IN4320 Machine Learning Exercise

Version 1.0.0

February 13, 2019

This exercise is an individual exercise, that means that you have to give your own answer, and you should not copy it from somebody else. You may discuss the problem with others, but you have to formulate your own anser.

You can use any programming languange that you prefer, but Matlab or Python are recommended. The responsible teacher can only help with Matlab.

Computational Learning Theory: boosting

Before you answer the following questions, you need to read the paper 'A decision-theoretic generalization of on-line learning and an application to boosting' by Y. Freund and R.E. Schapire, 1995 (also available from Brightspace, under Course Documents, Reading Material, Computation Learning Theory).

a. Show the equivalence between the formulation of Adaboost as it is given in the slides of the course, and the formulation as it is given in the paper.

b. Implement a weak learner **WeakLearn**: the decision stump, that is minimising the apparent error.

To find the optimal feature f and threshold θ, you have to do an *exhaustive search* for a training set. Convince yourself that it indeed optimizes what it should optimize.

Extend the implementation of the weak learner such that it accepts a weight per object. Convince yourself that it indeed optimizes what it should optimize.

Show your code.

c. Test the implementation of the decision stump on a simple dataset: generate two classes from two Gaussian distributions, where the means are $\mu_1 = [0, 0]^T$ and $\mu_2 = [2, 0]^T$, and the covariance matrices are identity matrices.) Make a scatterplot of the data, and give the optimal parameters obtained by your decision stump. Do the parameter values make sense?

Does the decision stump change if you rescale one of the features (for instance, when you multiply feature 2 by a factor of 10)?

d. Test the implementation on (an adapted version of) the Fashion NIST dataset. On Brightspace you will find a training and a test set, called `fashion57_train.txt` and `fashion57_test.txt`. They consist of the pixel values of small 28×28 images. In the training set, the first 32 rows are class 1, and the next 28 are class 2. The test set is larger: the first 195 rows contain samples from class 1, and the next 205 class 2. In Matlab you can read in the data using:

```
>> a = dlmread('fashion57_train.txt');
```

Train the decision stump on the training set and evaluate on the test set. What are the apparent error and the test error?

e. Implement the AdaBoost algorithm that is described in the paper of Freund and Shapire (use the notation and conventions from the paper, not the slides!), and implement the code to classify new and unseen data.

Show the code.

Convince yourself that it indeed optimizes what it should optimize.

f. Test your implementation on the simple dataset that you created in (c.). Find out which objects obtain a large weight w_i^t.[1] Keep the number of iterations T fixed to, say, $T = 100$.

g. Test the implementation on the Fashion dataset. What is the classification error on the test objects? How does this error depend on the number

[1] And no, it is *not* sufficient to just say 'object 13 and 7 have high weight'. Show me; do the more simple or more difficult objects get high weight?

of iterations T? If you take the classifier with the optimal T, which training objects get a high weight? Show them!

h. Show the learning curve for `n=[2,4,6,10,15,20]` training objects per class.