CSE522, Winter 2011, Learning Theory Lecture 4 - 01/16/2011

Estimation VS Approximation

Lecturer: Ofer Dekel *Scribe: Yanping Huang*

1 Some inequalities

Chebyshev's inequality: Let Z be a random variable with expected value μ and variance $\sigma^2 < \infty$. Then $\forall\ \epsilon > 0$, we have $Pr(|Z-\mu| \geq \epsilon) \leq \sigma^2/\epsilon^2$.

Proof. Let $X = (Z-\mu)^2 \geq 0$, then $E(X) = E[(Z-\mu)^2] = \sigma^2$. And from Markov's inequality we have

$$P(X \geq \epsilon^2) = P((Z-\mu)^2 \geq \epsilon^2) = P(|Z-\mu| \geq \epsilon) \leq E(X)/\epsilon^2 = \frac{\sigma^2}{\epsilon^2} \tag{1}$$

□

Hoeffding's inequality: Let $Z_1 \ldots Z_m$ are independent random variables. Assume that the Z_i are almost surely bounded: $Pr(Z_i \in [a,b]) = 1$ where $b-a=c$. Then, for the average of these variables $Z = \frac{1}{m}\sum_{i=1}^m Z_i$, we have $P(Z - E(Z) \geq \epsilon) \leq \exp(-\frac{2\epsilon^2}{c^2})$ for any $\epsilon > 0$.

Proof. Without loss of generality we assume $E(Z) = 0$ and $c = 1$ (or we can just let $Z' = \frac{Z-E(Z)}{c}$). Then from Markov's inequality we have

$$\begin{aligned}
Pr(Z \geq \epsilon) &= Pr(e^{4m\epsilon Z} \geq e^{4m\epsilon^2}) \leq \frac{E[e^{4m\epsilon Z}]}{e^{4m\epsilon^2}} \\
&= \frac{E[\prod_i e^{4\epsilon Z_i}]}{e^{4m\epsilon^2}} = \frac{\prod_i E[e^{4\epsilon Z_i}]}{e^{4m\epsilon^2}} \quad \text{(Second equality holds only when } Z_i\text{s are independent)} \\
&\leq \frac{\prod_i e^{2\epsilon^2}}{e^{4m\epsilon^2}} \quad \text{(Using Jensen's inequality)} \\
&= \exp(-2m\epsilon^2)
\end{aligned} \tag{2}$$

□

McDiarmid's Inequality: Suppose $Z_1, \ldots, Z_m$ are independent, the vector $Z = \{Z_1, Z_2, \ldots, Z_m\}$ and assume that f satisfies

$$\sup_{z_1,\ldots,z_m,z_i'} |f(z_1,\ldots,z_m) - f(z_1,\ldots,z_{i-1},z_i' z_{i+1},\ldots,z_m)| \leq \frac{d}{m}. \tag{3}$$

for any $1 \leq i \leq m$. In other words, replace the i-th coordinate x_i by some other value changes the value of f by at most d/m. Then f has the $\frac{d}{m}$-bounded property and satisfies the following inequality

$$Pr(f(Z) - E[f(Z)] \geq \epsilon) \leq \exp(-\frac{2m\epsilon^2}{d^2}) \tag{4}$$

for any $\epsilon > 0$.

2 Generalization Bounds for finite hypothesis space

2.1 Chernoff bound for a fixed hypothesis

In the context of machine learning theory, let a sample set $\mathcal{S} = Z$, each $Z_i = l(h; (x_i, y_i))$, and $f(Z) = \frac{1}{m}\sum_{i=1}^{m} l(h; (x_i, y_i)) = l(h; \mathcal{S})$. For a fixed hypothesis h, a loss function $l \in [0, c]$ and $\epsilon > 0$, we then have $Pr(|l(h; \mathcal{D}) - l(h; \mathcal{S})| \geq \epsilon) \leq 2\exp(-\frac{2m\epsilon^2}{c^2})$, where $\mathcal{D}$ denotes the distribution of the examples (x, y) and $\mathcal{S}$ is a sample set drawn from $\mathcal{D}$ with size m. Let $\delta = \exp(-\frac{2m\epsilon^2}{c^2})$. Then with probability at least $1 - \delta$, we have $|l(h; \mathcal{D}) - l(h; \mathcal{S})| \leq c\sqrt{\frac{\log(2/\delta)}{2m}}$.

