modified interactively by moving the slider in the Threshold frame of the Cost/Benefit Analysis window. The confusion matrix for the current value of the threshold is shown in the Confusion Matrix frame at the left bottom corner of the window.

Q8: what is the classification accuracy after you have minimized Cost/Benefits? What was it before you pressed that button? What is going on? In order to give an answer to this question and explain the corresponding phenomenon, let us take a look at the right side of the window. Its right bottom corner contains the ***Cost Matrix*** frame.

The left part of the frame contains the Cost matrix itself. Its four entries indicate the cost one should pay for decisions taken on the base of the classification model. The cost values are expressed in the table in abstract units, however in the case studies they can be considered in money scale, for example, in EUROs. The left bottom cell of the Cost matrix defines the cost of false positives. Its default value is 1 unit. In the case of virtual screening this corresponds to the mean price one should pay in order to synthesize (or purchase) and test a compound wrongly predicted by the model as "active". The right top cell of the Cost matrix defines the cost of false negatives. Its default value is 1 unit. In the case of virtual screening this corresponds to the mean price one should pay for

"throwing away" very useful compound and losing profit because of the wrong prediction taken by the classification model. It is also taken by default that one should not pay price for correct decisions.

It is clear that all these settings can be changed in order to match the real situation taking place in the process of drug design. The overall cost corresponding to the current value of the threshold is indicated at the right side of the frame. In order to find the threshold corresponding to the minimum cost, it is sufficient to press the button ***Minimize Cost/Benefit***. This explains the afore-mentioned difference in confusion matrices. The initial confusion matrix corresponds to the threshold 0.5, whereas the second confusion matrix results from the value of the threshold found by minimizing the cost function.
The current value of the cost is compared by the program with the cost of selecting the same number of instances at random. The difference between the values of the cost function between the random selection and the current value of the cost is called Gain. In the context of virtual screening, the Gain can be interpreted as the profit obtained by using the classification model instead of random selection of the same number of chemical compounds. Unfortunately, the current version of the Weka software does not provide the means of automatic maximization of the Gain function. However, this can easily be done interactively by moving the slider in the Threshold frame of the Cost/Benefit Analysis window.
The current model corresponds to the minimum value of the Cost function. Read the values for the current threshold from the right side of the ***Threshold*** frame.

Q9: "the current model (with the threshold obtained by minimizing the cost) specifies that it is optimal to select 3.9003 % of compounds in the course of virtual screening, and this ensures retrieving of 83.6735 % of active compounds. " Is this statement correct or incorrect?

Q10: What is the difference between cost sensitive and cost insensitive measures, in simple words?

Close the window with the Cost/Benefit Analysis

Task 6 – Statistical significance & paired t-test [max time: 20min]

Open the Experimenter interface.
The Experimenter has three panels: Setup, Run and Analyze.
To start a new experiment, click on New. Then on the line below, select the destination for your results – specify the following file name: "our_lab4" and choose CSV file. Underneath, select the junkbase dataset.
< http://stp.lingfil.uu.se/~santinim/ml/2016/Datasets/junk.arff >
To the right of the datasets select the algorithms to be tested. First, specify J48 with default parameters, then IBK with default parameters. The aim is to compare whether the numerical differences that we observe in the performance of the 2 models are statistically significant.

Now the experiment is ready for the first run.

If you wish, you can save your experimental settings by clicking the Save button at the top of panel (middle button).
To run the experiment, click the Run tab, which brings up a panel that contains a Start button: click it. A brief report is displayed when the operation is finished. The file "yourfilename.csv" contains the results. The CSV (Comma Separated Values[1]) can be read either in an ascii editor (the commas are column separators) or in a spreadsheet. Open the file in Excel. Each row represents one fold of the 10-fold crossvalidation (see the Fold column). You will see that each crossvalidation is run 10 times (the Run column) for each classifier, which makes 300 rows in all (plus the header row). Each row contains plenty of information. The spreadsheet is very informative, but the amount of information can be overwhelming.

For the time being we skip the detailed analysis of the dataset and we take a shortcut. A quick way to answer question "**How does J48 compare with the other classifiers**?" is to use the Analyse panel.
In the Analyze panel, click the Experiment button a the top right. Then click Perform test (near the bottom left). The result of a statistical significance test of the performance of the first learning scheme that we use as baseline (ie J48) versus IBk will be displayed in the large panel on the right.
We are comparing the percent correct statistic (default setting). The 2 methods are compared horizontally, as the heading of a little table. The numbers beside the scheme names identify which version of the scheme is being used (ignore them for the time being). The values in brackets at the beginning of the rows (ie. (100) is the number of experimental runs: 10 times tenfold crossvalidation.
The symbol placed beside a result indicates thea it is **statistically better (v)** or **worse (*)** than the baseline scheme (in this case J48 with default parameters) at the specified significance level (0.05 or 5%). The correct sampled t-test is used here (see Witten et al. (2011: 159).

Q11: Which is the worst method? How do you recognize from the output that it is worse than the other? Analyze and discuss the results and draw your conclusions on this experiment.

Additional Practice

Repeat Task 5 with the option "distanceWeighting" set to "Weight to 1-distance". What happens?
Read carefully Witten et al. (2011: Ch 5).
Read carefully Witten et al. (2011: Ch 13, The Experimenter).

[1] "A comma-separated values (CSV) file stores tabular data (numbers and text) in plain text. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format." Wikipedia.