CSE 546 HW #2

Sam Kowash

October 30, 2018

(1) A Taste of Learning Theory

1. Let $X \in \mathbb{R}^d$ a random feature vector, and $Y \in \{1, \ldots, K\}$ a random label for $K \in \mathbb{N}$ with joint distribution P_{XY}. We consider a randomized classifier $\delta(x)$ which maps a value $x \in \mathbb{R}^d$ to some $y \in \{1, \ldots, K\}$ with probability $\alpha(x, y) \equiv P(\delta(x) = y)$ subject to $\sum_{y=1}^K \alpha(x, y) = 1$ for all x. The risk of the classifier δ is

$$R(\delta) \equiv \mathbb{E}_{XY,\delta}\left[\mathbf{1}\{\delta(X) \neq Y\}\right],$$

which we should interpret as the expected rate of misclassification. A classifier δ is called deterministic if $\alpha(x, y) \in \{0, 1\}$ for all x, y. Further, we call a classifier δ_* a Bayes classifier if $\delta_* \in \arg\inf_\delta R(\delta)$.

If we first take the expectation over outcomes of δ, we find

$$R(\delta) = \mathbb{E}_{XY}\left[1 - \alpha(X, Y)\right],$$

since the indicator function is 1 except for the single outcome where $\delta(x) = y$, which occurs with probability $\alpha(x, y)$. It is then clear that minimizing $R(\delta)$ is equivalent to *maximizing* $\mathbb{E}_{XY}[\alpha(X, Y)]$.

2. We grab n data samples (x_i, y_i) i.i.d. from P_{XY} where $y_i \in \{-1, 1\}$ and $x_i \in \mathcal{X}$ where $\mathcal{X}$ is some set about which we make no further assumptions.