Machine Learning Course Notes

He Li, Ziheng Huang, Zehua Lai

2017 Fall

Contents

Chapter 0. Preliminary

0.1 Basic Inequality

Markov's Inequality. *Random variable $x > 0$, $\mathbb{E}[x]$ exists, $\forall k > 0$, we have*

$$\mathbb{P}(X \geq k) \leq \frac{\mathbb{E}[x]}{k} \tag{1}$$

Chebychev's Inequality. *Random variable x, $\mathbb{E}[x]$ exists, $\mathrm{Var}(x) = \sigma^2$, $\forall k > 0$, we have*

$$\mathbb{P}(|x - \mathbb{E}[x]| \geq k) \leq \frac{\sigma^2}{k} \tag{2}$$

Homework. *Random variable x, $x \sim \mathcal{N}(0,1)$. Define function $\Phi(u) = \mathbb{P}(x \geq u)$. Find elementary function f, g, s.t.*

$$g(u) \leq \Phi(u) \leq f(u)$$

Chernoff's Inequality. *Random variable $x > 0$, $\mathbb{E}[x], \mathbb{E}[x^2], ...$ known, then $\forall k > 0$,*

$$\mathbb{P}(x \geq k) \leq \min_{i \geq 1} \frac{\mathbb{E}[x^i]}{k^i} \tag{3}$$

Definition. ***Moment generating function*** *is defined as:*

$$M_x(t) := \mathbb{E}[e^{tx}] = 1 + t\mathbb{E}[x] + \frac{t^2}{2}\mathbb{E}[x^2] + ...$$

Then it is obvious that

$$\mathbb{P}(x \geq k) \leq \inf_{t>0} e^{-tk}\mathbb{E}[e^{tx}] \tag{4}$$

Proof. When $x \geq k$

$$e^{tk} \leq e^{tx}$$

Given the fact that $e^{tx} > 0$,

$$\int_k^\infty e^{tk}\mathrm{d}x \leq \int_k^\infty e^{tx}\mathrm{d}x \leq \mathbb{E}[e^{tx}]$$

Therefore,

$$\mathbb{P}(x \geq k) \leq e^{-tk}\mathbb{E}[e^{tx}]$$

□

Definition. *Random variable x, $P = (p_1, ..., p_n)$. Then* ***Entropy*** *is defined as*

$$H(x) = \sum_{i=1}^n p_i \log(\frac{1}{p_i}) \tag{5}$$

Definition. *Random variable x, $P = (p_1, ..., p_n)$, $Q = (q_1, ..., q_n)$. Then* ***Related entropy*** *or* ***K-L divergence*** *is defined as*

$$D(P\|Q) := \sum_{i=1}^n p_i \log(\frac{p_i}{q_i}) \tag{6}$$

Note that K-L divergence is not symmetric, $D(P\|Q) \neq D(Q\|P)$.

0.2 Concentration Inequality

Concentration Inequality is one kind of inequality aiming at giving bound for the following function,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}x_i - \mathbb{E}(x)\right| \geq \epsilon\right) \tag{7}$$

Chernoff Bound(Homework). *$x_1...x_n$ are independent Bernoulli variable, $\mathbb{E}\{x_i\} = p$. Then $\forall\delta > 0$,*

$$P(\frac{1}{n}\sum x_i - \mathbb{E}[\frac{1}{n}\sum x_i] \geq \delta) \leq \exp\{-nD_B^{(e)}(p+\delta\|p)\} \tag{8}$$

Proof. Denote $\hat{x} = \sum_{i=1}^{n} x_i$, then $M_{\hat{x}}(t) = (1-p+pe^t)^n$. Then,

$$P\left(\hat{x} \geq n(p+\delta)\right) \leq \inf_{t>0} e^{-tn(p+\delta)}(1-p+pe^t)^n$$

The right-hand side is to minimize $-t(p+\delta) + \ln(1-p+pe^t)$. Therefore,

$$\frac{pe^t}{1-p+pe^t} = p+\delta$$

Solve it, we find that

$$\begin{aligned}\inf_{t>0} e^{-tn(p+\delta)}(1-p+pe^t)^n &= \exp\left\{-t(p+\delta) + \ln\frac{1-p}{1-p-\delta}\right\} \\ &= \exp\{-nD_B^{(e)}(p+\delta\|p)\}\end{aligned}$$

