Machine Learning **Fall 2023**

Lecture 3: Concentration Inequality and Introduction to VC Theory

Lecturer: Liwei Wang *Scribe: Jiangyan Ma*

3.1 Concentration Inequality

3.1.1 Chernoff Bound

1. Let $X_1, X_2, \dots, X_n$ be i.i.d. Bernoulli random variables, $\mathrm{E}X = p$. Then,

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq \varepsilon\right] \leq \mathrm{e}^{-nD_B(p+\varepsilon\|p)}.$$

   Proof: Apply Chernoff's inequality and use $\mathrm{E}\mathrm{e}^{t\sum X_i} = (\mathrm{E}\mathrm{e}^{tX})^n = (p\mathrm{e}^t + 1 - p)^n$. ∎

2. Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables in $[0, 1]$, $\mathrm{E}X = p$. Then,

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq \varepsilon\right] \leq \mathrm{e}^{-nD_B(p+\varepsilon\|p)}.$$

   Proof: By Jensen's inequality, $\mathrm{E}\mathrm{e}^{t\sum X_i} = (\mathrm{E}\mathrm{e}^{tX})^n \leq (p\mathrm{e}^t + 1 - p)^n$. ∎

3. Let $X_1, X_2, \dots, X_n$ be independent random variables in $[0, 1]$, $\mathrm{E}X_i = p_i$. Let $p = \frac{1}{n}\sum_{i=1}^{n} p_i$. Then

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq \varepsilon\right] \leq \mathrm{e}^{-nD_B(p+\varepsilon\|p)}.$$

   Proof: By the AM-GM inequality,

$$\mathrm{E}\mathrm{e}^{t\sum X_i} = \prod_{i=1}^{n} \mathrm{E}\mathrm{e}^{tX_i} \leq \prod_{i=1}^{n} (p_i\mathrm{e}^t + 1 - p_i) \leq (p\mathrm{e}^t + 1 - p)^n.$$

∎

3.1.2 Additive Chernoff Bound

Since $D_B(p + \varepsilon \,\|\, p) \geq 2\varepsilon^2$ (left as homework), we also have

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq \varepsilon\right] \leq \mathrm{e}^{-2n\varepsilon^2}$$

in all cases.

3.1.3 Hoeffding Inequality

Let $X_1, X_2, \ldots, X_n$ be independent random variables, $X_i \in [a_i, b_i]$, $\mu := \mathrm{E}\frac{1}{n}\sum X_i$. Then

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \varepsilon\right] \leq \mathrm{e}^{-\frac{2n\varepsilon^2}{\sum(b_i-a_i)^2}}.$$

3.1.4 Draw without Replacement

Assume we have N numbers $a_1, a_2, \ldots, a_N \in \{0,1\}$. Randomly draw n numbers from $a_1, \ldots, a_N$.

1. If we *draw with replacement*, it is the same as the first case in Section 3.1.1.
2. If we *draw without replacement*, let $X_1, \ldots, X_n$ be the random variables obtained from draw with replacement, $Y_1, \ldots, Y_n$ be the random variables obtained from draw without replacement. We would like to prove $\frac{1}{n}\sum Y_i$ concentrates faster than $\frac{1}{n}\sum X_i$. In other words, we wish to prove

$$\mathrm{E}\mathrm{e}^{t(Y_1+\cdots+Y_n)} \leq \mathrm{E}\mathrm{e}^{t(X_1+\cdots+X_n)}. \tag{3.1}$$

   Expanding both sides gives us

$$\mathrm{E}\mathrm{e}^{t(Y_1+\cdots+Y_n)} = 1 + t\mathrm{E}\sum_i Y_i + \frac{t^2}{2}\mathrm{E}\sum_{i,j} Y_iY_j + \ldots,$$
$$\mathrm{E}\mathrm{e}^{t(X_1+\cdots+X_n)} = 1 + t\mathrm{E}\sum_i X_i + \frac{t^2}{2}\mathrm{E}\sum_{i,j} X_iX_j + \ldots.$$

   Apparently $\mathrm{E}\sum_i Y_i = \mathrm{E}\sum_i X_i$, $\mathrm{E}Y_iY_j = \Pr[Y_i=1, Y_j=1] \leq \Pr[X_i=1, X_j=1] = \mathrm{E}X_iX_j$, etc. Thus Equation (3.1) holds.

3.1.5 McDiarmid Lemma

Assume $f(x_1, \ldots, x_n)$ is a *stable function*, that is, for $\forall x_1, \ldots, x_n$, $\forall i$, $\forall x_i'$, we have

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i.$$

Then for independent random variables $X_1, \ldots, X_n$,

$$\Pr\left[f(X_1, \ldots, X_n) - \mathrm{E}f(X_1, \ldots, X_n) \geq \varepsilon\right] \leq \mathrm{e}^{-\frac{\varepsilon^2}{\sum c_i^2}}.$$

3.2 VC Theory: The First Theory of Generalization

3.2.1 Universal Approximation Theorem

Recall: Generalization, performance difference between training and test data. (over-fitting)

Representation power of Deep Neural network: Given any continuous target function $f(x)$, $x \in \mathbb{R}^d$. For any $\varepsilon > 0$, there exists a neural network $\mathrm{NN}(\cdot)$, such that $\|f(x) - \mathrm{NN}(x)\| \leq \varepsilon$. This is called the Universal Approximation Theorem.

3.2.2 An Oversimplified Setting

Suppose f is the learned classifier from training data $(x_1, y_1), \ldots, (x_n, y_n)$. The training error can be formulated as

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{I}[y_i \neq f(x_i)]$$

while the test error can be formulated as

$$\Pr[Y \neq f(X)] = \mathrm{E}\left(\mathrm{I}[Y \neq f(X)]\right).$$

The training error is the average of n Bernoulli random variables, while the test error is its expectation. By the concentration property, we expect the training error to converge to the test error as n increases. Then why would there be over-fitting? The reason is that f is learned from $(x_1, y_1), \ldots, (x_n, y_n)$, leading to $f(x_1), \ldots, f(x_n)$ being non-independent.

Let's consider a setting where we collect training data $(x_i, y_i)^n$ and learn $f \in \mathcal{F}$ to fit the training data. We call $\mathcal{F}$ the *hypothesis space* (A set of classifier, or a model). We assume $|\mathcal{F}| < \infty$. Under this oversimplified assumption, we can estimate the error using the union bound:

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{I}[y_i \neq f(x_i)] - \Pr[Y \neq f(X)] \geq \varepsilon\right] \leq |\mathcal{F}|\mathrm{e}^{-2n\varepsilon^2}.$$

Larger hypothesis space implies larger upper bound and thus larger probability of over-fitting. However, we know in realistic settings the hypothesis space is infinitely large. But from this we learnt that the size of $\mathcal{F}$ Highly determines the gap between two sets. The purpose of VC theory is to study the generalization error when $|\mathcal{F}| = \infty$.