1 Concentration inequalities, continued

1.1 Hoeffding's inequality, continued

Example:

Let h be a classifier and $f(z) = I(y \neq h(x))$ where $z = (x, y)$. Then Hoeffding's inequality implies

$$|R(h) - \hat{R}(h)| \leq \sqrt{\frac{1}{n}\log\left(\frac{2}{\delta}\right)} \text{ with probability at least } 1-\delta \text{ (i.e. w.h.p.)}$$

However, we'd be more interested in $\hat{R}(\hat{h}) - R(\hat{h}) \leq \sup_{h\in\mathcal{H}} \hat{R}(h) - R(h)$ where $\hat{h}$ is trained on the training set. This is a *uniform* thing that doesn't work with Hoeffding's (*pointwise*).

1.2 McDiarmid's inequality

Hoeffding's inequality was a special case of **McDiarmid's inequality**:

Suppose that

$$\sup_{z_1,\ldots,z_n,z_i'} |f(z) - f(z_1,\ldots,z_{i-1},z_i',z_{i+1},\ldots,z_n)| \leq c_i$$

Then

$$\mathbb{P}\left(|f(Z) - \mathbb{E}f(Z)| > \epsilon\right) \leq 2e^{-2\epsilon^2/\sum_{i=1}^n c_i}$$

Note that if $f(z_1,\ldots,z_n) = \frac{1}{n}\sum_{i=1}^n z_i$, then we get back Hoeffding's inequality.

Example (total variation distance):

Let $f(Z) = \sup_A |\hat{\mathbb{P}}(A) - \mathbb{P}(A)|$.

$$\begin{aligned}
|f(Z) - f(Z')| &= \left|\sup_A |\hat{\mathbb{P}}(A) - \mathbb{P}(A)| - \sup_A |\hat{\mathbb{P}}'(A) - \mathbb{P}(A)|\right| \\
&\leq \sup_A \left||\hat{\mathbb{P}}(A) - \mathbb{P}(A)| - |\hat{\mathbb{P}}'(A) - \mathbb{P}(A)|\right| \\
&\leq \sup_A \left|\hat{\mathbb{P}}(A) - \mathbb{P}(A) - (\hat{\mathbb{P}}'(A) - \mathbb{P}(A))\right| \\
&= \sup_A \left|\hat{\mathbb{P}}(A) - \hat{\mathbb{P}}'(A)\right|
\end{aligned}$$

Changing one observation changes f by at most $1/n$. Therefore,

$$\mathbb{P}\left(|f(Z) - \mathbb{E}f(Z)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

1.3 Sharper inequalities

The previous results don't use information about *where* probability mass lies. Hoeffding's inequality does not use any information about the random variables except that they are bounded. In fact, it is driven by the worst case: a random variable that puts all of its mass at the boundaries. If the variance of Z_i is small (i.e. we have more mass in the middle), we can get a sharper inequality from **Bernstein's inequality**.

This idea is that $\sum_{i=1}^n Z_i$ is approximately normally distributed with variance $v = \sum_{i=1}^n \mathbb{V}Z_i$. The tails of a $N(0, v)$ are of order $e^{-x^2/(2v)}$. Bernstein's inequality gives a tail bound that is a combination of a normal and a *penalty* for non-normality.

Lemma: Suppose that $|X| < c$ and $\mathbb{E}X = 0$. Then for any $t > 0$

$$\mathbb{E}[e^{tX}] \leq \exp\left\{t^2\sigma^2\left(\frac{e^{tc} - 1 - tc}{(tc)^2}\right)\right\}$$

where $\sigma^2 = \mathbb{V}X$

Proof: Let $F = \sum_{r=2}^{\infty} \frac{t^{r-2}\mathbb{E}(X^r)}{r!\sigma^2}$. Then,

$$\mathbb{E}[e^{tX}] = \mathbb{E}\Big[1 + tx + \sum_{r=2}^{\infty} \frac{t^r X^r}{r!}\Big] = 1 + t^2\sigma^2 F \leq e^{t^2\sigma^2 F}.$$

For $r \geq 2$, $\mathbb{E}[X^r] = E[X^{r-2}X^2] \leq c^{r-2}\sigma^2$ (all higher moments than the variance are killed by the a.s. bound, while the first two moments are computed as usual), so

$$F \leq \sum_{r=2}^{\infty} \frac{t^{r-2}c^{r-2}\sigma^2}{r!\sigma^2} = \frac{1}{(tc)^2}\sum_{r=2}^{\infty}\frac{(tc)^r}{r!} = \frac{e^{tc} - 1 - tc}{(tc)^2}.$$

Hence, $\mathbb{E}[e^{tX}] \leq \exp\left\{t^2\sigma^2\left(\frac{e^{tc}-1-tc}{(tc)^2}\right)\right\}$.

