CS281B/Stat241B (Spring 2006) Statistical Learning Theory **Lecture: 17**

Bounded Differences Inequality and Loss Classes

Lecturer: Peter Bartlett *Scribe: Jeremy Schiff*

1 Recap

Last lecture we discussed Hoeffding's Inequality for sums of bounded independent random variables:

$$S_n = \sum_{i=1}^{n} X_i \text{ where } X_i \in [a_i, b_i] \ .$$

Using Chernoff trick, we showed:

$$P\left[S_n - \mathbb{E}[S_n] > t_n\right] \leq \exp \frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}$$

This exploits the independence for relating moment generating functions.

2 Bounded difference inequality

Theorem 2.1. If $f : X^n \to \mathbb{R}$ satisfies $|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$, then for independent random variables $X_1, \ldots, X_n$

$$P\left[f - \mathbb{E}f \geq t\right] \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

This is a generalization of the Hoeffding Inequality.

This result is directly applicable to problems such as

- Largest Common Subsequence Problem - Altering one value in one of the sequences can affect the longest subsequence by at most 1.
- Traveling Salesman Problem - bounding the influence of any particular point on the length of the shortest tour.

Define:

$$V_i = \mathbb{E}[f|X_1, \ldots, X_i] - \mathbb{E}[f|X_1, \ldots, X_{i-1}]$$

Then:

$$f - \mathbb{E}f = \sum_{i=1}^{n} V_i$$

We can compute the bound:

$$
\begin{aligned}
P[f - \mathbb{E}f \geq t] \quad & \leq \inf_{s>0} \exp(-st)\mathbb{E}\exp(s \textstyle\sum_{i=1}^n V_i) \\
& = \inf_{s>0} \exp(-st)\mathbb{E}\left[\mathbb{E}[\exp(s \textstyle\sum_{i=1}^n V_i)]|X_1, \ldots, X_{n-1}\right] \\
& = \inf_{s>0} \exp(-st)\mathbb{E}\left[\exp(s \textstyle\sum_{i=1}^{n-1} V_i)\mathbb{E}[\exp(sV_n)]|X_1, \ldots, X_{n-1}]\right] \\
& \text{This next step is explained in detail below} \\
& \leq \inf_{s>0} \exp(-st)\mathbb{E}\left[\exp(s \textstyle\sum_{i=1}^{n-1} V_i)\exp(s^2 c_i^2/8)\right] \\
& \leq \inf_{s>0} \exp(-st + s^2/8 \textstyle\sum_{i=1}^{n-1} c_i^2) \\
& = \exp(-2t^2/\textstyle\sum_{i=1}^{n} c_i^2)
\end{aligned}
$$

Where

$$t = \frac{s}{4}\sum_{i=1}^{n} c_i^2$$

Define

$$L_i = \inf_x \mathbb{E}[f|X_1, \ldots, X_{i-1}, x] - E[f|X_1, \ldots, X_{n-1}]$$

$$U_i = \sup_x \mathbb{E}[f|X_1, \ldots, X_{i-1}, x] - E[f|X_1, \ldots, X_{n-1}]$$

Then obviously $L_i \leq V_i \leq U_i$.

We can see (using independence; also see notes for previous lecture):

$$
\begin{aligned}
& \quad U_i - L_i \\
& = \sup_x \mathbb{E}[f|X_1, \ldots, X_{i-1}, x] - E[f|X_1, \ldots, X_{n-1}] - \inf_x \mathbb{E}[f|X_1, \ldots, X_{i-1}, x] - E[f|X_1, \ldots, X_{n-1}] \\
& = \sup_x \textstyle\int f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x', x_{i+1}, \ldots, x_n) dP(x_{i+1}, ..., x_n) \\
& \leq c_i
\end{aligned}
$$

3 Bounded Difference Inequality Example

For Independent Random Variables $X_1, \ldots, X_n$, $\hat{\mathbb{E}}g = \frac{1}{n}\sum_{i=1}^n g(x_i)$ and a class of distributions F:

$$f(X_1, \ldots, X_n) = \sup_{g \in F} \mathbb{E}g - \hat{\mathbb{E}}g$$

If every $g \in F$ satisfies $g : X \to [a, b]$ then f has bounded differences with $c_i = (b-a)/n$. So, with probability at least $1 - \delta$,

$$\sup_{g \in F}(\mathbb{E}g - \hat{\mathbb{E}}g) \leq \mathbb{E}[\sup_{g \in F}(\mathbb{E}g - \hat{\mathbb{E}}g)] + (b - a)\sqrt{\frac{\ln(1/\delta)}{2n}}$$

4 Loss Classes

Choose $\hat{f} \in F$ to minimize $\hat{L}$ where

$$
\begin{aligned}
\hat{L}(f) \quad & = \hat{\mathbb{E}}l(Y, f(X)) \\
& = \tfrac{1}{n}\textstyle\sum_{i=1}^n l(y_i, f(X_i))
\end{aligned}
$$

$$\hat{f} = \mathrm{argmin}_{f \in F} \hat{L}(f)$$

We want to know how does $L(\hat{f})$ behave?

Define the loss class

$$l_F = \{(x, y) \mapsto l(y, f(x))\}$$

and the excess loss class

$$\begin{aligned} g_F &= \{g : f \in F\} \\ &= \{(x, y) \mapsto l(y, f(x)) - l(y, f^*(x)) : f \in F\} \end{aligned}$$

where

$$f^* = \mathrm{argmin}_{f \in F} \mathbb{E} l(Y, f(X))$$

If $\sup_{g \in l_F} |\mathbb{E} g - \hat{\mathbb{E}} g| \leq \epsilon$ then

$$\begin{aligned} \mathbb{E} l_{\hat{f}} - \mathbb{E} l_{f^*} &= \underbrace{(\mathbb{E} l_{\hat{f}} - \hat{\mathbb{E}} l_{\hat{f}})}_{\leq \epsilon} + \underbrace{(\hat{\mathbb{E}} l_{\hat{f}} - \hat{\mathbb{E}} l_{f^*})}_{\leq 0} + \underbrace{(\hat{\mathbb{E}} l_{f^*} - \mathbb{E} l_{f^*})}_{\leq \epsilon} \\ &\leq 2\epsilon \end{aligned}$$

5 The Glivenko-Cantelli Theorem

Definition. F is a Glivenko-Cantelli class if for all $\epsilon > 0$,

$$\lim_{n \to \infty} \sup_P P^n \{\sup_{f \in F} |\mathbb{E} f - \hat{\mathbb{E}}_n f| > \epsilon\} = 0$$

Glivenko-Cantelli Theorem. For any P on $\mathbb{R}$ with distribution function F, i.e. $F(t) = P\left[(-\infty, t]\right]$, we have almost surely that

$$F_n \to F \text{ uniformly on } \mathbb{R} \text{ as } n \to \infty \ ,$$

where F_n is empirical distribution function

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[X_i \leq t] \ .$$