Homework 2

1 The Compression Bound in the Non-Realizable Case

Assuming that the loss is bounded in $[0, 1]$. Prove that with probability at least $1 - \delta$, we have that if I is a compression set for T, then:

$$|L(\mathcal{A}(T)) - \hat{L}_{-I}(\mathcal{A}(T))| \leq \sqrt{\frac{(l+1)\log m + \log \frac{2}{\delta}}{2(m-l)}}$$

where l is the size of the compression set and the probability is with respect to a random draw of T.

2 Boosting and the Exponential Loss

Assume that we have a set of hypothesis

$$\mathcal{H} = \{h_1(\cdot), h_2(\cdot), \ldots h_k(\cdot)\}$$

where each h is a mapping from $\mathcal{X}$ to $\{-1, 1\}$ (in class, we considered slightly more general hypothesis class which mapped to $[-1, 1]$). Our weak learner learner will only use hypothesis from $\mathcal{H}$.

The AdaBoost algorithm we presented in class is equivalent to the following algorithm:

Algorithm 1 AdaBoost

Input parameters: T, $\mathcal{H}$

Initialize $w_1 \leftarrow \frac{1}{m}\mathbf{1}$
Initialize the classifier $f_0(x) = 0$
for $t = 1$ to T **do**

1. Choose the hypothesis h_t as follows:

$$h_t = \text{argmin}_{h \in \mathcal{H}} \sum_{i=1}^{m} w_{t,i} \mathbf{1}\left[h(x_i) \neq y_i\right]$$

(this step can be considered the implementation of the weak learner).

2. Calculate the error

$$\gamma_t = \sum_{i=1}^{m} w_{t,i} \mathbf{1}\left[h_t(x_i) \neq y_i\right]$$

3. Set

$$\beta_t = \log \frac{1-\gamma_t}{\gamma_t}$$

and update the weights

$$w_{t+1,i} = \frac{w_{t,i} e^{\beta_t \mathbf{1}[h_t(x_i) \neq y_i]}}{Z_t}, \quad Z_t = \sum_{i=1}^{m} w_{t,i} e^{\beta_t \mathbf{1}[h_t(x_i) \neq y_i]}$$

4. Update the classifier:

$$f_t(x) = f_{t-1}(x) + \beta_t h_t(x)$$

end for
OUTPUT the classifier $\text{sgn}\left(f_T(x)\right)$

Let us now provide a different interpretation of Boosting, as a greedy procedure of adding features for a particular loss function.

2.1 Forward Stagewise Additive Modeling

The Forward Stagewise algorithm for a general loss function ϕ is as follows:

Algorithm 2 Forward Stagewise Additive Modeling

Input parameters: T, $\mathcal{H}$, ϕ

Initialize the classifier $f_0(x) = 0$
for $t = 1$ to T **do**

1. Compute

$$(h_t, \alpha_t) = \text{argmin}_{\alpha \in \mathbb{R}, h \in \mathcal{H}} \sum_{i=1}^{m} \phi(y_i, f_{t-1}(x_i) + \alpha h(x_i))$$

2. Update the classifier:

$$f_t(x) = f_{t-1}(x) + \alpha_t h_t(x)$$

end for
OUTPUT the classifier $\text{sgn}\,(f_T(x))$

The interpretation of this stagewise algorithm is that at each step the algorithm greedily adds an $h \in \mathcal{H}$ to the current hypothesis to minimize the ϕ-risk. This problem will show that AdaBoost is a forward stagewise algorithm, when we consider the exponential loss:

$$\phi(y, y') = \exp(-y'y)$$

The interpretation is that AdaBoost just greedily adds features $h \in \mathcal{H}$ at each iteration to minimize the exponential loss.

Prove that AdaBoost is equivalent to the forward stagewise algorithm, when ϕ is the exponential loss. The most natural proof is an inductive one. For $T = 1$, show that the algorithms are identical. Hint: For any fixed value of $\alpha > 0$, consider what the argminimizer would be (importantly, note that it is not a function of α). Now, for this h_t find the argminimizer over α. Compare the signed output (recall sgn function is not sensitive to the scale). Now for the inductive step, you will also need to understand how the weights $w_{t,i}$ are related to the stagewise algorithm.

2.2 The Exponential Loss

Now let us understand the exponential loss a little better (again, here $\mathcal{Y} = \{-1, 1\}$). To do this, find the Bayes optimal predictor under ϕ. In particular, find:

$$f^* \in \text{argmin}_f \mathbb{E}\left[\phi(y, f(x))\right]$$

where the minimization is over all functions. Clearly, the answer must be stated in terms of the underlying distribution $\Pr(x, y)$. Now interpret this quantity.

3 Lower Bounds for the Experts Problem

For the experts problem with bounded losses, we have seen an algorithm that has expected regret $O(\sqrt{T \log k})$ when there are k experts. Recall that at time t, an experts algorithm $\mathcal{A}$ chooses a distribution p_t over the experts (based on the past history) and chooses an expert i_t according to p_t. The regret $R_T(\mathcal{A})$ of an algorithm $\mathcal{A}$ is defined as

$$R_T(\mathcal{A}) := \sum_{t=1}^{T} l_{t,i_t} - \min_i \sum_{t=1}^{T} l_{t,i} \ .$$

Since we are deriving lower bounds, let us assume that the losses $l_{t,i}$ actually belong to $\{0, 1\}$. Prove the following two claims by choosing an appropriate distribution of the losses.

- For $T \leq \log_2 k$ and any algorithm $\mathcal{A}$,

$$\mathbb{E}\left[R_T(\mathcal{A})\right] \geq \frac{T}{2} \ .$$

- Suppose $k = 2$. For any algorithm $\mathcal{A}$, we have

$$\mathbb{E}\left[R_T(\mathcal{A})\right] = \Omega(\sqrt{T}) \ .$$

4 VC Dimension

- Show that the VC dimensions of convex polygons with d vertices in $\mathbb{R}^2$ is exactly $2d + 1$.
- Given an upper bound on the VC dimension of closed ellipsoids in $\mathbb{R}^d$.
- Give an example of a class $\mathcal{F}$ such that $|\mathcal{F}| = \infty$ yet $\mathrm{VCdim}(\mathcal{F}) = 1$.

5 Glivenko-Cantelli Theorem

Suppose we have an i.i.d. sample $X_1, \ldots, X_m$ drawn from some probability distribution over $\mathbb{R}$ with distribution function $F(t)$. That is,

$$F(t) := \mathbb{P}\left(X_i \leq t\right) \ .$$

The empirical distribution function is an approximation to F obtained using the sample,

$$\hat{F}_m(t) := \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\left[X_i \leq t\right] \ .$$

A famous theorem of probability theory (the Glivenko-Cantelli theorem) states that $\hat{F}_m$ converges *uniformly* to F. That is,

$$\sup_{t \in \mathbb{R}} \left| F(t) - \hat{F}_m(t) \right| \to 0 \ \text{ almost surely} \ .$$

Using results proved in the class, show that, with probability at least $1 - \delta$,

$$\sup_{t \in \mathbb{R}} \left| F(t) - \hat{F}_m(t) \right| \leq 2\sqrt{\frac{2\ln(m+1)}{m}} + \sqrt{\frac{\ln(1/\delta)}{2m}} \, .$$