Reinforcement Learning February 27, 2019

Recitation 1

Lecturer:Yishay Mansour *TA: Lee Cohen*

Part 1: Learning Theory[1]

Recap: Probability Theory and Inequalities

Lemma 1 - The union bound
Let $A_1, A_2, ..., A_k$ be k different events (they can be dependent events). Then -

$$P(A_1 \cup ... \cup A_k) \leqslant P(A_1) + ... + P(A_k)$$

Lemma 2 - Markov's inequality
If X is a nonnegative random variable and $a > 0$, then the probability that X is no less than a, is no greater than the expectation of X divided by a:

$$P(X \geqslant a) \leqslant \frac{E(X)}{a}$$

Lemma 3 - Chebyshev's inequality
Let X be a random variable with finite expected value μ and finite non-zero variance σ. Then for any real number $k > 0$,

$$P\left(|X - \mu| \geqslant k\sigma\right) \leqslant \frac{1}{k^2}$$

Lemma 4 - Chernoff bound
Let $Z_1, ..., Z_m$ be m independent and identical distributed (i.i.d) random variables in range $[0, 1]$. Let $\hat{\phi} = (1/m)\sum_{i=1}^{m} Z_i$ be the empirical mean of these random variables, and let any $\gamma > 0$ be fixed. Then -

$$P\left(\left|\phi - \hat{\phi}\right| > \gamma\right) \leqslant 2exp(-2\gamma^2 m)$$

ERM: Empirical Risk Minimization

Let's restrict our attention to binary classification in which the labels are $y \in \{0, 1\}$. We assume we are given a training set $S = \left\{\left(x^{(i)}, y^{(i)}\right); i = 1, ..., m\right\}$ of size m, where the training examples $\left(x^{(i)}, y^{(i)}\right)$ are drawn i.i.d from some distribution D.
For a hypothesis h, we define the **empirical risk** (or the training error) to be -

$$\hat{\varepsilon}\left(h\right) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left\{h(x^{(i)}) \neq y^{(i)}\right\} \quad .$$

[1] taken from "cs229 notes4" by Andrew Ng©

Where $\mathbb{1}$ is the indicator function, i.e.,

$$\mathbb{1}\left(h(x^{(i)}) \neq y^{(i)}\right) = \begin{cases} 1 & h(x^{(i)}) \neq y^{(i)} \\ 0 & else \end{cases}$$

We also define the generalization error to be - $\varepsilon(h) = P_{(x,y)\sim D}\left(h(x) \neq y\right)$. Moreover, we define the **hypothesis class** H used by a learning algorithm to be the set of all classifiers considered by it (for example - in neural networks, then we could let H be the set of all classifiers representable by some neural network architecture.)

Let's start by considering a learning problem in which we have a finite hypothesis class $H = \{h_1, ..., h_k\}$ consisting of k hypotheses. Empirical risk minimization can now be thought of as a minimization over the class of functions H, in which the learning algorithm picks the hypothesis: $\hat{h} = \underset{h \in H}{argmin}\ \hat{\varepsilon}(h)$. The empirical risk minimization selects $\hat{h}$ to be whichever of these k functions has the smallest training error.

We would like to give guarantees on the generalization error of $\hat{h}$. Take $h_i \in H$, now consider a Bernoulli random variable Z whose distribution is defined as follows. We're going to sample $(x^{(j)}, y^{(j)}) \sim D$. Then we set $Z_j = \mathbb{1}\left\{h_i(x^{(j)}) \neq y^{(j)}\right\}$, Z_j indicate whether h_i misclassifies the sample $(x^{(j)}, y^{(j)})$.
The training error can be written - $\hat{\varepsilon}\left(h_i\right) = \frac{1}{m}\sum_{i=1}^{m} Z_j$. Thus $\hat{\varepsilon}(h)$ is exactly the mean of the m random variables Z_j that are drawn i.i.d from a Bernoulli distribution with mean $\hat{\varepsilon}(h_i)$. Hence, we can apply the Chernoff inequality, to obtain -

$$P\left(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma\right) \leq 2exp(-2\gamma^2 m) \quad .$$

This shows that, for our particular h_i, training error will be close to generalization error with high probability, assuming m is large. We can prove that this will be true for simultaneously for all $h \in H$. To do so, let A_i denote the event that $|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma$. We've already show that, for any particular A_i, it holds true that $P(A_i) \leq 2exp(-2\gamma^2 m)$. Thus, using the union bound, we have that -

$$\begin{gathered} P\left(\exists h \in H, (|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma\right) = \\ = P\left(A_1 \cup ... \cup A_k\right) \leqslant \sum_{i=1}^{k} P(A_i) \leqslant \\ \leqslant \sum_{i=1}^{k} 2exp(-2\gamma^2 m) = 2k \cdot exp(-2\gamma^2 m). \end{gathered}$$

And the probability that we estimate the error "well enough" for every $h_i \in H$ is

$$\begin{gathered} P\left(\forall h \in H, (|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma\right) = \\ P\left(\neg\exists h \in H, (|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma\right) \\ \geqslant 1 - 2k \cdot exp(-2\gamma^2 m). \end{gathered}$$

Now, we can calculate the minimal number of samples, m, to guarantee that with probability of at least $1 - \delta$, training error will be within γ of generalization error, for any given $\gamma, \delta \in [0, 1]$ and every $h \in H$. This is done by setting $\delta = 2k \cdot \exp(-2\gamma^2 m)$. Solving for m, we find that $m \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$, hence logarithmic in k, the number of hypotheses in H.