CSE522, Winter 2011, Learning Theory Lecture 15 - 02/22/2011

Doob Martingales and online learning

Lecturer: Ofer Dekel *Scribe: Karthik Mohan*

1 Background on Expectation

Conditional probability: $\mathsf{P}(A|B) = \frac{\mathsf{P}(A\cap B)}{\mathsf{P}(B)}$.
Expectation: Let $Z : \Omega \to \mathbb{R}$ be a random variable. $\mathsf{E}[Z] = \sum_{z\in\Omega} \mathsf{P}(Z = z)z$.
Conditional Expectation 1: $\mathsf{E}[Z|Y = y] = \sum_z \mathsf{P}(Z = z|Y = y)z$. This expectation is a function of y and hence a number.
Conditional Expectation 2: $\mathsf{E}[Z|Y] = \sum_z \mathsf{P}(Z = z|Y)z$. This expectation is a function of Y and hence a random variable.
Example 1

$$X = \begin{cases} 1 & \text{w.p. } 1/2 \\ 0 & \text{w.p. } 1/2 \end{cases} \qquad Y = \begin{cases} 2 & \text{w.p. } 1/2 \\ 0 & \text{w.p. } 1/2 \end{cases} \tag{1}$$

Let $Z = X + Y$. Note that, $\mathsf{E}[Z] = \frac{1}{4}[0 + 1 + 2 + 3] = 3/2$.

$$\mathsf{E}[Z|X] = \begin{cases} \mathsf{E}[Z|X = 0] & \text{w.p.} & \mathsf{P}(X = 0) = 1/2 \\ \mathsf{E}[Z|X = 1] & \text{w.p.} & \mathsf{P}(X = 1) = 1/2 \end{cases} \tag{2}$$

Note that $\mathsf{E}[Z|X] = \mathsf{E}[Y] + X = 1 + X$.

Lemma 1 (Law of Total Expectation)**.** $\forall X, Y$ $\mathsf{E}[X] = \mathsf{E}[\mathsf{E}[X|Y]]$.

Proof.

$$\begin{aligned} \mathsf{E}[X] &= \textstyle\sum_x \mathsf{P}(X = x)x \\ &= \textstyle\sum_x(\sum_y \mathsf{P}(X = x, Y = y))x \\ &= \textstyle\sum_x(\sum_y \mathsf{P}(X = x|Y = y)\mathsf{P}(Y = y))x \\ &= \textstyle\sum_y \mathsf{P}(Y = y)\sum_x \mathsf{P}(X = x|Y = y)x \\ &= \textstyle\sum_y \mathsf{P}(Y = y)\mathsf{E}[X|Y = y] \\ &= \mathsf{E}[\mathsf{E}[X|Y]] \end{aligned} \tag{3}$$

where the second equality follows from total probability. □

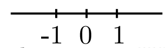
Example 2 Let $U_1, U_2, \ldots, U_m$ be random variables. Let $X = f(U_1, U_2, \ldots, U_m)$ and $Z = E[X|U_1, U_2, \ldots, U_k]$. Then, $\mathsf{E}[X] = \mathsf{E}_{U_1,U_2,\ldots,U_k}[\mathsf{E}_{U_{k+1},\ldots,U_m}[X]] = \mathsf{E}_{U_1,\ldots,U_k}[\mathsf{E}[X|U_1, U_2, \ldots, U_k]] = \mathsf{E}[Z]$. The previous expression also follows from the law of total expectation.

2 Background on Martingales

Definition 2. *A sequence of random variables* $(W_i)_{i=0}^m$ *is a martingale w.r.t another sequence of random variables* $(U_i)_{i=1}^m$ *if*

$$\begin{aligned} & \mathsf{E}[|W_i|] && < \infty \\ \forall i \quad & \mathsf{E}[W_{i+1}|U_1, U_2, \ldots, U_i] && = W_i \end{aligned} \tag{4}$$

Example 3 Consider a random walk on real line:



Let W_i denote the position after i steps with the inital position $W_0 = 0$. Let the random walk be described by:

$$W_{i+1} = \begin{cases} W_i + 1 & \text{w.p. } 1/2 \\ W_i - 1 & \text{w.p. } 1/2 \end{cases} \tag{5}$$

Let,

$$U_i = \begin{cases} 1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases} \tag{6}$$

Note that, $W_{i+1} = W_i + U_{i+1}$, $i = 0, \ldots, m-1$. Also, $(W_i)_{i=0}^m$ is a martingale w.r.t $(U_i)_{i=1}^m$ since $\mathsf{E}[|W_i|] < \infty$ and $\mathsf{E}[W_{i+1}|U_1, U_2, \ldots, U_i] = \mathsf{E}[\sum_{j=1}^{i+1} U_j | U_1, U_2, \ldots, U_i] = \sum_{j=1}^{i} U_i + \mathsf{E}[U_{i+1}] = W_i$.

3 Doob Martingale

Definition 3. *Let $(U_i)_{i=1}^m$ be a sequence of random variables and $f(U_1, U_2, \ldots, U_m)$ be a function such that $\mathsf{E}[|f(U_1, U_2, \ldots, U_m)|] < \infty$. The doob martingale is defined as $(W_i)_{i=0}^m$ where, $W_i = \mathsf{E}[f(U_1, \ldots, U_m)|U_1, U_2, \ldots, U_i]$ for $1 \leq i \leq m$ and $W_0 = \mathsf{E}[f(U_1, \ldots, U_m)]$.*

Note that the randomness is incrementally revealed in W_i as i goes from 0 (W_0 is a scalar) to m (W_m is a function of $U_1, U_2, \ldots, U_m$).