The above inequality says that for each hypothesis $h \in \mathcal{H}$, there exists a set $\mathcal{S}$ of samples that satisfies the bound $c\sqrt{\frac{2/\delta}{2m}}$ with probability at least $1-\delta$. However, these sets may be different for different hypotheses. For a fixed observed sample $\mathcal{S}^*$, this inequality may not hold for all hypotheses in $\mathcal{H}$ including the hypothesis $h_{ERM} = \arg\min_{h\in\mathcal{H}} l(h; \mathcal{S}^*)$ with minimum empirical risk. Only some hypotheses in $\mathcal{H}$ (not necessarily h_{ERM}) will satisfy this inequality.

2.2 Uniform bound

To overcome the above limitation, we need to derive a *uniform* bound for all hypotheses in $\mathcal{H}$. As shown in Fig 1, we define the set $\Omega_i = \{\mathcal{S} \sim \mathcal{D} : |l(h_i; \mathcal{D}) - l(h_i; \mathcal{S})| > \epsilon\}$ be the set contains all "bad" samples for which the bound fails. For each i, $Pr(\Omega_i) \leq \delta$. If $|\mathcal{H}| = k$, we can write $Pr(\Omega_1 \cup \ldots \cup \Omega_k) \leq \sum_{i=1}^{k} Pr(\Omega_i)$. As a result, we obtain the uniform bound:

$$\begin{aligned} P(\forall h \in \mathcal{H} : l(h; \mathcal{D}) - l(h; \mathcal{S}) \leq \epsilon) &\leq 1 - \sum_i P(|l(h_i; \mathcal{D} - l(h_i, \mathcal{S})| > \epsilon)| \\ &= 1 - 2k\exp(-\frac{2m\epsilon^2}{c^2}). \end{aligned} \tag{5}$$

Finally we have the theorem for a uniform bound

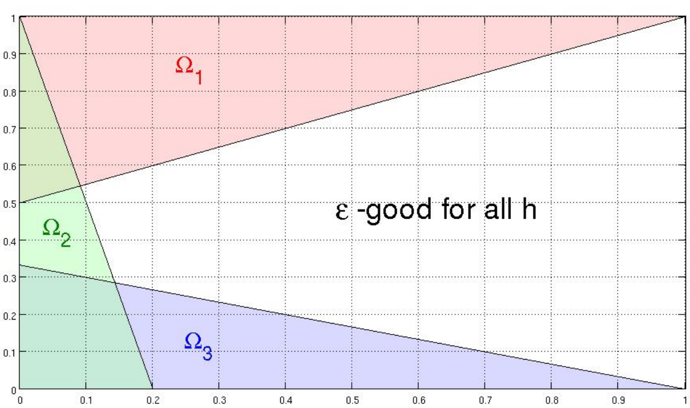

Figure 1: Set diagram for Ω_i and the ϵ uniformly good set of samples.

Theorem 1. *If the size of the hypothesis space $|\mathcal{H}| = k$, the loss function $l \in [0, c]$, and $\mathcal{S}$ is the sample set drawn from distribution $\mathcal{D}$ with $|\mathcal{S}| = m$. Then $\forall\delta > 0$ and $\forall h \in \mathcal{H}$, with probability at least $1 - \delta$,*

$$|l(h; \mathcal{D} - l(h; \mathcal{S})| \leq c\sqrt{\frac{\log(2/\delta) + \log(k)}{2m}} \tag{6}$$

2.3 Excess Risk

Define the excess risk for any hypothesis $h \in \mathcal{H}$ be $l(h;\mathcal{D}) - \min_{h\in\mathcal{H}} l(h;\mathcal{D})$. From Theorem 1 we have $l(h_{ERM};\mathcal{D}) \leq \min_{h\in\mathcal{H}} l(h;\mathcal{D}) + 2\epsilon$, as shown in Fig 2, where $\epsilon = c\sqrt{\frac{\log(2/\delta)+\log(k)}{2m}}$.

