

CS 374 Assignment #8 Evaluation of Learning Algorithms

Due the week of October 31, 2005

This week you will step back from your study of individual classifier-learning algorithms in order to evaluate them both empirically and analytically. The readings and exercises will focus on evaluation methodology, but the broader goal of this assignment is to give you the opportunity to reflect on all that you've already learned this semester.

1 Assessing and Comparing Classifier-Learning Algorithms

1.1 Reading

Please read the following:

- Alpaydin, Chapter 14 (Assessing and Comparing Classification Algorithms) and
- Mitchell, Chapter 5 (Evaluating Hypotheses).

The order in which you read these is not important. The Alpaydin reading covers more than the chapter in Mitchell, but the explanations in Mitchell are more clear.

1.2 Exercises

To check your understanding of the readings, please do the following exercises:

- 2 and 3 on page 349 of Alpaydin;
- 5.1-5.4 on page 152 of Mitchell.

1.3 Report on the Evaluation of Learning Algorithms

Your central task this week is to write a report that discusses and compares the classifier-learning algorithms you have studied this semester, including:

- Naive Bayes
- C4.5
- k Nearest Neighbor
- Backpropagation for multilayer perceptrons
- SMO

Your report should include an empirical evaluation as outlined below. It should also include a general analysis and discussion of the pros and cons of each of the methods evaluated. You might even find it useful to consider some of the major theorems that relate to each of the algorithms.

1.3.1 Empirical Evaluation: Learning Curves

When assessing a classifier-learning algorithm, it is useful to know not only its best performance for a given application domain, but also how quickly it approaches that best performance. Your evaluation should begin by applying each of the algorithms above to three different data sets. Specifically, you should use Weka to evaluate: J48, 1-NN, 3-NN, NaiveBayes, SMO, and MultilayerPerceptron. For each of these algorithms you should use default settings.

The three data sets to be evaluated are `ionosphere`, `vehicle`, and `sick`. You can find them in

```
~andrea/shared/cs374/HW8
```

Each data set has several training files and one or more test files:

`ionosphere`:

```
ionosphere-50.arff      50 training examples
ionosphere-100.arff    100 training examples
ionosphere-150.arff    150 training examples
ionosphere-200.arff    200 training examples
ionosphere-250.arff    250 training examples
ionosphere-test.arff   test data
```

`vehicle`:

```
vehicle-50.arff        50 training examples
vehicle-100.arff       100 training examples
vehicle-150.arff       150 training examples
vehicle-200.arff       200 training examples
vehicle-250.arff       250 training examples
vehicle-test.arff      test data
```

`sick`:

```
sick-50.arff           50 training examples
sick-100.arff          100 training examples
sick-150.arff          150 training examples
sick-200.arff          200 training examples
sick-250.arff          250 training examples
sick-test.arff         test data
```

I generated each of these data sets from larger data files that are available in the same directory. Each of the data files includes a description of the data and, more generally, the application domain.

You should run each of the learning algorithms on each training data file, evaluating the results on the corresponding test data file. Your report should include tables summarizing your results for each data set as follows:

```
ionosphere:
N          J48  1-NN  3-NN  NB  SOM  MLP
50         aaa  bbb  ccc  ddd  eee  fff
100        aaa  bbb  ccc  ddd  eee  fff
150        aaa  bbb  ccc  ddd  eee  fff
200        aaa  bbb  ccc  ddd  eee  fff
250        aaa  bbb  ccc  ddd  eee  fff
```

where `aaa` gives the error rate of J48, `bbb` gives the error rate of 1-Nearest Neighbor, etc. The sizes of the test files given above are the same. You can find additional test data in the same location, if you're interested in seeing more.

When running the algorithms on the data sets above, you might also find it useful to note the time required to learn each model. It is also interesting to consider the confusion matrix given for each evaluation run on a test set.

If you prefer, you can include graphs of the data above, rather than tables. I recommend using gnuplot. Version 3.7 is installed on the machines in the lab. You can find documentation on gnuplot at

<http://www.duke.edu/~7Ehpgavin/gnuplot.html>

1.3.2 Empirical Evaluation: Comparison of Average Error Rates

Your report should also include comparisons of the average error rates of each of the algorithms on several data sets. The specific data sets to be considered are:

- `ionosphere`
- `vehicle`
- `auto`
- `vote`

These can all be found in

`~andrea/shared/cs374/UCI`

This directory includes many data sets from the UC Irvine Machine Learning Repository, which have been converted into arff format for Weka. Your evaluation should certainly include the specific data sets named here, but you should also feel free to extend your evaluation to other data sets. I selected the four data sets above for several reasons:

- They include both large and small numbers of examples.
- They include both continuous and nominal attributes.
- Some have missing attribute values.
- They include both two-class and multi-class problems.
- All of the algorithms will run to completion on them in a reasonable amount of time.

Use the Weka Experimenter (recall the exercise you did in order to find the optimal value for `k` for the `k`-Nearest Neighbor assignment). You should perform 10-fold cross validation for each algorithm for each data set. You do not need to repeat this more than one time. (So be sure to adjust the parameter in the right half of the window that asks for the number of times to repeat. The default is 10 – change it to 1.) Include the results of your experiments in the report you write. How do the algorithms compare to each other? Which differences in error rates are statistically significant?

1.3.3 General Analysis of Learning Algorithms

In addition to performing the tests specified above, you should take this assignment as an opportunity to reflect on the general advantages and disadvantages of each of the algorithms you have studied this semester. Be sure to include these in your report. As noted above, you might even want to refer to some of the major theorems you have proved.

1.3.4 Optimizing the Algorithms

As an extra challenge (this is not a requirement for this assignment), try out various parameter settings for each of the learning algorithms to find those that work best on the data sets above.

2 Challenges to Evaluation Methodology

Many researchers in the field of machine learning are concerned with making evaluation methodology better. Sometimes work toward that goal considers ways in which some of the statistics we report might be misleading. Others propose new frameworks for evaluating our algorithms.

2.1 Reading

Please read “The Case Against Accuracy Estimation for Comparing Induction Algorithms” by Provost, Fawcett, and Kohavi. This paper appeared in ICML-98.

2.2 Reaction

Write a brief one-page summary and reaction to the paper. Are you convinced by the authors’ arguments?