MIT 9.520: Statistical Learning Theory **Fall 2014**

Problem Set 2

Student: Brando Miranda

Problem 1 Please write your analysis on Problem 1 here

Problem 2 Please write your analysis on Problem 2 here

Problem 3 Please write your analysis on Problem 3 here

Problem 4 Please write your analysis on Problem 4 here a)

Lemma 1: Let $X_i = I_S[f_i]$, then $\mathbb{E}[X_i] = I[f]$ and by chebyshev's bound therefore:

$$Pr[|I_s[f_i] - I[f_i]| \geq \epsilon|] \leq \frac{Var[X_i]}{\epsilon^2}$$

Proof: Let $X_i = I_S[f_i] = \frac{1}{n}\sum_{j=1}^{n} V(f_i, z_j)$, where z_j is the random samples/data and f_i is a non random (and fixed) function in $\mathcal{H}$. Then, if we take the expectation of X_i wrt to the distribution of z we get:

$$\mathbb{E}[X_i] = \mathbb{E}[\frac{1}{n}\sum_{j=1}^{n} V(f_i, z_j)] = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[V(f_i, z_j)] = \mathbb{E}[V(f_i, z_j)] = I[f_i]$$

Thus, we can use chebyshev's and thus the following statement is true (concluding proof of lemma):

$$Pr[|I_s[f_i] - I[f_i]| \geq \epsilon|] \leq \frac{Var[X_i]}{\epsilon^2}$$

□

Theorem: Given the conditions in the question, the the upper bound we are looking for is as follows:

$$Pr[\sup_{f \in \mathcal{H}} |I_s[f] - I[f]| \geq \epsilon|] \leq \frac{N(c^2 - 2cM + M^2)^2}{n\epsilon^2}$$

Proof: If the largest difference between the empirical risk and the expected risk is larger than ϵ, then that means that the defect (i.e. difference of empirical risk and generalization error) of one of the functions in this finite set is larger than ϵ. i.e. At least one of the defects is larger than ϵ. In equations it reads as follows:

$$Pr[\sup_{f \in \mathcal{H}} |I_s[f] - I[f]| \geq \epsilon|] = Pr[\cup_{i=1}^{N} |I_s[f_i] - I[f_i]| \geq \epsilon|]$$

by the union bound:

$$Pr[\sup_{f\in\mathcal{H}} |I_s[f] - I[f]| \geq \epsilon|] \leq \sum_{i=1}^{N} Pr[|I_s[f_i] - I[f_i]| \geq \epsilon|]$$

By Lemma 1 we know:

$$Pr[|I_s[f_i] - I[f_i]| \geq \epsilon|] \leq \frac{Var[X_i]}{\epsilon^2}$$

Therefore using this upper bound we can further upper bound our equation above by:

$$\sum_{i=1}^{N} Pr[|I_s[f_i] - I[f_i]| \geq \epsilon|] \leq \sum_{i=1}^{N} \frac{Var[I_s[f_i]]}{\epsilon^2}$$

Now if we can upper bound the variance, we are done.

$$Var[I_s[f_i]] = Var[\frac{1}{n}\sum_{j}^{n} V(f_i, z_j)]$$

Since the function f_i is fixed (i.e. it was NOT chosen based on the training data) and the only randomness involved is with selecting samples/data from z, then z is the only random variable. However, since its part of the training data and it was sampled in an iid way, then z's are independent. Which makes the cost functions $V(f_i, z_j)$ independent. Therefore, we can use the "linearity" of variance when the random variables are independent to yield the following statement:

$$Var[\frac{1}{n}\sum_{j}^{n} V(f_i, z_j)] = \frac{1}{n^2}\sum_{j}^{n} Var[V(f_i, z_j)] = \frac{1}{n} Var[V(f_i, z_j)]$$

Now lets use the definition variance to upper bound the above:

$$Var[V(f_i, z_j)] = \mathbb{E}[V(f_i, z_j)^2] - \mathbb{E}[V(f_i, z_j)]^2$$

To upper bound the above we need to upper bound $\mathbb{E}[V(f_i, z_j)^2]$ and lower bound $\mathbb{E}[V(f_i, z_j)]^2$.

It is easy to lower bound $\mathbb{E}[V(f_i, z_j)]^2$ because according to the question, it is the squared loss function, which can never be less than zero. Therefore, the lower bound is:

$$\mathbb{E}[V(f_i, z_j)]^2 \geq 0$$

To upper bound $\mathbb{E}[V(f_i, z_j)]^2$ we need to substitute in the definition of the squared loss function:

$$\mathbb{E}[V(f_i, z_j)^2] = \mathbb{E}[f_i(x)^2 - 2f_i(x)y + y^2]$$

Since $sup_x \in X|f(x)| \leq C$ and the max value of any y is M, then we have:

$$\mathbb{E}[V(f_i, z_j)^2] \leq \mathbb{E}[(C^2 - 2CM + M^2)^2] = (C^2 - 2CM + M^2)^2$$

Setting the terms we desired to be their maximum value and minimum value respectively, we get:

$$Var[V(f_i, z_j)] = \mathbb{E}[V(f_i, z_j)^2] - \mathbb{E}[V(f_i, z_j)]^2 \leq (C^2 - 2CM + M^2)^2 - 0 = (C^2 - 2CM + M^2)^2$$

$$Var[V(f_i, z_j)] \leq (C^2 - 2CM + M^2)^2$$

Yielding:

$$Var[I_s[f_i]] \leq \frac{1}{n}(C^2 - 2CM + M^2)^2$$

However, our goal is to bound the following:

$$Pr[\sup_{f \in \mathcal{H}} |I_s[f] - I[f]| \geq \epsilon|] \leq \sum_{i=1}^{N} \frac{Var[I_s[f_i]]}{\epsilon^2}$$

Now that we have an upper bound on the variance of the empirical risk, lets upper bound the desired probability:

$$Pr[\sup_{f \in \mathcal{H}} |I_s[f] - I[f]| \geq \epsilon|] \leq \sum_{i=1}^{N} \frac{1}{n} \frac{(C^2 - 2CM + M^2)^2}{\epsilon^2} = \frac{N}{n} \frac{(C^2 - 2CM + M^2)^2}{\epsilon^2}$$

Yielding the desired upper bound. □

Problem 5 (MATLAB) Please write your analysis on Problem 5 here