□

Note

1. if x is distributed on $[0,1]$ and $\mathbb{E}\{x_i\} = p$, x is not Bernoulli distribution, then by Jessan's inequality we have $\mathbb{E}[e^{tx}] < ...$, we know it is better.
2. if x are not identical distributed, only with independency the result remains the same.
3. **Additive Chernoff bound(Homework)**

$$\exp\{-nD_B^{(e)}(p+\delta\|p)\} \leq e^{-2n\delta^2} \tag{9}$$

Proof. Denote $f(\delta) = D_B(p+\delta\|p) - 2\delta^2$, $f(0) = 0$.

$$\begin{aligned}f'(\delta) &= \ln(\frac{p+\sigma}{p}) - \ln(\frac{1-p-\delta}{1-p}) - 4\delta \\ f''(\delta) &= \frac{2}{(p+\delta)(1-p-\delta)} - 4\end{aligned}$$

Therefore, $f'(0) = 0$ and $f''(\delta) \geq 0$ if $\delta \geq 0$. Therefore,

$$f(\delta) \leq f(0) = 0$$

Therefore,

$$\exp\{-nD_B(p+\delta\|p)\} \leq e^{-2n\delta^2}$$

□

Hoeffding Inequility. *$x_1...x_n$ are independent, and distributed on $[a_i, b_i]$, $\mathbb{E}\{x_i\} = p$. Then,*

$$P(\frac{1}{n}\sum x_i - \mathbb{E}[\frac{1}{n}\sum x_i] \geq \epsilon) \leq \exp\{-\frac{2n^2\epsilon^2}{\sum(b_i - a_i)^2}\} \tag{10}$$

Proof. We first prove $\mathbb{E}[e^{t(x-\mu)}] \leq \exp\{\frac{1}{8}t^2(b-a)^2\}$.

From Jenson's inequility, $\mathbb{E}[e^{t(x-\mu)}]$ is maximized when x is distributed on boundary of domain, which makes x a Bernoulli distribution. Assume $P(x = b) = p$,

$$\mathbb{E}[e^{t(x-\mu)}] = pe^{t(b-\mu)} + (1-p)e^{t(a-\mu)}$$
$$F(x) = \mathbb{E}[e^{t(x-\mu)}] - \exp\{\frac{1}{8}t^2(b-a)^2\}$$

with some calculation, we can find that $\max F(x) \leq 0$. Therefore, $\mathbb{E}[e^{t(x-\mu)}] \leq \exp\{\frac{1}{8}t^2(b-a)^2\}$. Then,

$$\begin{aligned}
P\left(\frac{1}{n}\sum_{i=1}^n x_i - \frac{1}{n}\sum_{i=1}^n \mu_i \geq \epsilon\right) &= P\left(\sum_{i=1}^n x_i - \sum_{i=1}^n \mu_i \geq n\epsilon\right) \\
&\leq \inf_{t>0} e^{-tn\epsilon}\mathbb{E}[e^{t(\sum_{i=1}^n x_i - \sum_{i=1}^n \mu_i)}] \\
&= \inf_{t>0} e^{-tn\epsilon}\prod_{i=1}^n \mathbb{E}[e^{t(x_i-\mu_i)}] \\
&\leq \inf_{t>0} \exp\{-tn\epsilon + \frac{1}{8}t^2\sum_{i=1}^n (b_i - a_i)^2\} \\
&\leq \exp\{\frac{-2n^2\epsilon^2}{\sum_{i=1}^n (a_i - b_i)^2}\}
\end{aligned}$$

□

Question: without independency, is concentration here exists? No, since $\mathbb{E}(AB) \neq \mathbb{E}(A)\mathbb{E}(B)$.

Definition. *Random variables $S_0...S_n...$ are **martingale** if $\forall i$, (fair game)*

$$\mathbb{E}[S_i|S_{i-1}, ...S_0] = S_{i-1}$$

*Denote $X_i = S_i - S_{i-1}$, X_i is called **martingale difference**.*

Azuma's Inequality. *$x_1...x_n$ are martingale difference, $|x_i| \leq C_i$, $S_0 = 0$, $\mathbb{E}\{x_i\} = p$. Then,*

$$P(\frac{1}{n}\sum x_i - \mathbb{E}[\frac{1}{n}\sum x_i] \geq \epsilon) \leq \exp\{-\frac{2n^2\epsilon^2}{\sum C_i^2}\} \tag{11}$$

