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

• Choose two gaussian densities $p_{-1}, p_{+1}$ 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 $\tau$ 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 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 readability of your code.

\[\text{http://en.wikipedia.org/wiki/Normal_distribution}\]
\[\text{http://en.wikipedia.org/wiki/Error_function}\]
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 \(\varepsilon\) 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.