Machine Learning	Prof. Dr. Bernt Schiele
	Dr. Paul Swoboda & Dr. Gerard Pons-Moll
Exercise Sheet 9 - 17.12.2018	due: 07.01.2019, 14:15

Exercise 17 - GentleBoost

You are supposed to implement GentleBoost as discussed in the lecture 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. In GentleBoost this is done via weighted least squares. 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. Derive the optimality condition for a and c for minimizing the weighted least squares loss $L(f)$,

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

where $\gamma \in \mathbb{R}^n$ are the weights.

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

```
[a,b,c,minError]=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) derives the optimal decision stump, that is derive the optimal parameters a, b, c, of

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

(as usual $X \in \mathbb{R}^{n\times d}$ and $Y \in \mathbb{R}^n$).

c. **(4 Points)** Write a Matlab-function

```
[W,aparam,bparam,cparam]=GentleBoost(X,Y,MaxIter)
```

which given the training data X, Y and the number of maximal iterations `MaxIter` 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 ($k \leq$ `MaxIter`).

As the weak learner use the decision stump of b).

d. **(2 Points)** Apply the GentleBoost classifier with `MaxIter` = 100 to the USPS data using one-versus-all classification. Save your predictions `Pred` and the corresponding test error `TestError` in the file `USPSResults`. Plot the training and test error as a function of the number of iterations done. Save the plot as `PlotTrainTestError`. How does the test error compare to the one of the support vector machine ?

Hints:

a. For the implementation of the function `FitStump`

   - You do not need to check all possible thresholds b - think about how many different possible thresholds b exist which yield different results for the weighted least squares error ? Compute for each possible threshold first the optimal parameters a and b and then the corresponding weighted least squares error. Take the threshold which yields the smallest error.
   - The function `cumsum` which computes the cumulative sum of a vector might be useful.
   - One can very efficiently implement this function using vectorization. There is no need for any for loop !

b. For the implementation of the function `GentleBoost` draw the weight vector w uniformly from the unit sphere (`w=randn(dim,1); w=w/norm(w);`)

Submission instructions

- We accept both handwritten and electronic submissions. So you can choose what is more convenient for you. In any case, you should specify full names and immatriculation IDs of all team members. Obviously, programming tasks you can submit only electronically.
- Handwritten submissions should be submitted in the lecture hall of Monday's lecture (before the lecture starts).
- Electronic submissions should be zipped, containing the m-files (`Basis` etc.), your plots (png files) and the matlab data files (.mat) and emailed to the corresponding tutor:

  a. Apratim Bhattacharyya (Wednesday 8-10): abhattac@mpi-inf.mpg.de
  b. Maksym Andriushchenko (Thursday 8-10): s8mmandr@stud.uni-saarland.de
  c. Max Losch (Friday 16-18): mlosch@mpi-inf.mpg.de

  If not all 3 students belong to the same tutorial group, then you should email your submission to **only** one tutor (e.g. to the tutor of the first author of your homework), so please do not put other tutors in copy of the email.

  The email subject must have the following form: "[ML18/19 Exercise] Sheet X", where X is the number of the current exercise sheet. Then please specify in the email full names and immatriculation IDs of all team members. Then please attach all your files as a single zip archive, which consists of your immatriculation IDs, e.g. "2561234_2561235_2561236.zip".

- Reminder: you should submit in groups of 3. Otherwise, we will later on merge the groups smaller than 3 students.