Definition. *Random variables $S_0...S_n...$ are **super martingale** if $\forall i$,*

$$\mathbb{E}[S_i|S_{i-1}, ...S_0] \leq S_{i-1}$$

*Denote $X_i = S_i - S_{i-1}$, X_i is called **super martingale difference**.*

Note that **Azuma's Inequality** also holds for super martingale difference. Given that $x_1...x_n$ are super martingale difference, $|x_i| \leq C_i$, $S_0 = 0$, $\mathbb{E}\{x_i\} = p$. Then,

$$P(\frac{1}{n}\sum x_i - \mathbb{E}[\frac{1}{n}\sum x_i] \geq \epsilon) \leq \exp\{-\frac{2n^2\epsilon^2}{\sum C_i^2}\}$$

Definition. *$x_1...x_n$ are independent random variable, function f is called **stable** if it satisfies that $\forall i$, $\forall x_1...x_n; x_1, ..., x_i', ..., x_n$,*

$$|f(x_1, ..., x_n) - f(x_1, ..., x_i', ..., x_n)| \leq c_i \tag{12}$$

McDiarmid Lemma. *$x_1...x_n$ are independent random variable, f is stable, then*

$$\mathbb{P}\left\{f(x_1,...,x_n) - \mathbb{E}[f(x_1,...,x_n)] \geq \epsilon\right\} \leq \exp\{-\frac{2\epsilon^2}{\sum C_i^2}\} \tag{13}$$

Draw with/without replacement. *$a_1,...,a_N \in \{0,1\}$ are random variables, uniformly distributed. Draw n numbers from $a_1,...,a_N$, denote $x_1,...,x_n$. Consider $P(\frac{1}{n}\sum x_i - \mathbb{E}[\frac{1}{n}\sum x_i] \geq \epsilon)$*

1. Draw with replacement: $x_1,...,x_n$ independent, Chernoff bound holds.

2. Draw without replacement: $x_1,...,x_n$ not independent, Chernoff bound holds. Actually, it's more concentrated.

0.3 σ^2-subgaussian

Definition. *X is called σ^2-subgaussian if*

$$\log E[e^{\lambda(X-\mathbb{E}X)}] \leq \frac{1}{2}\lambda^2\sigma^2 = \log\left\{MGF(\mathcal{N}(0,\delta^2))\right\} \tag{14}$$

where MGF is moment generating function.

Chernoff Bound. *If X is σ^2-subgaussian, then,*

$$\mathbb{P}[X > \mathbb{E}X + t] \leq e^{-\frac{t^2}{2\sigma^2}} \tag{15}$$

Proof. Using moment generating function,

$$\begin{aligned}
\mathbb{P}[X > \mathbb{E}X + t] &\leq \inf_{\lambda>0} e^{-\lambda t}\mathbb{E}\left[e^{\lambda[X-\mathbb{E}(X)]}\right] \\
&\leq \inf_{\lambda>0} e^{-\lambda t}e^{\frac{1}{2}\lambda^2\sigma^2} \\
&= e^{-\frac{t^2}{2\sigma^2}}
\end{aligned}$$

□

Hoeffding. *If $a \leq X \leq b$ then X is $\frac{1}{4}(b-a)^2$-subgaussian.*

Proof. Let $\log \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] = \phi(\lambda)$, then

$$\begin{aligned}
\phi'(\lambda) &= \frac{\mathbb{E}\left[(X-\mathbb{E}X)e^{\lambda(X-\mathbb{E}X)}\right]}{\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]} \\
\phi''(\lambda) &= \frac{\mathbb{E}\left[(X-\mathbb{E}X)^2e^{\lambda(X-\mathbb{E}X)}\right]}{\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]} - \frac{\mathbb{E}\left[(X-\mathbb{E}X)e^{\lambda(X-\mathbb{E}X)}\right]}{\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]^2} \\
&\leq \frac{\mathbb{E}\left[(X-\mathbb{E}X)^2e^{\lambda(X-\mathbb{E}X)}\right]}{\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]} \\
&\leq \frac{1}{4}(a-b)^2
\end{aligned}$$

Here

$$\frac{\mathbb{E}\left[(X-\mathbb{E}X)e^{\lambda(X-\mathbb{E}X)}\right]}{\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]^2} \geq \frac{\mathbb{E}\left[(X-\mathbb{E}X)\right]}{\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]^2} = 0$$

because when $X - \mathbb{E}(X) > 0$, $e^{\lambda(X-\mathbb{E}X)} > 1$ and when $X - \mathbb{E}(X) \leq 0$, $e^{\lambda(X-\mathbb{E}X)} \leq 1$. And $\frac{\mathbb{E}\left[(X-\mathbb{E}X)^2e^{\lambda(X-\mathbb{E}X)}\right]}{\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]} \leq \frac{1}{4}(a-b)^2$ can be proved by the change of probability measure. □