This Lemma can be used to show...

Bernstein's inequality: If $|Z_i| \leq c$ a.s. and $\mathbb{E}Z_i = \mu$, then for all $\epsilon > 0$,

$$\mathbb{P}\left(|\overline{Z} - \mu| > \epsilon\right) \leq 2\exp\left\{-\frac{n\epsilon^2}{2\sigma^2 + 2c\epsilon/3}\right\} \text{ where } \sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{V}Z_i.$$

Compare this to **Hoeffding's inequality:**

$$\mathbb{P}\left(|\overline{Z} - \mu| > \epsilon\right) \leq 2\exp\left\{-\frac{n\epsilon^2}{2c^2}\right\}$$

- If $\sigma^2 >> 2c\epsilon/3$, then $\log(\text{Bernstein}) \asymp -\frac{n\epsilon^2}{2\sigma^2} \leq -\frac{n\epsilon^2}{4c^2} \asymp \log(\text{Hoeffding})$
- If $\sigma^2 << 2c\epsilon/3$, then $\log(\text{Bernstein}) \asymp -\frac{3n\epsilon^2}{4c\epsilon} = -\frac{3n\epsilon}{4c} \leq -\frac{n\epsilon^2}{4c^2} \asymp \log(\text{Hoeffding})$

This implies that the Bernstein bound is like an exponential for large ϵ and normal for small ϵ.

Proof of Bernstein's inequality:

For simplicity, assume that $\mu = 0$. From the Lemma, we know

$$\mathbb{E}[e^{tX}] \leq \exp\left\{t^2\sigma^2\left(\frac{e^{tc}-1-tc}{(tc)^2}\right)\right\} \text{ where } \sigma_i^2 = \mathbb{E}[X_i^2]$$

So then

$$\begin{aligned}
\mathbb{P}(\bar{X}_n > \epsilon) &= \mathbb{P}(\sum_{i=1}^n X_i > n\epsilon) = \mathbb{P}(e^{t\sum_{i=1}^n X_i} > e^{tn\epsilon}) \\
&= e^{-tn\epsilon}\mathbb{E}[e^{t\sum_{i=1}^n X_i}] = e^{-tn\epsilon}\prod_{i=1}^n \mathbb{E}[e^{tX_i}] \\
&\leq e^{-tn\epsilon} exp\left\{nt^2\sigma^2\left(\frac{e^{tc}-1-tc}{(tc)^2}\right)\right\}
\end{aligned}$$

Set $t = (1/c)\log(1+\epsilon c/\sigma^2)$ to obtain

$$\mathbb{P}(\bar{X}_n > \epsilon) \leq \exp\left\{-\frac{n\sigma^2}{c^2}h\Big(\frac{c\epsilon}{\sigma^2}\Big)\right\}$$

where $h(u) = (1+u)\log(1+u) - u$. Bernstein's inequality then follows from $h(u) \geq u^2/(2+2u/3)$ for $u \geq 0$.

The most useful part of Bernstein's inequality is the associated **PAC bound**:

With probability at least $1-\delta$

$$|\overline{Z} - \mu| \leq \sqrt{\frac{2\sigma^2\log(1/\delta)}{n}} + \frac{2c\log(1/\delta)}{3n}$$

since $|Z_i| \leq c$, $\sigma^2 = n^{-1}\sum_{i=1}^n \mathbb{V}Z_i$.

In particular, if the variance is small enough,

$$\sigma^2 \leq \frac{2c^2\log(1/\delta)}{9n} \quad \Rightarrow \quad |\overline{Z} - \mu| \leq \frac{4c\log(1/\delta)}{3n}$$

That is, we get a n-decay instead of $\sqrt{n}$-decay, which is a big win!

References

1. http://www.stat.cmu.edu/∼larry/=sml/Concentration.pdf