Statistical Machine Learning | 19 — Concentration of Measure | 11/04/14

Lecturer: Prof. Homrighausen | Scribe: Ryan Haunfelder

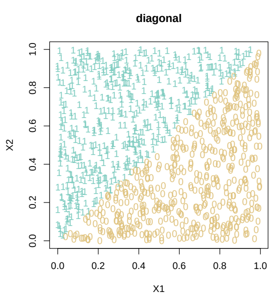
1 Henry Scharf-Subsampling Aggregation:Sagging(or Surging)

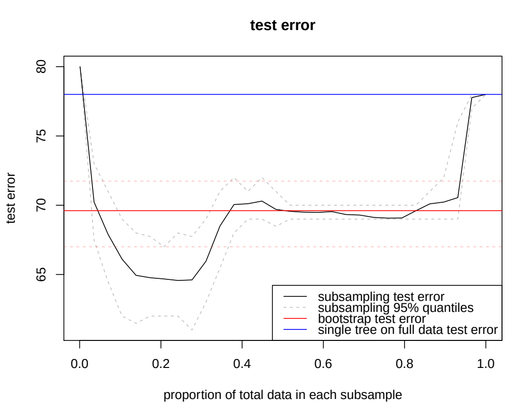
Sagging, or bagging with a bootstrap subsample with $b < n$ was investigated by Henry. He shared a result on consistency and compared this method to traditional bagging through simulation. Both methods were tested on logit and diagonal models.

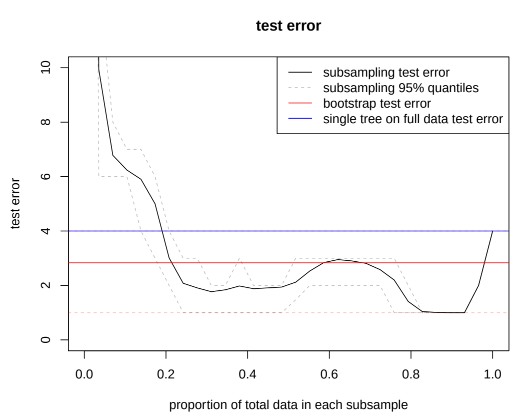
Logit Model

$$
\begin{aligned}
Y_i &\ Bernoulli(p_i) \\
logit(p) &= X^T\beta \\
(X_1, X_2, X_3, X_4) &\sim N(\mu, \mathbf{\Sigma_x}) \\
X_5 &\sim Poisson(\lambda_x) \\
X_6 &\sim Uniform(0, 1)
\end{aligned}
$$

Diagonal Model







2 Concentration of Measure: Introduction

The core of modern machine learning theory rests with the following structure:

1. Concentration inequalities: Show that a random quantity is close to its mean with high probability
   (a) Hoeffding's- If X_i are almost surely bounded, that is $P(X_i \in [a_i, b_i]) = 1$, then
   $$P\left(\bar{X} - E(\bar{x})\right) \leq exp\left(-\frac{2n^2t^2}{\sum\limits_{i=1}^{n}(b_i - a_i)^2}\right)$$
   (b) Bernstein's- Three inequalities similar in flavor to,
   $$P\left(\sum_{i=1}^{n} X_i > t\right) \leq exp\left(-\frac{\frac{1}{2}t^2}{\sum E(X_j^2) + \frac{1}{3}Mt}\right)$$
   (c) McDiarmid's- For any function f that when any argument is changed has a value that changes less than c, i.e.,
   $$\sup_{x_1, x_2, \ldots, x_n, \hat{x}_i} \|f(x_1, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_{i-1}, \hat{x}_i, \ldots, x_n)\| \leq c_i$$
   the inequality,
   $$P(f(X_1, X_2, \ldots, X_n) - E(f(X_1, X_2, \ldots, X_n)) \geq \epsilon) \leq exp\left(-\frac{2\epsilon^2}{\sum\limits_{i=1}^{n} c_i^2}\right)$$
2. Uniform bounds: Guarantee that a set of random quantities are all simultaneously close to their means with high probability. These will relate the the test error and training error.
   (a) VC-dimension- The number of points that an algorithm can shatter.
   (b) Rademacher complexity-Richness of a class of functions.
   (c) Covering/bracketing numbers-The number of spherical balls needed to cover a space.

Goal: (concentration inequalities) + (complexity measure) = uniform coverage of a stochastic process (i.e. $\sup_{t \in T} X_t$).

2.1 Motivation

Suppose we have data $\mathcal{D}$ and a loss function ℓ_f and we wish to find a function $\hat{f}$ that can predict a new Y from an X

Form the excess risk, or difference between the true risk of the estimator that minimizes the emperical risk and the that the minimizes the true risk,

$$\mathcal{E}(\hat{f}) = \mathbb{P}\ell_{\hat{f}} - \inf_{f \in \mathcal{F}} \mathbb{P}\ell_f$$

and $\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{\mathbb{P}}\ell_f$

$\hat{\mathbb{P}} = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$ is the empirical measure.

