Machine Learning	Dr. Mario Fritz and Dr. Björn Andres
Exercise Sheet 11 - 26.01.2015	due: 02.02.2015

Exercise 25 - 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. **(3 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

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

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);`)

As usual, please zip all files into one and name it using the filename convention and then send it to your respective tutor.

Exercise 26 - Properties of entropy and mutual information

- **(5 Points)** Prove Lemma 7 on page 121 of the lecture notes, which says

Lemma 1

- $H(Y|X) = H(Y,X) - H(X)$,
- $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$,
- $I(X;Y) = I(Y;X)$ *and* $I(X;Y) \geq 0$,
- $I(X;Y) = 0$ *if and only if* X *is independent of* Y,
- $I(X;Y\,|\,Z) = 0$ *if and only if* X *and* Y *are conditionally independent given* Z.

Hints:

a. It is helpful for the second part of this exercise to use Jensen's inequality. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function and $g : \mathbb{R}^m \to \mathbb{R}^d$ be any real-valued measurable function, then

$$f(\mathbb{E}[g(X)]) \leq \mathbb{E}[f(g(X))].$$

for any distribution of X with values in $\mathbb{R}^m$. In integral form this inequality reads,

$$f\left(\int_{\mathbb{R}^m} g(x)\,p(x)\,dx\right) \leq \int_{\mathbb{R}^m} f(g(x))\,p(x),dx.$$

For a strictly convex function, we have for any distribution of X equality,

$$f\big(\mathbb{E}[g(X)]\big) = \mathbb{E}\big[f\big(g(X)\big)\big] \quad \text{if and only if} \quad g(x) = const., \text{ almost everywhere..}$$