Machine Learning Jun.-Prof. Matthias Hein

Exercise Sheet 10 - 17.01.2012 due: 24.01.2012

Exercise 1 - GentleBoost

In this exercise you are implementing GentleBoost using a real-valued decision stump as the weak base classifier.

a. **(3 Points)** In the base step of the boosting procedure one has to fit the weak classifier to the weighted training data. As the weak classifier f we use a decision stump:

$$f(x) = a\mathbb{1}_{\langle w,x\rangle+b>0} + c,$$

with $a, b, c \in \mathbb{R}$ and $w \in \mathbb{R}^d$. Assume that w is fixed and the parameters a, b, c are computed as minimizers of the weighted least squares loss $L(f)$,

$$L(f) = \sum_{i=1}^{n} \gamma_i \big(Y_i - f(X_i)\big)^2,$$

given some weights $\gamma \in \mathbb{R}^n$. Derive the optimality condition for a and c. How do you then compute the optimal a, b and c efficiently?

b. **(3 Points)** Write a Matlab-function

```
[a,b,c,minLoss]=FitStump(X,Y,w,gamma)
```

which given the fixed vector $w \in \mathbb{R}^d$ and the weights $\gamma \in \mathbb{R}^n$ (n is the number of training points) returns the optimal decision stump, i.e. the optimal parameters a, b, c, of

$$f(x) = a\mathbb{1}_{\langle w,x\rangle+b>0} + c.$$

as well as the corresponding loss $L(f)$ (as usual $X \in \mathbb{R}^{n\times d}$ and $Y \in \mathbb{R}^n$).

c. **(2 Points)** Write a Matlab-function

```
[W,aparam,bparam,cparam]=GentleBoostTrain(X,Y,k)
```

which given the training data X, Y and the number of maximal iterations `k` returns

$$\texttt{W} \in \mathbb{R}^{d\times k}, \texttt{aparam} \in \mathbb{R}^k, \texttt{bparam} \in \mathbb{R}^k, \texttt{cparam} \in \mathbb{R}^k,$$

where k is the number of used weak classifiers.

As the weak learner use the decision stump of b), where the weight vector w is drawn uniformly from the unit sphere (`w=randn(d,1); w=w/norm(w);`)

d. **(1 Point)** Write a Matlab-function

```
f=GentleBoostClassify(X,W,aparam,bparam,cparam)
```

which given the parameters $\texttt{W} \in \mathbb{R}^{d\times k}, \texttt{aparam} \in \mathbb{R}^k, \texttt{bparam} \in \mathbb{R}^k, \texttt{cparam} \in \mathbb{R}^k$ returns the prediction $f \in \{-1, 1\}^n$ on the testing data X.

e. **(2 Points)** Apply your method to the `diabetes.mat` dataset from the last exercise sheet. Use 5-fold cross validation to determine the best value of number of weak classifiers k from the set $\{10, 20, \ldots, 1000\}$. Report k and the corresponding cross validation error, training error (retrained on the full training set) and test error. Compare the results of the classification to the result of kernel ridge regression from the last exercise sheet. Submit the code you used to generate the results.

Hints: For the implementation of the function `FitStump`

- The function `cumsum` which computes the cumulative sum of a vector might be useful.
- One can very efficiently implement this function using the power of vectorization. There is no need for any for loop !

Exercise 2 - Classification of handwritten digits

(3 Points) The problem deals with the classification of handwritten digits (10 classes). The file `USPS.mat` contains training and test data `Xtrain` resp. `Xtest`, where each row corresponds to an image of size 16 × 16, as well as the corresponding labels `Ytrain` resp. `Ytest`.

Write a matlab script `GentleBoostOneVersusAll.m` which solves the multi-class problem using GentleBoost in a **one-versus-all** scheme. Apply the GentleBoost classifier with $k = 1000$ to the USPS data. Save your prediction `Pred` and the corresponding test error `TestError` on the test set in a file `USPSResults.mat`.

Visually inspect the digits which have been misclassified (use `VecToImage(X,16,16,0,1,1)`). How do you judge the result?

Exercise 3 - T-test

(2 Points) The file `Exercise3.mat` contains the predictions `Yboost` using gentle boost as well as `Ysvm` using the support vector machine with Gaussian kernel, as well as the true labels `Ytest` on a subset of the USPS dataset from exercise 2. One observes an error of 3.7% of boosting and 2.1% of the SVM.

Perform a T-test as covered in the lecture using significance level $\alpha = 0.05$ to compare the performance (in terms of true error) of these two classifiers. State the null hypothesis. Report the value of the test statistic. Do you reject the null hypothesis? (Use the t-distribution table here: `http://www.blogforgood.net/wp-content/uploads/2011/09/T-Table.jpg`).

Submit the code used to perform the T-test.

Submission: Zip the files and send them to `tb@cs.uni-saarland.de`, using the naming convention introduced in previous exercise sheets, and using as subject line `ML Sheet 10`.