The Bayes Predictor

$l(y,z) = \alpha(y-z)_+ + (1-\alpha)(z-y)_+$ for $\alpha \in (0,1)$

$(z)_+ = \max(0,z)$

	$y \geq z$	$y \leq z$
$\alpha(y-z)$	+ve	~~+ve~~ −ve
$(1-\alpha)(z-y)$	−ve	+ve
Result	$\alpha(y-z)$	$(1-\alpha)(z-y)$

$$\therefore l(y,z) = \begin{cases} \alpha(y-z), & y \geq z \\ (1-\alpha)(z-y), & y < z \end{cases}$$

For Bayes predictor we minimize the conditional expected loss

$$f^*(x) = \underset{z}{\operatorname{argmin}} \; E_{y|X=x}[l(Y,z)]$$

$$E[l(Y,z)|X=x] = \alpha E[(y-z)_+ |X=x] + (1-\alpha)E[(z-y)_+ |X=x]$$

$$= \alpha \int_z^{\infty} (y-z) P_{Y|X}(y|x)\,dy + (1-\alpha)\int_{-\infty}^{z} (z-y) P_{Y|X}(y|x)$$

To minimize this

$$\frac{d}{dz} E[l(Y,z)|X=x] = 0$$

$$\frac{d}{dz} I(z) = \int_{a(z)}^{b(z)} f(y,z)\,dy$$

$$= f(b(z),z)\,b'(z) - f(a(z),z)\,a'(z) + \int_{a(z)}^{b(z)} \frac{\partial}{\partial z} f(y,z)\,dy$$

For term

$$I = \alpha \int_{z}^{\infty} (y-z) P_{Y/X}(y|x)\,dy$$

$$a(z) = z \Rightarrow a'(z) = 1$$

$$b(z) = \infty \Rightarrow b'(z) = 0$$

$$\frac{d}{dz} I(z) = \alpha\Big[\underbrace{\cancel{(y-z) P_{Y/X}(y/x)}}_{\text{Vanishes } b'(z)=0}.0 - \underbrace{\cancel{(y-z) P_{Y/X}}}_{\text{Vanishes at } y=z}(y/x).1 + \int_{z}^{\infty} \frac{\partial}{\partial z}(y-z) P_{Y/X}(y|x)\Big]$$

$$\frac{d}{dz} I(z) = -\alpha P(Y \leq z \,|\, X = x)$$

$$= -\alpha\left(1 - F_{Y/X}(z/x)\right)$$

where $F_{Y/X}(z/x) = P(Y \leq z | X = x)$ is the CDF

For Term 2

$$I = (1-\alpha) \int_{-\infty}^{z} (z-y) P_{Y/X}(y|x)\,dy$$

$$a(z) = -\infty \Rightarrow a'(z) = 0$$

$$b(z) = z \Rightarrow b'(z) = 1$$

$$\frac{d}{dz} I = (1-\alpha)\Big[\underbrace{\cancel{(z-y) P_{Y/X}}}_{\text{Vanishes at } z=y}(y|x)dy.1 + \underbrace{\cancel{(z-y) P_{Y/X}}}_{\text{vanishes as } a'(z)=0}(y|x)dy.0 + \int_{z}^{\infty} \frac{\partial}{\partial z}(z-y) P_{Y/X}(y|x)\,dy\Big]$$

$$= (1-\alpha) P_{Y/X}(y|x)$$

$$= (1-\alpha) F_{Y/X}(z,x)$$

Combining both we get

$$-\alpha\left(1-F_{Y/X}(z/x)\right)+(1-\alpha)\,F_{Y/X}(z,x)$$

$$-\alpha+\alpha F+F-\alpha F$$

$$-\alpha+F=0$$

$$F=\alpha$$

$$F_{Y/X}(z/x)=\alpha$$

lets pick $\alpha = 0.7$

We compute $L(z)=\alpha\int_z^1 (y-z)\,dy+(1-\alpha)\int_0^z (z-y)\,dy$

Firsts

$$\int_z^1 (y-z)\,dy = \int_z^1 y\,dy - z(1-z)$$

$$=\left[\frac{y^2}{2}\right]_z^1$$

$$=\left(\frac{1}{2}-\frac{z^2}{2}\right)$$

$$= z(1-z)$$

Second integral $\int_0^z (z-y)\,dy = z^2-\frac{z^2}{2}$

$$=\frac{z^2}{2}$$

$$\therefore\ L(z)=\alpha\left(\frac{1}{2}-\frac{z^2}{2}-z(1-z)\right)+(1-\alpha)\frac{z^2}{2}$$

Plug in $\alpha=0.7$, plot or test several z values and you'll find the loss is minimized at $z=0.7$

Confirming $f^*=0.7$

Concentration and Inequalities I

Problem A

Let $z_1, z_2 \dots z_n$ be independent random variables such that $z_i \in [a_i, b_i]$

$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} z_i \qquad \mu = E[\hat{\mu}]$

$= \frac{1}{n}\sum_{i=1}^{n} E[z_i]$

then for any $\varepsilon > 0$

$$P(|\hat{\mu} - \mu| \geq \varepsilon) \leq 2\, e^{\frac{-2n^2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

If $z_i \in [0, 1]$

the $P(|\hat{\mu} - \mu| \geq \varepsilon) \leq 2\, e^{-2n\varepsilon^2}$

Markov's inequality: for a random variable x and $t > 0$

$$P(x \geq \varepsilon) \leq \frac{E[e^{tx}]}{e^{t\varepsilon}}$$

Apply to sum $\sum z_i$ using exponential moments

$$P(\hat{\mu} - \mu \geq \varepsilon) = P\left(\sum_{i=1}^{n} z_i - E[z_i] \geq n\varepsilon\right)$$

Use chernoff's method

$$P\left(\sum (z_i - E[z_i]) \geq n\varepsilon\right) \leq \inf_{t>0} \exp(-tn\varepsilon) . E\left[\exp\left(t \sum (z_i - E[z_i])\right)\right]$$

$$E\left[\exp\left(t \sum_{i=1}^{n} z_i - E[z_i]\right)\right] = \prod_{i=1}^{n} E\left[\exp\left(t(z_i - E[z_i])\right)\right]$$

By Hoeffding's lemma

$$E\left[e^{t(x - E(x))}\right] \leq \exp\left(\frac{t^2 (b-a)^2}{8}\right)$$

here $z_i \in [0, 1] \Rightarrow a = 0 \quad b = 1$

$$\therefore E\left[\exp\left(t(z_i - E[z_i])\right)\right] \leq \exp\left(\frac{t^2}{8}\right)$$

$$E[e^{tS_n}] \leq \prod_{i=1}^{n} e^{t^2/8}$$

$$= e^{nt^2/8}$$

$$\therefore P(S_n \geq n\varepsilon) \leq \frac{E[e^{(tS_n)}]}{e^{tn\varepsilon}}$$

$$\leq \frac{e^{nt^2/8}}{e^{tn\varepsilon}}$$

$$\leq \exp\left(\frac{nt^2}{8} - tn\varepsilon\right)$$

we minimize $\frac{nt^2}{8} - tn\varepsilon$

$$\frac{d}{dt}\left(\frac{nt^2}{8} - tn\varepsilon\right) = 0$$

$$\frac{2nt}{8} - n\varepsilon = 0$$

$$\frac{\cancel{n}t}{4} = \cancel{n}\varepsilon$$

$$t = 4\varepsilon$$

$$P(S_n > n\varepsilon) \leq \exp\left(\frac{n16\varepsilon^2}{8} - 4\varepsilon^2 n\right)$$

$$\leq \exp(2n\varepsilon^2 - 4n\varepsilon^2)$$

$$\leq -2n\varepsilon^2$$

$$\therefore P(|\hat{p} - p| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}$$

Given $n = 500$, $\varepsilon = 0.05$

Plugging in values

$$P(|\hat{p} - p| > \varepsilon) \leq 2e^{-2\times 500 \times 0.05^2}$$

$$\leq 0.1642$$

$$P(|\hat{P} - Y| \geq \varepsilon) \leq 2\exp(2n\varepsilon^2)$$

$$2\exp(2n\varepsilon^2) \leq \delta$$

$$\exp 2n\varepsilon^2 \leq \frac{\delta}{2}$$

$$\log \exp 2\varepsilon^2 n \leq \log \frac{\delta}{2}$$

$$-2\varepsilon^2 n \leq \log \delta - \log 2$$

$$-n \leq \frac{\log \delta - \log 2}{2\varepsilon^2}$$

$$n \leq \frac{\log 2 - \log \delta}{2\varepsilon^2}$$

In most practical situations Hoeffding's inequality is not tight. Hoeffding's inequality is a worst case bound which uses only range of random variables and makes no assumptions about the distribution. Hence the bound is always safe and guaranteed but loose. Inequalities like Bernstein or Chernoff's are more likely to provide tighter bound.

Given $f(Z_1, Z_2 \dots Z_n) = \max_i Z_i$ where $Z_i \in [0,1]$

Hoeffding's inequality is specifically designed to bound the deviation of the sample mean of independent bounded random variables

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

It relies of independence of Z_i and f is sum or average. Hence Hoeffding's inequality cannot be applied in this given setup.

Problem B

Let $z_1, z_2 \cdots z_n$ be independent random variables with
$z_i \in [0, 1]$
Var (z_i)

Sample mean $\hat{p} = \frac{1}{n} \sum_{i=1}^{n} z_i$
True mean $p = E[\hat{p}]$

Then for any $\varepsilon > 0$ Bernstein's inequality states

$$P(|\hat{p} - p| \geq \varepsilon) \leq 2 \exp\left[\frac{-n\varepsilon^2}{2\sigma^2 + \frac{2}{3}\varepsilon}\right]$$

Comparison with Hoeffding

Assumptions : Hoeffding inequality guarantees only boundedness
Bernstein's inequality guarantees boundedness and bounded variance

Sharpness : The Hoeffding inequality assumes only that the variables are bounded not how they are distributed within those bounds. But Bernstein's inequality takes variance into account.

When variance $\sigma^2 \ll 1$ the denominator becomes much smaller so bound is sharper.

Given $n = 500$ $\sigma^2 = 0.04$ $\varepsilon = 0.05$

$$P(|\hat{p} - p| \geq \varepsilon) \leq 2 \exp\left(\frac{-500 \times 0.05^2}{2 \times 0.04 + \frac{2}{3} 0.05}\right)$$

$$\leq 3.25 \times 10^{-5}$$

Compared to the bound in problem A this is tighter

Given $z_i \in [0,1]$ independent but not necessarily identically distributed

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} z_i \qquad \mu = E[\hat{\mu}]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[z_i]$$

$$\sigma_i^2 = Var(z_i)$$

$$V = Var(\hat{\mu})$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2$$

$$P(|S_n| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2 + \frac{2}{3}n\varepsilon}\right)$$

where $S_n = \sum_{i=1}^{n} X_i$ $\quad \hat{\mu} - \mu = \frac{1}{n} S_n$ $\quad t = n\varepsilon$

$$\cancel{P(|S_n|} \quad P(|\hat{\mu} - \mu| \geq \varepsilon) \leq 2\exp\left(\frac{-n^2\varepsilon^2}{2\sum_{i=1}^{n}\sigma_i^2 + \frac{2}{3}n\varepsilon}\right)$$

$$\sum_{i=1}^{n}\sigma_i^2 = n^2 V$$

$$\therefore P(|\hat{\mu} - \mu| \geq \varepsilon) \leq 2\exp\left(\frac{-n^2\varepsilon^2}{2n^2V + \frac{2}{3}n\varepsilon}\right)$$

$$P(|\hat{\mu} - \mu| \geq \varepsilon) \leq 2\exp\left(\frac{-\varepsilon^2}{2V + \frac{2\varepsilon}{3n}}\right)$$

Concentration and Inequalities II

Problem A

$j \in \{1, 2 \ldots d\}$

$\hat{P}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$

By Hoeffding inequality

$$P(|\hat{P}_j - P_j| \geq \varepsilon/\sqrt{d}) \leq 2\exp\left(\frac{-n\varepsilon^2}{2d}\right)$$

Applying Union bound over all d coordinates

$$P(|\hat{P} - P|_\infty \geq \varepsilon/\sqrt{d}) \leq 2d\exp\left(\frac{-n\varepsilon^2}{2d}\right)$$

Since $|\hat{P} - P|_2 \leq \sqrt{d}\,|\hat{P} - P|_\infty$

$$P(|\hat{P} - P|_2 \geq \varepsilon) \leq 2d\exp\left(\frac{-n\varepsilon^2}{2d}\right)$$

Dependence on d

The bound deteriorates as d increases because of the $\sqrt{d}$ term in the norm conversion and the d factor in the union bound. For higher dimensional settings this bound becomes loose and require alternative methods like matrix concentration inequalities.

Problem B

For $z_i \in [0, 1]$ Hoeffding gives

$$P(|\hat{\mu} - \mu| \geq \varepsilon) \leq 2\exp(-2n\varepsilon^2)$$

$$2e^{-2n\varepsilon^2} \leq 0.05$$

$$e^{-2n\varepsilon^2} \leq 0.025$$

$$-2n\varepsilon^2 \leq \ln 0.025$$

$$\cancel{2n\varepsilon^2 \geq} -n \leq \frac{1}{2\varepsilon^2} \ln 0.025$$

$$-n \leq \frac{\ln 0.025}{2 \times 0.0025}$$

$$-n \leq \frac{\ln 0.025}{0.005} \Rightarrow n \geq \frac{\ln(1/0.025)}{0.005}$$

$$\approx 737.78$$

$$\approx 738 \text{ samples}$$

For $\sigma^2 \leq 0.25$ Bernstein gives

$$P(|\hat{\mu} - \mu| \geq \varepsilon) \leq 2\exp\left(\frac{-n\varepsilon^2}{2\sigma^2 + \frac{2}{3}\varepsilon}\right) \leq 0.05$$

$\varepsilon = 0.05 \quad \sigma^2 = 0.25$

$$\exp\left(\frac{-n \times 0.0025}{2 \times 0.25 + \frac{2}{3} \times 0.05}\right) \leq 0.025$$

$$n \geq \frac{0.5333 \ln(1/0.025)}{0.0025}$$

$$\approx 787.4$$

$$\approx 788 \text{ samples}$$

Higher the confidence lower will be the tail probability and need more samples. Smaller error tolerance implies tighter estimation and need more samples.

Hoeffding variance is safe but loose and possible conservative. Bernstein adapts to variance. Bernstein is better when variance is low.

Bounding generalization error for Ridge regression

$$\mathcal{F} = \{ f_\theta(x) = \varphi(x)^T\theta : \theta \in \mathbb{R}^d, |\theta|_2 \le D \}$$

In lectures the generalization error bound was derived using Rademacher complexity for hypothesis class $\mathcal{F}$ without regularization ($\lambda = 0$)

$$R_n(\mathcal{F}) \le \frac{DR}{\sqrt{n}}$$

where $\|\varphi(x_i)\|_2 \le R$ $\|\theta\|_2 \le D$ The generalization error bound with probability at least $1-\delta$ is

$$\text{Generalization error} \le O\left(\frac{DR}{\sqrt{n}} + B\sqrt{\frac{\log(1/\delta)}{n}} \right)$$

$$\therefore \hat{\theta} = \underset{\|\theta\|_2 \le D}{\arg\min} \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \varphi(x_i)^T\theta)^2 + \lambda\|\theta\|_2^2 \right]$$

To incorporate the regularization term we can consider the effective norm of θ under regularization. The regularization term penalizes large θ, so the effective hypothesis class becomes smaller. The Rademacher complexity for this regularized class can be bounded by

$$R_n(\mathcal{F}) \le \frac{D_{eff} R}{\sqrt{n}}$$

where D_{eff} is the effective norm of θ under regularization. For ridge regression, the solution satisfies $\|\theta\|_2 \leq \min(D, B/\lambda)$ since the regularization term dominates when λ is large. Thus we can approximate $D_{eff} = \min(D, B/\lambda)$

The modified generalization error bound becomes:

$$\text{Generalization error} \leq O\left(\frac{\min(D, B/\lambda)R}{\sqrt{n}} + B\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

As $\lambda \to 0$

The regularization term vanishes and $D_{eff} \to D$
The bound reduces to the original bound without regularization:

$$\text{Generalization error} \leq O\left(\frac{DR}{\sqrt{n}} + B\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

This is the case where model is less constrained

As $\lambda \to \infty$

The regularization term dominates forcing $\theta \to 0$
The effective norm $D_{eff} \to 0$

$$\text{Generalization error} \leq O\left(B\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

Here the model is overly constrained leading to underfitting.

Given ~~$\lambda = 1/\sqrt{n}$~~

$D = 1 \quad R = 2 \quad B = 1 \quad d = 10$

$n = 1000 \quad \delta = 0.05 \quad \lambda = \frac{1}{\sqrt{n}}$

$= \frac{1}{\sqrt{1000}}$

$$D_{eff} = \min(D, B/\lambda)$$

$$= \min(1, 1/1/\sqrt{1000})$$

$$= \min(1, \sqrt{1000})$$

$$= \min(1, 31.6)$$

$$= 1$$

Generalization error $\leq \frac{D_{eff} R}{\sqrt{n}} + B\sqrt{\frac{\log(1/\delta)}{n}}$

$$\leq \frac{1 \times 2}{\sqrt{1000}} + \frac{B\sqrt{\log(1/0.05)}}{\sqrt{1000}}$$

$$\leq \frac{2}{1000} + \sqrt{\log 20}$$

$$\leq 0.118$$

$\therefore$ Generalization error bound is approximately 11.8%.

Rademacher Complexity

Problem A

The empirical Rademacher Complexity $\hat{R}_n(H)$ of a class of real valued functions H defined on a domain z, given a sample $z_1, z_2 \dots z_n \in Z$ is defined as

$$\hat{R}_n(H) = E_\varepsilon\left[\sup_{h\in H} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(z_i)\right]$$

Where $\varepsilon_1, \varepsilon_2 \dots \varepsilon_n$ are iid

Basic properties

a) Scaling : For any scalar $c \in R$

$$\hat{R}_n(cH) = E_\varepsilon\left[\sup_{h\in H} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (c h(z_i))\right]$$

$$= |c| . E_\varepsilon\left[\sup_{h\in H} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(z_i)\right]$$

$$= |c| \hat{R}_n(H)$$

The absolute value arises because the supremum is sensitive to the sign of c

b) If $H \subseteq G$

$$\hat{R}_n(H) = E_\varepsilon\left[\sup_{h\in H} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(z_i)\right] \le E_\varepsilon\left[\sup_{g\in G} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g(z_i)\right]$$

$$= \hat{R}_n(G)$$

Since the supremum over a larger set G cannot be smaller than that over a subset H

Problem B

Given the class of linear Function

$$\mathcal{F} = \{ f_\theta(x) = \theta^T x : \|\theta\|_2 \leq D\}$$ with $\|x\|_2 \leq R$ the empirical Rademacher Complexity is:

$$\hat{R}_n(\mathcal{F}) = E_\varepsilon \left[\sup_{\|\theta\|_2 \leq D} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \theta^T x_i \right]$$

Using the duality of norms $(\sup_{\|\theta\|_2 \leq D} \theta^T v = D\|v\|_2)$ we have

$$\hat{R}_n(\mathcal{F}) = E_\varepsilon \left[\frac{D}{n} \left\| \sum_{i=1}^{n} \varepsilon_i x_i \right\|_2 \right]$$

By Jensen's inequality and the linearity of expectation:

$$\hat{R}_n(\mathcal{F}) \leq \frac{D}{n} \sqrt{E_\varepsilon \left[\left\| \sum_{i=1}^{n} \varepsilon_i x_i \right\|_2^2 \right]} = \frac{D}{n} \sqrt{\sum_{i=1}^{n} \|x_i\|_2^2}$$

$$\leq \frac{DR\sqrt{n}}{n}$$

$$= \frac{DR}{\sqrt{n}}$$

For $d=100$ $D=1$ $R=1$ $n=500$

$$\hat{R}_n(\mathcal{F}) \leq \frac{1.1}{\sqrt{500}}$$

$$\approx 0.0447$$

Problem c

Given the class $\mathcal{F}$ of single hidden layer ReLU networks

$$\mathcal{F} = \left\{ f(x) = \sum_{j=1}^{m} a_j \sigma(w_j^T x) : \|w_j\|_2 \le B, \|a\|_1 \le A \right\}$$

Symmetrization: The empirical Rademacher complexity is

$$\hat{R}_n(\mathcal{F}) = E_\varepsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) \right]$$

Contraction lemma:

$$\hat{R}_n(\mathcal{F}) \le A \cdot E_\varepsilon \left[\sup_{\|w_j\|_2 \le B} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i w_j^T x_i \right]$$

Norm and maxima

$$\hat{R}_n(\mathcal{F}) \le \frac{AB}{n} E_\varepsilon \left[\max_i \left| \sum_{i=1}^{n} \varepsilon_i x_i \right| \right] \le \frac{AB\sqrt{2\log(2m)}}{\sqrt{n}}$$

The $\sqrt{\log m}$ term arises from bounding the maximum over m hidden units

a) $\sqrt{\log m}$ term: Increasing m (no. of hidden units) increases the bound, but dependence is logarithmic which grows slowly.

b) As n increases the bound reduces because $\frac{1}{\sqrt{n}}$ increases generalization by decreasing the models capacity.

c) The constrains A and B control the models capacity. Smaller A or B reduces the bound suggesting regularization strategies like weigth decay or ~~at~~ architectural choices. to limit A and B

Dependence on input dimension d

The derived bound $\frac{DR}{\sqrt{n}}$ does not explicitly depend on d. This suggest that increasing the number of features does not necessarily increase the Rademacher complexity, provided the norms D and R are controlled. However, in practice higher d may lead to larger $|\theta|_2$ or $|x|_2$ indirectly affecting the bound.

as $n \to \infty$

As $n \to \infty$ the bound $\frac{DR}{\sqrt{n}} \to 0$. This decay is beneficial for generalization because it implies that the model's capacity to fit random noise diminishes with more data, reducing overfitting.

Tightness of the bound

The bound is tight in this example because it capture the worst case scenario where x_i are alligned and Rademacher variables ε_i maximize the norm. But in practice it may be smaller due to randomness in ε_i