This can be interpreted in two ways:

- Expectation: Let f be a function, then we write

$$\hat{\mathbb{P}} f = \int f \, d\hat{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

- Measure: Let C be a (measurable) set, then we write

$$\hat{\mathbb{P}} C = \int \mathbf{1}_C d\hat{\mathbb{P}} = \frac{1}{n} |\{i : X_i \in C\}|$$

These notions are used interchangeably, and motivate using $\mathbb{P}$ for both probability and expectation

Apply the $2 - \epsilon$ technique:

$$\begin{aligned}
\mathcal{E}(\hat{f}) &= \mathbb{P}\ell_{\hat{f}} - \hat{\mathbb{P}}\ell_{\hat{f}} + \hat{\mathbb{P}}\ell_{\hat{f}} - \inf_{f \in \mathcal{F}} \mathbb{P}\ell_f \\
&\leq \mathbb{P}\ell_{\hat{f}} - \hat{\mathbb{P}}\ell_{\hat{f}} + \hat{\mathbb{P}}\ell_{f_*} - \mathbb{P}\ell_{f_*} \\
&\leq 2 \sup_{f \in \mathcal{F}} |\mathbb{P}\ell_f - \hat{\mathbb{P}}\ell_f|
\end{aligned}$$

Where f_* is such that $\mathbb{P}\ell_{f_*} = \inf_{f \in \mathcal{F}} \mathbb{P}\ell_f$

We can then use a concentration inequality to bound the excess risk. So, fixing an $\epsilon > 0$

$$\begin{aligned}
\mathbb{P}(\mathcal{E}(\hat{f}) > 2\epsilon) &\leq \mathbb{P}(| \sup_{f \in \mathcal{F}} |(\hat{\mathbb{P}} - \mathbb{P})\ell_f| > \epsilon) \\
&\leq \frac{\mathbb{E} \sup_{f \in \mathcal{F}} |(\hat{\mathbb{P}} - \mathbb{P})\ell_f|}{\epsilon}
\end{aligned}$$

Conclusion:

$$\mathbb{P}(\mathcal{E}(\hat{f}) > 2\epsilon) \leq \epsilon^{-1} \mathbb{E} \sup_{f \in \mathcal{F}} |(\hat{\mathbb{P}} - \mathbb{P})\ell_f| = \epsilon^{-1} \mathbb{E} \|\hat{\mathbb{P}} - \mathbb{P}\|_{\mathcal{F}}$$

We can bound the excess risk of an estimator $\hat{f}$ by bounding the supremum of the difference between the empirical measure and true measure

Note that:

- Using the previous notation, $X_t = (\hat{\mathbb{P}} - \mathbb{P})\ell_f$, and $T = \mathcal{F}$
  Sometimes the index set is considered $\mathcal{L} = \{\ell_f : f \in \mathcal{F}\}$
- The stochastic process $\mathbb{G} = \sqrt{n}(\hat{\mathbb{P}} - \mathbb{P})$ is the empirical process

The stochastic process $(\hat{\mathbb{P}} - \mathbb{P})\ell_f$ is zero mean and hence we know by the SLLN that for all $f \in \mathcal{F}$

$$(\hat{\mathbb{P}} - \mathbb{P})\ell_f \to 0 \text{ a.s}$$

assuming $\mathbb{P}\ell_f$ exists.

However, this does not give us uniform control because this does not imply that the supremum goes to zero.

Definition 2.1. *We call an index set $\mathcal{F}$ a <u>Glivenko-Cantelli</u> class if*

$$\sup_{f \in \mathcal{F}} |(\hat{\mathbb{P}} - \mathbb{P})\ell_f| = \|\hat{\mathbb{P}} - \mathbb{P}\|_{\mathcal{F}} \to 0 \ \textit{a.s}$$

A classical example is the empirical CDF

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,t]}(X_i) = \hat{\mathbb{P}} f_t$$

where $f_t(x) = \mathbf{1}_{(-\infty,t]}(x)$

Often, we are attempting to estimate a functional of the true CDF with a plug-in version using the empirical CDF

True CDF: $F(t) = \mathbb{P}(X \le t)$

Example 2.2. *Let $\theta = \theta(\mathbb{P})$ given by the median: $\theta = \theta(\mathbb{P})$ is argmin of $\mathbb{P}(-\infty, x] = \inf_x F(x)$ subject to $F(x) \ge 1/2$*