Azuma. *X_t are random variables, $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = X_{t-1}$ (martingale), $\Delta_t = X_t - X_{t-1}$. If $a_t \le \Delta_t \le b_t$, then X_t is $\frac{1}{4}\sum_{i=1}^{t}(b_i - a_i)^2$-subgaussian.*

Proof. $\mathbb{E}X_t = 0$

$$\begin{aligned}
\mathbb{E}e^{\lambda X_t} &= \mathbb{E}e^{\lambda(X_{t-1}+\Delta_t)} \\
&= \mathbb{E}\left\{\mathbb{E}\left[e^{\lambda(X_{t-1}+\Delta_t)}|\mathcal{F}_{t-1}\right]\right\} \\
&= \mathbb{E}\left\{e^{\lambda X_{t-1}}E\left[e^{\lambda\Delta_t}|\mathcal{F}_{t-1}\right]\right\} \\
&\le \mathbb{E}\left[e^{\lambda X_{t-1}}\right]e^{\frac{1}{8}\lambda^2(b_t-a_t)^2} \\
&... \\
&\le \exp\left\{\frac{1}{8}\sum_{i=1}^{t}(b_i-a_i)^2\right\}
\end{aligned}$$

□

McDiarmid. *$X_1, X_2, ..., X_n$ independent r.v. $f(x_1, x_2, ..., x_n)$*

$$D_i f = sup_{x_{-i}} sup_{x,y}|f(x_{-i}, x) - f(x_{-i}, y)|$$

then $f(X_1, X_2, ..., X_n)$ is $\frac{1}{4}\sum_{i=1}^{n} D_i f^2$-subgaussian.[1]

Proof. Let $Z_i = E[f(X_1, ..., X_n)|X_1, ..., X_i]$, then $f - E[f] = \sum_{i=1}^{n} Z_i - Z_{i-1}$. Note that

$$E[Z_i|X_1, ..., X_{i-1}] = E[f(X_1, ..., X_n)|X_1, ..., X_{i-1}] = Z_{i-1}$$

Z_i is martingale (Doob martingale). □

Draw with/without replacement. $E[e^{\lambda(X_1+...+X_n)}] \le E[e^{\lambda(X'_1+...+X'_n)}]$.

Proof. Taylor expansion: $E[(X_1 + ... + X_n)^i] \le E[(X'_1 + ... + X'_n)^i]$ □

[1] Probability in High Dimension (Princeton).

Chapter 1. VC Dimension

1.1 Empirical Risk Minimiaztion

We take some simple classifiers as examples. Consider $(x_i, y_i)_{i=1}^n$, $x_i \in \mathbb{R}^d, x = (x^{(1)}, ..., x^{(d)})$ are called instance, $y \in \{\pm 1\}$ are called label.

1. Decision Tree: hypothesis space $\mathcal{F} = \{\text{all decision tree}, \text{depth} \leq \alpha\}$
2. Linear Classifier: hypothesis space $\mathcal{F} = \left\{(\omega, b) : \omega \in \mathbb{R}^d, b \in \mathbb{R}, \|\omega\| = 1\right\}$, then classifier is:

$$f(x) = \text{sgn}(\omega^T x + b)$$

Definition. *Given $\mathcal{F}$, training data $(x_i, y_i)_{i=1}^n$. **Empirical Risk Minimiaztion(ERM)** is a algorithm, which finds $f \in \mathcal{F}$, s.t.*

$$\hat{f} := \underset{f \in \mathcal{H}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} I\left[y_i \neq f(x_i)\right] \tag{16}$$

Definition. ***Empirical Error*** *measures $\hat{f}$'s performance on training data, denoted as*

$$\mathbb{P}_S(y_i \neq f(x_i)) := \frac{1}{n} \sum_{i=1}^{n} I\left[y_i \neq \hat{f}(x_i)\right] \tag{17}$$

where $S = (x_i, y_i)_{i=1}^n$

Generalization Error *measures $\hat{f}$'s performance on test data, denoted as*

$$\mathbb{P}_D(y_i \neq f(x_i)) = \underset{(x,y)\sim D}{\mathbb{E}} \{I(y \neq \hat{f}(x))\} \tag{18}$$

Note that a small empirical error cannot indicate a small generalization error(cannot using Chernoff bound), since on training data, $z_i = I\left[y_i \neq \hat{f}(x_i)\right]$ are not independent.