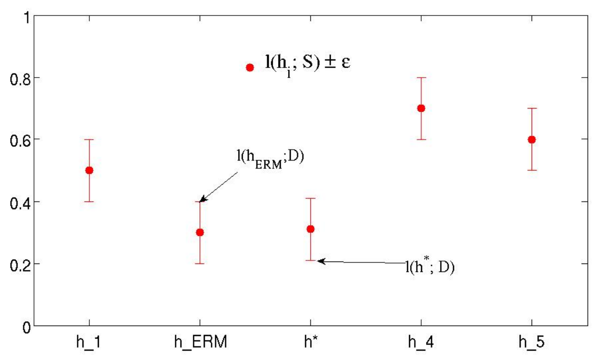

Figure 2: The empirical risk $l(h_i;\mathcal{S})$ and the range for the true risk $l(h_i;\mathcal{D})$. The difference between $l(h_{ERM};\mathcal{D})$ and $l(h^*;\mathcal{D})$ is at most 2ϵ where $h^* = \arg\min_{h\in\mathcal{H}} l(h;\mathcal{D})$

2.4 Estimation VS Approximation

First we define the Bayesian risk $\min_{\text{all possible h}} l(h;\mathcal{D})$ and the Bayesian hypothesis $\arg\min l(h;\mathcal{D})$, a hypothesis that attains the Bayesian risk. Sometimes some errors may be inevitable, the Bayesian risk may be strictly positive.

Take the binary classification task for example, where $y \in \{-1, 1\}$ be the lab, and the loss function is zero-one $l(h;(x,y)) = \mathbf{1}_{h(x)\neq y}$. The risk for h can be written as:

$$\begin{aligned} E[\mathbf{1}_{h)x\neq y}] &= E_x[E[\mathbf{1}_{h(x)\neq y}|X=x]] \qquad (7) \\ \text{where} \quad [E[\mathbf{1}_{h(x)\neq y}|X=x] &= Pr(h(x)\neq y|X=x) \\ &= \begin{cases} Pr(Y=-1|X=x) & \text{if } h(x)=+1 \\ Pr(Y=+1|X=x) & \text{if } h(x)=-1 \end{cases} \end{aligned}$$

We have the Bayesian hypothesis that minimizes the above risk

$$h_{Bayes} = \begin{cases} +1 & \text{if } Pr(y=1|x=x) > 0.5 \\ -1 & \text{otherwise} \end{cases} \qquad (8)$$

if we know the distribution $\mathcal{D} = Pr(Y|X)Pr(X)$.

For a hypothesis space $\mathcal{H}$, we define the approximation error as $l(h^*,\mathcal{D}) - l(h_{Bayes},\mathcal{D})$ and the estimation error as $l(h_{ERM};\mathcal{D}) - l(h^*;\mathcal{D})$ where $h^* = \arg\min_{h\in\mathcal{H}} l(h;\mathcal{D})$.

We observe that as the size of hypothesis space $k = |\mathcal{H}|$ increases, the approximation error may decreases while the estimation error will increase. Consider the following two scenarios demonstrated in Fig 3,

- Scenario 1: $k = 10^{10}$, $h_{ERM} = \arg\min_{h\in\mathcal{H}} l(h;S)$. Then with probability at least $1-\delta$, from the excess risk theorem we have

$$l(h_{ERM},\mathcal{D}) \leq \min_{h\in\mathcal{H}} l(h,\mathcal{D}) + 2c\sqrt{\frac{2\log(2/\delta)+\log(10^{10})}{2m}}$$

.