Then, we might estimate $\theta(\mathbb{P})$ with $\hat{\theta} = \theta(\hat{\mathbb{P}})$ by plugging in F_n The Glivenko-Cantelli theorem says that

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \to 0 \ a.s.$$

If we write $\mathcal{F} = \{f_t : f_t(x) = \mathbf{1}_{(-\infty,t]}(x), t \in \mathbb{R}\}$, then

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| = \|F_n - F\|_{\mathbb{R}} = \|\hat{\mathbb{P}} - \mathbb{P}\|_{\mathcal{F}}$$

and hence $\mathcal{F}$ is a Glivenko-Cantelli (G.C.) class.
Technical condition: *$\mathbb{P}(-\infty, t] > 1/2$ for each $t > \theta(\mathbb{P})$. This forces a continuity property that |median($\mathbb{P}$) - median($\mathbb{P}$')| $< \epsilon$ if $\mathbb{P}$ and $\mathbb{P}$' are uniformly close. This means we must be able to separate the median from it's neighbor.*

Example 2.3. *As $\mathcal{F}$ is G.C., for all $\delta > 0$, for n large enough*

$$\sup_t |\hat{\mathbb{P}}(-\infty, t] - \mathbb{P}(-\infty, t]| < \delta$$

Fix $\epsilon > 0$. Choose δ such that

$$\begin{aligned}
\mathbb{P}(-\infty, \theta - \epsilon] &< \frac{1}{2} - \delta \quad (\textit{This is always possible}) \\
\mathbb{P}(-\infty, \theta + \epsilon] &> \frac{1}{2} + \delta \quad (\textit{This requires condition})
\end{aligned}$$

Now,

$$\mathbb{P}(-\infty, \hat{\theta}] > \underbrace{\hat{\mathbb{P}}(-\infty, \hat{\theta}] - \delta}_{\textit{uniform closeness}} \ge 1/2 - \delta$$

Hence, $\hat{\theta} > \theta - \epsilon$ as BWOC:

$$\hat{\theta} \le \theta - \epsilon \Rightarrow \mathbb{P}(-\infty, \hat{\theta}] \le \mathbb{P}(-\infty, \theta - \epsilon] < 1/2 - \delta$$

Also, it can be shown that $\hat{\theta} \le \theta + \epsilon$.

Hence, uniform closeness of F_n to F shows that the sample and populations medians are close.

Note, we have asked for much more than needed, sometimes this can be too much This gets refined to a rate of convergence by the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality

$$\mathbb{P}(\|F_n - F\|_\infty > \epsilon) \le 2e^{-2n\epsilon^2}$$

Both the constants 2 cannot be improved upon. [2]

This result, along with the previous discussion gets us a rate of convergence for the median.

Definition 2.4. *Let ψ be a non-decreasing, convex function such that $\psi(0) = 0$. Then, define the norm,*

$$\|X\|_\psi = \inf\left\{C > 0 : \mathbb{E}\psi\left(\frac{|X|}{C}\right) \leq 1\right\}$$

$\|.\|_\psi$ is a norm.

Proof. First, for any constant a,

$$\begin{aligned}\|aX\|_\psi &= \inf\left\{C > 0 : \mathbb{E}\psi\left(\frac{|aX|}{C}\right) \leq 1\right\} \\ &= \inf\left\{C > 0 : \mathbb{E}\psi\left(\frac{|X|}{C/|a|}\right) \leq 1\right\} \\ &= |a|\|X\|_\psi\end{aligned}$$

Let $\|X\|_\psi = x$ and $\|X\|_\psi = y$.

Then,

$$\begin{aligned}\mathbb{E}\psi\left(\frac{|X+Y|}{x+y}\right) &\leq \mathbb{E}\left(\psi\left(\frac{|X|}{x}\right) + \psi\left(\frac{|Y|}{y}\right)\right) && \text{(Since } \psi \text{ is non-decreasing)} \\ &= \mathbb{E}\left(\psi\left(\frac{x}{x+y}\frac{|X|}{x} + \frac{y}{x+y}\frac{|Y|}{y}\right)\right) \\ &\leq \frac{x}{x+y}\mathbb{E}\left(\psi\left(\frac{|X|}{x}\right)\right) + \frac{y}{x+y}\mathbb{E}\left(\psi\left(\frac{|Y|}{y}\right)\right) && \text{(By convexity)} \\ &\leq 1\end{aligned}$$

Therefore,

$$\|X+Y\|_\psi \leq \|X\|_\psi + \|Y\|_\psi$$

and hence it is a norm. ■

There are two main cases

- L_p norm: $\psi(x) = x^p \Rightarrow \|X\|_\psi = \|X\|_p = (\mathbb{E}|X|^p)^{1/p}$
- p-Orlicz: $\psi_p(x) = e^{x^p} - 1$

Two important facts:

- $\|X\|_{\psi_p} \leq \|X\|_{\psi_q}(\log 2)^{1/q-1/p}$, for $p \leq 2$
- $\|X\|_p \leq p!\|X\|_{\psi_1}$

This allows us to interchange results about various norms, as long as we don't care about constants By Markov's inequality

$$\begin{aligned}\mathbb{P}(|X| > x) &\leq \frac{\mathbb{E}\psi(|X|)/\|X\|_\psi}{\psi(x)/\|X\|_\psi} \\ &\leq \frac{1}{\psi(x)/\|X\|_\psi} \\ &= \begin{cases}\|X\|_p x^{-p} & \text{if } \psi(x) = x^p \\ \frac{1}{e^{(x/\|X\|_\psi)^p}-1} \asymp e^{-(x/\|X\|_\psi)^p} & \text{if } \psi(x) = \psi_p(x)\end{cases}\end{aligned}$$

Hence, Orlicz norms allow us to encode the tail behavior of a random variable.

In fact, it works as an if and only if:

If $\mathbb{P}(|X| > x) \leq Ce^{-cx^p}$ then $\|X\|_{\psi_p} \leq ((1+C)c^{-1})^{1/p} < \infty$

3 Concentration of Measures

For showing results about empirical processes or performance guarantees for algorithms, we want results of the form

$$\mathbb{P}\left(|f(Z_1, \ldots, Z_n) - \mu_n(f)| > \epsilon\right) < \delta_n$$

where $\delta_n \to 0$ and $\mu_n(f) = \mathbb{E}f(Z_1, \ldots, Z_n)$.

For statistical learning theory, we need uniform bounds

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |f(Z_1, \ldots, Z_n) - \mu_n(f)| > \epsilon\right) < \delta_n$$

Suppose $\mu = \mathbb{E}Z < \infty$ and $\mathbb{P}(Z \geq 0) = 1$. Then for any $\epsilon > 0$

$$\mathbb{E}Z = \int_0^\infty Z d\mathbb{P} \geq \int_\epsilon^\infty Z d\mathbb{P} \geq \epsilon \int_\epsilon^\infty d\mathbb{P} = \epsilon \mathbb{P}(Z > \epsilon)$$

Yielding Markov's inequality

This can be transformed to Chebyshev's inequality by using the variance

$$\mathbb{P}(|Z - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2} \Rightarrow \mathbb{P}(|\overline{Z} - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

Observation: This is nice, but does not decay exponentially fast. However, it only makes a second moment assumption A different transformation occurs via a Chernoff bound. For any $t > 0$

$$\mathbb{P}(Z > \epsilon) = \mathbb{P}\left(e^{tZ} > e^{t\epsilon}\right) \leq e^{-t\epsilon}\mathbb{E}[e^{tZ}]$$

This is the moment generating function. Here we see increasing moment conditions giving tighter bounds

This can be minimized over t as it is arbitrary

$$\mathbb{P}(Z > \epsilon) \leq \inf_{t>0} e^{-t\epsilon}\mathbb{E}[e^{tZ}]$$

This is main content of the paper by Hoeffding [1]: Suppose $Z \in [a, b]$, then for any t

$$\mathbb{E}[e^{tZ}] \leq e^{t\mu + t^2(b-a)^2/8}$$

Using this, we find Hoeffding's inequality

$$\mathbb{P}\left(|\overline{Z} - \mu| > \epsilon\right) \leq 2e^{-2n\epsilon^2/(b-a)^2}$$

Proof sketch: Let Z be zero mean

$$\begin{aligned}
\mathbb{P}\left(\overline{Z} > \epsilon\right) &= \mathbb{P}\left(e^{t\overline{Z}} > e^{t\epsilon}\right) \\
&\leq e^{-t\epsilon}\mathbb{E}e^{t\overline{Z}} \\
&= e^{-t\epsilon}\prod_{i=1}^{n}\mathbb{E}e^{tn^{-1}Z_i} \\
&\leq e^{-t\epsilon}e^{(t/n)^2(b-a)^2/8} \quad \text{(Now, minimize over } t \text{ and symmetrize)}
\end{aligned}$$

We can let the upper and lower limits change with i: $Z_i \in [a_i, b_i]$

Also, we can invert this probability statement into a PAC bound: with probability at least $1 - \delta$

$$|\overline{Z} - \mu| \leq \sqrt{\frac{c}{2n} \log\left(\frac{2}{\delta}\right)}$$

where $c = n^{-1} \sum_i (b_i - a_i)^2$

Compare to Chebyshev, which has growth

$$|\overline{Z} - \mu| \leq \sqrt{\frac{\sigma^2}{n\delta}}$$

References

[1] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30.

[2] Massart, P. (1990). The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality.