Statistical Machine Learning
Assignment 3

Please send me

- the **original script** detailing your computations.
  - The script must be **documented**, i.e. the code corresponding to each answer must be delimited and your loops/variables briefly explained.
  - The script must be **executable**: by just running your script, all results should appear **automatically**.
  - Do not use external functions, everything must be coded **by yourself** using elementary linear algebra functions and standard libraries.
- A document (.doc, .pdf) which will contain your answer and your analysis. Do not put your source code in that document. Illustrations, graphs, *etc.* are welcome.

This homework is due **Jan 9th (Tue.) noon**

Send your homework to marcocuturicameto+report@gmail.com. Please put the word `report` in the title of your email.

Exercise 1: Classification - Hoeffding's and V.C Bounds

- Choose two gaussian densities[1] p_{-1}, p_{+1} on $\mathbb{R}$ with unit variance and mean in $[-1, 1]$. We consider a pair of random variables (X, Y) where the density of (X, Y) is defined by the following: $p(Y = 1) = 0.65$ and the density of $p(X|Y = 1)$ is equal to p_{+1} while $p(X|Y = -1)$ is equal to p_{-1}.
- Consider $N = 20$ different linear classifiers on $\mathbb{R}$, that is step functions defined by a threshold τ and a sign $t \in \{-1, 1\}$ as

$$f_{t,\tau}(x) = \begin{cases} t \text{ if } x > \tau \\ -t \text{ if } x \leq \tau \end{cases} .$$

  Choose $t \in \{-1, 1\}$ and $\tau \in [-2, 2]$ randomly and uniformly.
- Give a detailed illustration of Hoeffding's bound for the supremum of the difference of the empirical risk and the true risk for the set of N functions considered above, by sampling 200 sets of $n = 20, 50, 100$ independent observations of (X, Y). In order to do so, you will need to compute the true risk of each of the Heaviside functions (the Error function[2] might be useful) and sample randomly from the densities p_{-1} and p_{+1}. Try to split these steps using short subroutines to improve overall readibility of your code.

[1] http://en.wikipedia.org/wiki/Normal_distribution

[2] http://en.wikipedia.org/wiki/Error_function

- We have studied Vapnik Chervonenkis bounds for infinite families of functions. Give an expression for this bound when considering all possible translations and multiplications by $\{-1, 1\}$ of the Heaviside-functions. Your bound should only depend on the threshold ε and sample size n. Find a condition on N for which the VC bound is tighter (that is, provides a lower bound) than Hoeffding's bound.