Theorem 4. *A doob martingale is a martingale.*

Proof. For a doob martingale, $\mathsf{E}[|W_i|] = \mathsf{E}_{U_1,\ldots,U_i}[|\mathsf{E}_{U_{i+1},\ldots,U_m}[f(U_1, \ldots, U_m)]|] \leq \mathsf{E}[|f(U_1, \ldots, U_m)|] < \infty$. Also, $\mathsf{E}[W_{i+1}|U_1, \ldots, U_i] = \mathsf{E}[\mathsf{E}[f(U_1, \ldots, U_m)|U_1, \ldots, U_{i+1}]|U_1, \ldots, U_i] = \mathsf{E}_{U_{i+1},\ldots,U_m}[\mathsf{E}_{U_{i+2},\ldots,U_m}[f(U_1, \ldots, U_m)]] = \mathsf{E}_{U_{i+1},\ldots,U_m}[f(U_1, \ldots, U_m)] = \mathsf{E}[f(U_1, \ldots, U_m)|U_1, \ldots, U_i] = W_i$. □

Definition 5. *We say that a martingale $(W_i)_{i=0}^m$ has $\frac{c}{m}$-bounded differences ($\frac{c}{m}$-Lipschitz) if $|W_{i+1} - W_i| \leq \frac{c}{m}$.*

Fact 6. ***Hoeffding-Azuma**: For Doob Martingales with $\frac{c}{m}$-bounded differences, $\mathsf{P}(W_m - W_0 < \epsilon) \leq e^{\frac{-2m\epsilon^2}{c^2}}$. Thus, for all $\delta > 0$ w.p. $\geq 1 - \delta$ over the random draws $U_1, \ldots, U_m$,*

$$W_m \leq W_0 + c\sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{7}$$

For general martingales, $\mathsf{P}(W_m - W_0 > \epsilon) \leq e^{-\frac{m\epsilon^2}{2c^2}}$.

4 Online Learning

Assume the samples, $S \in \mathcal{D}^m$. Let $(\mathcal{X}, \mathcal{Y})$ denote the set of all possible feature vectors and labels respectively. The general form of online learning is as follows:

1. Start with $h_0 \in \mathcal{H}$.

2. For $i = 1, 2, \ldots, m$,

(a) Receive $x_i \in \mathcal{X}$.

(b) Predict $h_{i-1}(x_i)$.

(c) Receive $y_i \in \mathcal{Y}$.

(d) Suffer loss $l(h_{i-1};(x_i,y_i))$.

(e) Update $h_i \leftarrow \mathcal{A}(h_{i-1};(x_i,y_i))$ (where $\mathcal{A}$ denotes the online algorithm).

Remarks
Now let $f(S) = \frac{1}{m}[\sum_{i=1}^{m} l(h_{i-1};\mathcal{D}) - \sum_{i=1}^{m} l(h_{i-1};(x_i,y_i))]$. Define $\sum_{i=1}^{m} l(h_{i-1};(x_i,y_i))$ to be the cumulative loss. We make the following remarks.

1. $l(h_{i-1};\mathcal{D})$ is a random variable (since h_{i-1} is a function of $(x_j,y_j)_{j=1}^{i-1}$ that are drawn i.i.d from $\mathcal{D}$).

2. $l(h_{i-1};(x_i,y_i))$ is a random variable (with randomness in $(x_j,y_j)_{j=1}^{i}$).

3. $\mathsf{E}[l(h_{i-1};(x_i,y_i))|(x_j,y_j)_{j=1}^{i-1}] = \mathsf{E}_{(x_i,y_i)}[l(h_{i-1};(x_i,y_i))] = l(h_{i-1};\mathcal{D})$.

4. If $l \in [0,c]$ then $W_i = \mathsf{E}[f(S)|(x_j,y_j)_{j=1}^{i}]$ is a Doob martingale with $\frac{c}{m}$-bounded differences.

Theorem 7. *$\forall \delta > 0$, w.p. $\geq 1-\delta$ over the random (i.i.d) sampling of $S \in \mathcal{D}^m$,*

$$f(S) \leq c\sqrt{\frac{\log(\frac{1}{\delta})}{2m}} \tag{8}$$

Proof. Remark 4 states that $W_i = \mathsf{E}[f(S)|(x_j,y_j)_{j=1}^{i}]$ is a doob martingale. $W_0 = \mathsf{E}[f(S)]$ and $W_m = f(S)$. Thus, by Hoeffding-Azuma inequality(Fact 6), it follows that w.p. $\geq 1-\delta$,

$$\begin{aligned} W_m - W_0 &\leq c\sqrt{\frac{\log(\frac{1}{\delta})}{2m}} \\ f(S) - \mathsf{E}[f(S)] &\leq c\sqrt{\frac{\log(\frac{1}{\delta})}{2m}} \end{aligned} \tag{9}$$

It also follows from Remark 3 that $\mathsf{E}[f(S)] = 0$ and hence the theorem follows. □

Remark
Let $\frac{1}{m}[\sum_{i=1}^{m} l(h_{i-1};\mathcal{D})]$ denote the average risk of $\{h_0,\ldots,h_{m-1}\}$. Also let $\mathcal{Q}$ be the uniform distribution over $\{h_0,\ldots,h_{m-1}\}$. Then, the previous theorem implies that w.p. $\geq 1-\delta$,

$$l(\mathcal{Q};\mathcal{D}) \leq \frac{1}{m}\sum_{i=1}^{m} l(h_{i-1};(x_i,y_i)) + c\sqrt{\frac{\log(\frac{1}{\delta})}{2m}} \tag{10}$$