Assignment 7

Machine Learning, Summer term 2013, Ulrike von Luxburg

Solutions due by June 3

Goal of this assignment is to think about how to compare classifiers. Some questions may not have a concrete answer. Think about those problems and try to use your intuition about algorithms. The aim is to get a feeling for how to play with data and different algorithms. You are encouraged to run algorithms several times with different parameters and settings to learn different aspects of an algorithm. You are also welcome to use available resources in the Internet.

Exercise 1 (Discussion, 3 points) Assume you are going to buy a classifier. How would you compare different classifiers? What are the relevant criteria? What is a good setup of for experiments, and how to evaluate the results?

Exercise 2 (Compare in practice, 15 points) In this exercise you should compare different classifiers in practice. This contains several steps:

1. Selecting candidate classifiers: Consider at least three different classifiers
   - kNN
   - Linear SVM and a kernel SVM
   - One more algorithm of your choice.
2. Choosing datasets: You should use at least three different datasets. The size of the training and the test datasets are also a design choice. "Consistent" classifiers will all perform very good when we have "enough" training points.
   - USPS dataset
   - Breast cancer dataset
   - One or two **new** data sets. You can bring your own data set or use UCI repository. `http://archive.ics.uci.edu/ml/`
     Note that these datasets would need preprocessing (If all attributes are real, you can easily import it into matlab).
3. Run the algorithms on the data sets and compare the results. Use cross validation to select parameters. You can use any available implementation of your classifiers (Hint: First try with a reduced number of points and/or reduced number of features).
4. Try to "understand" the results and the data. Some questions that you should ask:
   - Which of the multiclass approaches works better?
   - Are there some easy and some difficult classes?
   - How variable are the test errors?
   - Is there one algorithm that is clearly "better" on these data sets?
   - How dependent are the results of the parameters?
5. Visualization: You should come up with ways to visualize your results (e.g. plot all kinds of statistics).

Web page: `http://www.informatik.uni-hamburg.de/ML/contents/people/luxburg/teaching/2013-ss-vorlesung-ml/`
Login: "machine", Password: "learning"