0.1 Recap

There are three significant parts of Learning: representation, optimization and generalization.

Recall the formulation of supervised learning, the assumption is: i.i.d. data $D(X, Y)$. For both training data $(x_1, y_1) \cdots (x_n, y_n)$, and test data $(x_{n+1}, y_{n+1}) \cdots$. And the most important thing is that all observation is training data.

Preparation We know two simple inequalities followed:

Theorem 0.1 (Markov Inequality) *X is a non-negative random variable, EX exists, then $\forall k > 0$*

$$\mathbb{P}(X \geq k) \leq \frac{EX}{k}$$

And

Theorem 0.2 (Chebyshev Inequality) *X is a random variable, $EX, \mathrm{Var}(X)$ exist, $\mathrm{Var}(X) = \sigma^2$, then $\forall k > 0$*

$$\mathbb{P}(|X - EX| \geq k) \leq \frac{\sigma^2}{k^2}$$

Then when we have more information about moments of X, can we get a better bound about the tail probability? For the case that we know finitely many moments of X, by adding a parameter t, we can actually get the following estimation:

Proposition 0.3 *Random variable $X \geq 0$, $EX, EX^2 \cdots EX^r$ exist, $\forall k > 0$*

$$\mathbb{P}(X \geq k) = \mathbb{P}(X^t \geq k^t), \forall t = 1, 2, \ldots, r$$

Hence,

$$\mathbb{P}(X \geq k) \leq \min_{t \in [r]} \frac{EX^t}{k^t}$$

For the case that we know all moments of X, we may need a better way to use all the information of moments. Similar to the use of generating function in solving a_n for $a_{n+2} = pa_{n+1} + qa_n$, a good way to use all the information of moments is to use moment generating function.

Definition 0.4 (Moment Generating Function) *X is a random variable. All moments of X exist. Then the moment generating function of X is defined as*

$$Ee^{tX} = 1 + tEX + \frac{t^2}{2}EX^2 + \cdots$$

Using moment generating function, with the method of adding a parameter t, we can get following well-known inequality, which gives us a much more better upper bound of the tail probability.

Theorem 0.5 (Chernoff Inequality) *X is a random variable, Ee^{tX} exists. Then $\forall k > 0$*

$$\mathbb{P}(X \geq k) = \mathbb{P}(e^{tX} \geq e^{tk}) \leq \frac{Ee^{tX}}{e^{tk}}$$

Hence,

$$\mathbb{P}(X \geq k) \leq \inf_{t > 0} \frac{Ee^{tX}}{e^{tk}}$$

0.2 Concentration Inequality

Consider that $X, X_1, X_2 \dots X_n$ are i.i.d. Bernoulli random variables, $EX = \mathbb{P}(X = 1) = p$.

Use Chebyshev inequality, we get

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - EX\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i)}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2}$$

But by central limit theorem, we guess the decay of this probability should be $e^{-O(n)}$. How to show it's true? Some concepts and the Chernoff inequality will be useful.

Definition 0.6 (Entropy) *X is a random variable with probability distribution $(p_1, p_2, \dots, p_n)$, then the entropy of X, denoted as $H(X)$, is defined by*

$$H(X) := -\sum_{i=1}^{n} p_i \log_2 p_i (bits)$$

Or

$$H(X) := -\sum_{i=1}^{n} p_i \ln p_i (nats)$$

Definition 0.7 (Relative Entropy) *$\mathcal{P} = (p_1, p_2, \dots, p_n)$ and $\mathcal{Q} = (q_1, q_2, \dots, q_n)$ are two probability distributions. The relative entropy is defined by*

$$D(P||Q) := \sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i} (bits)$$

Relative entropy is one way of describing the difference between two distributions. Actually it's non-negative.

Proposition 0.8 *By Jensen's inequality,*

$$D(P||Q) := -\sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i} \geq -\log_2\left(\sum_{i=1}^{n} p_i \frac{q_i}{p_i}\right) = 0$$

For the convenience of notation, we define

Definition 0.9 (Bernoulli Relative Entropy) *$\mathcal{P} = (p, 1-p), \mathcal{Q} = (q, 1-q)$ are Bernoulli distributions. The Bernoulli relative entropy is defined by*

$$D_B(P||Q) := p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}$$

0.2.1 Chernoff Bound

We are ready to get the Chernoff bound.

Theorem 0.10 (Chernoff Bound) *$X, X_1, X_2 \ldots X_n$ are i.i.d. Bernoulli random variables, $EX = \mathbb{P}(X = 1) = p, \mathrm{Var}(x) = p(1-p)$. By Chernoff inequality, we have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - EX \geq \varepsilon\right) = \mathbb{P}\left(\sum_{i=1}^{n} X_i \geq n(p+\varepsilon)\right) \leq \inf_{t>0} Ee^{t\sum_{i=1}^{n} X_i} e^{-nt(p+\varepsilon)}$$

$$Ee^{t\sum_{i=1}^{n} X_i} = E\left(\prod_{i=1}^{n} e^{tX_i}\right) = \prod_{i=1}^{n} Ee^{tX_i} = \left(Ee^{tX}\right)^n = \left(pe^t + 1 - p\right)^n$$

$$\inf_{t>0} Ee^{t\sum_{i=1}^{n} X_i} e^{-nt(p+\varepsilon)} = \inf_{t>0} \left(pe^t + 1 - p\right)^n e^{-nt(p+\varepsilon)} = e^{-nD_B(p+\epsilon||p)}$$

The proof of the last equation is left as exercise, so we omit the proof. It's not hard, and here is a hint for students: Function $\log(x)$ is strictly increasing, so the minimum points of $\log(f)$ and f are the same. Then get the minimum point of $\log(f)$ by derivation.

What if X is not a Bernoulli random variable? Intuitively, it should concentrate around the expectation more easily than Bernoulli case. To show this rigorously, we can use Jensen's inequality.

Proposition 0.11 *$X, X_1, X_2 \ldots X_n$ are i.i.d. random variables with $X \in [0,1], EX = p \in [0,1]$, then*

$$Ee^{tX} = Ee^{t(X\cdot 1 + (1-X)\cdot 0)} \leq E\left(Xe^t + 1 - X\right) = \left(pe^t + 1 - p\right)$$

So,

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - EX \geq \varepsilon\right) &\leq \inf_{t>0} Ee^{t\sum_{i=1}^{n} X_i} e^{-nt(p+\varepsilon)} \\ &\leq \inf_{t>0} \left(pe^t + 1 - p\right)^n e^{-nt(p+\varepsilon)} \\ &= e^{-nD_B(p+\epsilon||p)}\end{aligned}$$

The more general case that $X_1, X_2 \ldots X_n$ are independent random variables, $X_i \in [0,1], EX_i = p_i \in [0,1], p = \frac{1}{n}\sum_{i=1}^{n} p_i$ is left as exercise.