Weka – Naive Bayes

Lab5 (in-class): 13:15-15:00 (CHOMSKY)

ACKNOWLEDGEMENTS: INFORMATION, EXAMPLES AND TASKS IN THIS LAB COME FROM SEVERAL WEB SOURCES.

Learning objectives

In this assignment you are going to:

- naïve bayes (simple, NB)
- multiple roc curves
- debugging

Task 1: Warming up –NaiveBayesSimple [maxtime: 10min]

Start Weka, launch the Explorer window and select the "Preprocess" tab.
Open the junk dataset.
< http://stp.lingfil.uu.se/~santinim/ml/2016/Datasets/junk.arff >

Q1: How many attributes?

Select the Classify tab to get into the Classification tab of Weka. Click on Choose→bayes and hover on to NaiveBayesSimple. Read carefully the content of the tooltip. Then run the classifier.

Q2: What happens? What would you suggest doing when you come across such a situation?

Task 2: Naïve Bayes [maxtime: 15min]

Reload the junk dataset with all attributes. Choose **bayes** and hover on **Naïve Bayes**. Read carefully the content of the tooltip. Run this classifier on the junk dataset.

Fill up the following table and read the "Warnings" below:

Classifier	**Acc**	**k-stat**	**Avg. P**	**Avg. R**	**Avg. F**	**Avg. AUC**
J48, default						
IBk, default						
NaïveBayes, default						

Warnings. Please note that all the evaluation metrics are somehow controversial. Your evaluation of the results must always be argued and supported by the knowledge you have of the data and by your domain expertise.

- Accuracy (in Weka : "Correctly Classified Instances") is the weakest metric because it is unstable since it is affected by class unbalance. Several tips and solutions are proposed by researchers and practitioners (eg. see: "*Evaluation of classification performance on small, imbalanced datasets* "[1], or "*8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset*"[2], "*What is the chance level accuracy in unbalanced classification problems?*"[3] and many many more).

- Cohen's kappa, although widely used for inter-rater agreement, is sometimes heavily criticized (eg. see here: Kappa Statistic is not Satisfactory for Assessing the Extent of Agreement Between Raters[4]).
- Accuracy, P/R, F, kappa tacitly assume equal error cost. This is not wise, since some errors are more costly than others! (see Section 5.7 in the weka book)
- ROC curves indicates the performance of a classifier without regard to class distribution or error cost. However ROC-curves and AUC metric are not appropriate for all problems (eg. see: "AUC: a misleading measure of the performance of predictive distribution models"[5].

- **ATT!** you cannot compare accuracy with AUC (ROC and AUC represent the average performance on all possible thresholds). These metrics measure different things! Please, read carefully how ROC curves are computed on pages 172-174 of the weka book, 3rd ed.
- well, use all evaluation metrics with a grain of salt...

Q3: which one is the most robust classifier, in your opinion? why? How do you explain the different results (Think of the underlying mathematical models and their inductive biases). Put forward your interpretations of the results.

Task 3: Statistical Significance [maxtime: 15min]

Quickly check whether the differences in performance are statistically significant.

Go to the experimenter interface, setup the experiment in the simple mode. Create an output cvs file. Choose the j48 as baseline, then ibk and naïve bayes. Run the classifiers on the junk dataset.

---

1 < https://people.inf.ethz.ch/bkay/talks/Brodersen_2010_06_21.pdf >

2 < http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/ >

3 < http://stats.stackexchange.com/questions/148149/what-is-the-chance-level-accuracy-in-unbalanced-classification-problems >

4 < http://www.agreestat.com/research_papers/kappa_statistic_is_not_satisfactory.pdf >

5 < ftp://gis.msl.mt.gov/Maxell/Models/Predictive_Modeling_for_DSS_Lincoln_NE_121510/Modeling_Literature/Loboetal2008_AUC.pdf >

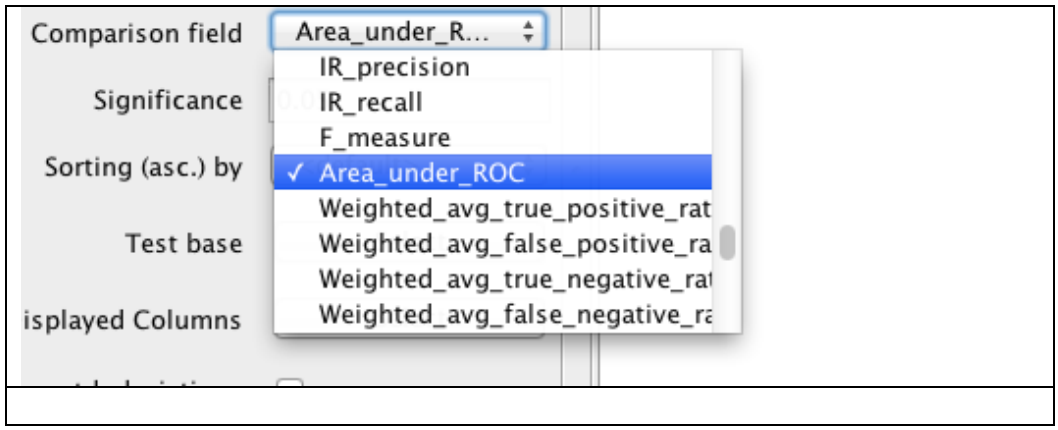
In Weka experimenter application, for the simple mode option:

In the section "iteration control", choose Data set first.

Att! What is the exact difference between “Data set first” and “Algorithm first”, where can we use either “Data set first” or “ Algorithm first”. This only makes a difference if you have multiple algorithms and datasets and if you store results in a database.

"Datasets first" will run the first algorithm in your list on all datasets before proceeding to the next algorithm, and so on. "Algorithms first" will run all algorithms in your list on the first dataset before proceeding to the next dataset, and so on. If you store results in a file then they will only become available when all runs are finished, so this won't make a difference. If you store results in a database, then you may be interested in seeing all results for the first dataset as quickly as possible, so you'd choose "Algorithms first".[6]

Choose the AUC as comparison in the Comparison field. See figure below.



Choose 0.05 significance level.

Run and Analyse the experiment

Q4: what happens?

Open the csv file, and try and fix the error.

Remember: 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%).

Q5: What does the snippet below mean?

```
------------------------
     (v/ /*) |  (0/0/1)
```

[6] <http://weka.8497.n7.nabble.com/iteration-control-td32663.html >

Task 4: multiple ROC Curves [maxtime: 30min]

***Preliminaries**: This task is based on the Weka Tutorial included in Lecture 5, namely "Weka Tutorial 30: Multiple ROC Curves (Model Evaluation)"*
< https://www.youtube.com/watch?v=rZHw3gGe7DA >.

The Knowledge Flow interface allows you to design configurations for streamed data processing[7]. The Knowledge Flow interface lets you drag boxes representing learning algorithms and data sources around the screen and join them together into the configuration you want. It enables you to specify a data stream by connecting components representing data sources, preprocessing tools, learning algorithms, evaluation methods, and visualization modules. If the filters and learning algorithms are capable of incremental learning, data will be loaded and processed incrementally. Most of the Knowledge Flow components will be familiar from the Explorer. For example, the Classifiers panel contains all of Weka's classifiers, the Filters panel contains the filters, and so on. The components for visualization and evaluation have not yet been encountered. Read their their characteristics and descriptions in the related chapter. You establish the knowledge flow by configuring the individual components and connecting them up by right-clicking the various component types. *This interface is described in Ch 12 of the weka book 3rd ed. and Ch 11 of the weka book 2nd ed.*

Hands-on

Open the Knowledge Flow interface.

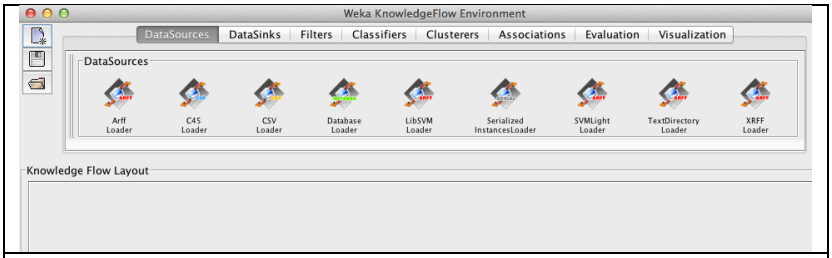


Fig. 1: Knowledge Flow Interface

Pick up the components.

- Click on ArffLoader (DataSource tab). The mouse cursor changes to crosshairs to signal that are ready to place the component. Do this by clicking anywhere in the lower part of the screen, called *canvas* or *Knowledge Flow Layout.*
- Select ClassAssigner (Evaluation tab) and place it in the canvas.
- Select ClassValuePicker (Evaluation tab) and place it in the canvas.
- Select CrossValidationFoldMaker (Evaluation tab) and place it in the canvas.
- Select NaiveBayes and J48 (Classifiers tab) and place it in the canvas.
- Select ClassifierPerformanceEvaluator (Evaluation tab) and place it in the canvas. We need two instances of the ClassifierPerformanceEvaluator because we are going to use two classifiers in this task.
- Select ModelPerformanceChart (Visualization tab) and place it in the canvas.

Create the connections.

- Right-click on the ArffLoader (canvas) and choose "dataset". Draw a connection (ie a line) to the ClassAssigner. Then create a connection between ClassAssigner and

[7] Stream processing is a computer programming paradigm that allows some applications to more easily exploit a limited form of parallel processing. SP treats data not as static tables or files, but as a continuous infinite stream of data.

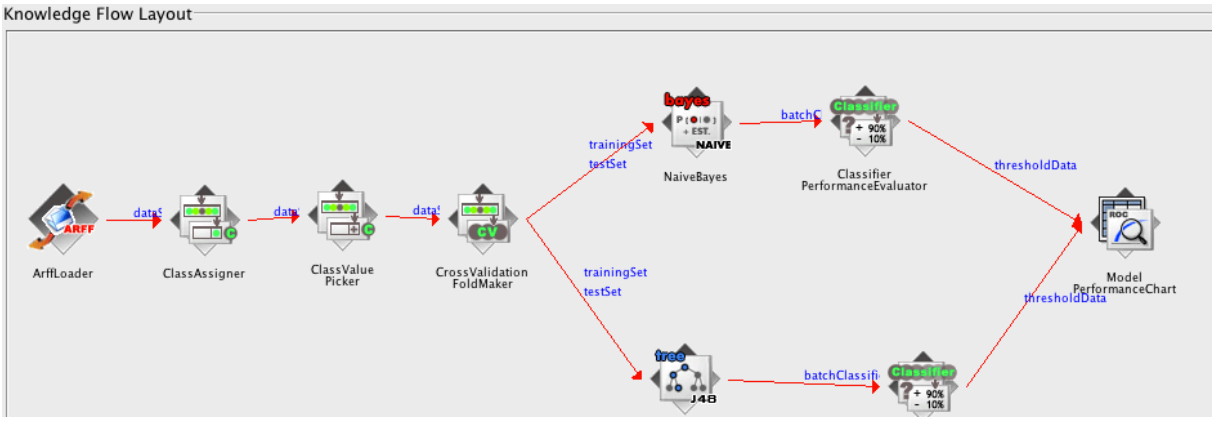
ClassValuePicker, and between ClassValuePicker and CrossValidationFoldMaker.

- Right-click on CrossValidationFoldMaker, choose Training set and create a connection with the NB classifier. Right-click again and choose Test set and create a connection with the NB classifier.
- Repeat these steps to create connections between the CrossValidationFoldMaker and J48.
- Right-click on the NB classifier, select BatchClassifier and create a connection with the ClassifierPerformanceEvaluator.
- Repeat these steps to create a connection between J48 and ClassifierPerformanceEvaluator.
- RightClick on the ClassifierPerformanceEvaluator (conneted to NB), select ThresholdData and create a connection between ClassifierPerformanceEvaluator and ModelPerformanceChart.
- Repeat these steps for ClassifierPerformanceEvaluator connected to J48.

Load the data.

- Right-click on the ArffLoader, choose Configure and choose the junk dataset.
- Right-click on ClassAssigner, choose Configure and choose the class attribute (last attribute).
- Right-click on ClassValuePicker, choose Configure and select the "junk" class.
- Right-click on CrossValidationFoldMaker, choose Configure. Confirm that you are making a 10-fold-crossvalidation with the default random seed value 1.

The stream that you have just created should look somewhat similar to the stream shown in the picture below.



Start the processing.

- Right-click on ArffLoader and choose StartLoading. The results are displayed in ModelPerformanceChart.
- Right-click on ModelPerformanceChart and choose "Show chart".

Q6: Which curve is better? why? motivate your answer.

Task 5: Resuming Cost/Gain analysis

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

Reload the ace dataset into the Explorer interface.

The dataset contains 1846 instances, 1025 attributes including the class activity that has 2 values: nonactive and active.

The ace dataset is highly unbalanced, since the number of non-active compounds is much larger the number of active compounds.
Go throw the following tutorial that inspired Task 6 in Lab4 < http://masterchemoinfo.u-strasbg.fr/Documents/TutoChemo/classification.pdf > (random forests excluded).

Additional practice

Weka knowledge flow tutorial
< http://software.ucv.ro/~eganea/AIR/KnowledgeFlowTutorial-3-5-8.pdf >

Additional reading: About Normal Distribution and Standard deviation equal to 0:
< http://statistics.about.com/od/Mathstat/fl/When-Is-the-Standard-Deviation-Equal-to-Zero.htm >
< http://www.mathplanet.com/education/algebra-2/quadratic-functions-and-inequalities/standard-deviation-and-normal-distribution >

Solution Task 3