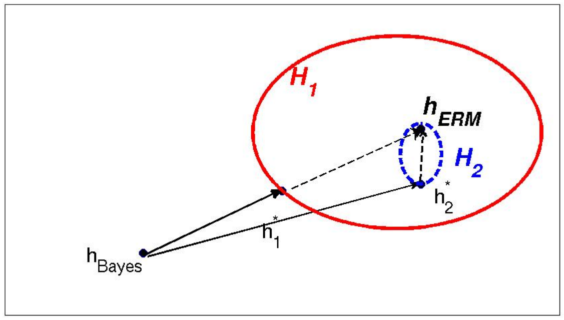

Figure 3: The approximation errors are shown in solid arrows pointing from the Bayesian hypothesis h_{Bayes} to h_i^*, where $h_i^* = \arg\min_{h_i \in \mathcal{H}_i} l(h_i; \mathcal{D})$. The estimation errors are shown in dotted arrows pointing from h_i^* to h_{ERM}. Suppose $h_{ERM} \in \mathcal{H}_1 \cap \mathcal{H}_2$ and $|\mathcal{H}_2| << |\mathcal{H}_1|$. The figure demonstrates the effect of the size of hypothesis space on the approximation error and the estimation error.

- Scenario 2: $k = 3$, $\mathcal{H} = \{h_{ERM}, h_1, h_2\}$. Now we have

$$l(h_{ERM}, \mathcal{D}) \leq \min_{h \in \mathcal{H}} l(h, \mathcal{D}) + 2c\sqrt{\frac{2\log(2/\delta) + \log(3)}{2m}}$$

3 Generalization Bound for infinite hypothesis space

Theorem 2. *If the size of the hypothesis space is infinite, $|\mathcal{H}| = \infty$, the loss function $l \in [0, c]$, and $\mathcal{S}$ is the sample set drawn from distribution $\mathcal{D}$ with $|\mathcal{S}| = m$. Then $\forall \delta > 0$ and $\forall h \in \mathcal{H}$, with probability at least $1 - \delta$,*

$$|l(h; \mathcal{S}) - l(h; \mathcal{D})| \leq \epsilon(\delta) \tag{9}$$

Proof. To apply Hoeffding's inequality 2, we define $f(S) = \max_{h \in \mathcal{H}}[l(h; \mathcal{D} - l(h; \mathcal{S})]$.

First we show that $f(\mathcal{S})$ is $\frac{c}{m}$ bounded. $\forall h$, we change one example in $\mathcal{S} \rightarrow \mathcal{S}'$. $l \in [0, c]$, thus $|l(h, \mathcal{S}') - l(h, \mathcal{S}| \leq \frac{c}{m}$. $l(h, \mathcal{D})$ remains the same, we have $|f(\mathcal{S}) - f(\mathcal{S};)| \leq \frac{c}{m}$.

Next we apply McDiarmid's inequality,

$$\begin{aligned} Pr(|f(\mathcal{S}) - E[f(\mathcal{S})]| &\geq \epsilon) \leq 2\exp(-\frac{2m\epsilon^2}{c^2}) = \delta \\ f(s) &\leq E[f(\mathcal{S})] + c\sqrt{\log(2/\delta)/2m} \end{aligned}$$

where $E[f(\mathcal{S})] = E_{\mathcal{S}}\{\max_{h \in \mathcal{H}}[l(h; \mathcal{D} - l(h; \mathcal{S})]\}$. The expectation is taken over all possible samples.

Third, we show that $\max_{i \in \mathcal{I}} E(X_i) \leq E(\max_{i \in \mathcal{I}} X_i)$. For $\forall j \in \mathcal{I}$, $x_j \leq \max x_i$. Then $E(X_j) \leq E(\max_i X_i)$. Finally we have $\max_j E(X_j) \leq E(\max_i X_i)$. Using this lemma, we can show

$$\begin{aligned} E_{\mathcal{S}}[f(\mathcal{S})] &= E_{\mathcal{S}}[\max_h[l(h; \mathcal{D}) - l(h; \mathcal{S})]] \\ &= E_{\mathcal{S}}[\max_h[E_{\mathcal{S}'} - l(h; \mathcal{S})]] \\ &\leq E_{\mathcal{S}} E_{\mathcal{S}'}[\max_h(l(h; \mathcal{S}' - l(h; \mathcal{S}))] \end{aligned} \tag{10}$$

□