1.2 Finite Hypothesis Space

We consider finite hypothesis space first.

Union Bound. *Consider $\mathcal{F}$ is finite, $|\mathcal{F}| < \infty$. ERM learns $\hat{f} \in \mathcal{F}$, then*

$$P\left\{\mathbb{P}_D(y_i \neq f(x_i)) - \mathbb{P}_S(y_i \neq f(x_i)) \geq \epsilon\right\} \leq |\mathcal{F}| e^{-2n\epsilon^2} \tag{19}$$

Proof. Fix $f \in \mathcal{F}$, $P\left\{\mathbb{P}_D(y_i \neq f(x_i)) - \mathbb{P}_S(y_i \neq f(x_i)) \geq \epsilon\right\}$ satisties Chernoff bound.

$$P\left\{\mathbb{P}_D(y_i \neq f(x_i)) - \mathbb{P}_S(y_i \neq f(x_i)) \geq \epsilon\right\} \leq e^{-2n\epsilon^2}$$

Therefore, we have union bound

$$P\left\{\exists f \in \mathcal{F},\ \mathbb{P}_D(y_i \neq f(x_i)) - \mathbb{P}_S(y_i \neq f(x_i)) \geq \epsilon\right\} \leq |\mathcal{F}| e^{-2n\epsilon^2}$$

□

1.3 VC Bound

Now we consider the situation in infinite hypothesis space. Denote $z_i = (x_i, y_i)$, $\phi_f(z_i) = I(y_i \neq f(x_i)$. Consider:

$$\mathbb{P}\left(\sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_i \phi_f(z_i) - \mathbb{E}[\phi_f(z)]\right| \geq \epsilon\right) \tag{20}$$

Step I (Double Sample Trick)

Proposition. *$X_1, ..., X_n, X_{n+1}, ..., X_{2n}$ are i.i.d Bernoulli R.V., $\mathbb{E}(X) = p$. Denote $\nu_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\nu_2 = \frac{1}{n}\sum_{i=n+1}^{2n} X_i$. If $n \geq \frac{\ln 2}{\epsilon^2}, \epsilon > 0$, then*

$$\frac{1}{2}\mathbb{P}\left(|\nu_1 - p| \geq 2\epsilon\right) \leq \mathbb{P}\left(|\nu_1 - \nu_2| \geq \epsilon\right) \leq 2\mathbb{P}\left(|\nu_1 - p| \geq \frac{\epsilon}{2}\right) \tag{21}$$

Proof. Right inequality:

$$|\nu_1 - \nu_2| \geq \epsilon \Longrightarrow |\nu_1 - p| \geq \frac{\epsilon}{2} \text{ or } |\nu_2 - p| \geq \frac{\epsilon}{2}$$

Therefore,

$$\begin{aligned}
\{|\nu_1 - \nu_2| \geq \epsilon\} &\subset \{|\nu_1 - p| \geq \frac{\epsilon}{2}\} \cup \{|\nu_2 - p| \geq \frac{\epsilon}{2}\} \\
\mathbb{P}\left(|\nu_1 - \nu_2| \geq \epsilon\right) &\leq \mathbb{P}\left(|\nu_1 - p| \geq \frac{\epsilon}{2} \cup |\nu 2 - p| \geq \frac{\epsilon}{2}\right) \\
&\leq \mathbb{P}\left(|\nu_1 - p| \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(|\nu_2 - p| \geq \frac{\epsilon}{2}\right) \\
&= 2\mathbb{P}\left(|\nu_1 - p| \geq \frac{\epsilon}{2}\right)
\end{aligned}$$

Similarily, left inequality:

$$|\nu_1 - p| \geq 2\epsilon, |\nu_2 - p| \leq \epsilon \Longrightarrow |\nu_1 - \nu_2| \geq \epsilon$$

□

Lemma(Homework).

$$\begin{aligned}
&\frac{1}{2}\mathbb{P}\left(\sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} \phi_f(z_i) - \mathbb{E}[\phi_f(z)]\right| \geq 2\epsilon\right) \\
\leq &\mathbb{P}\left(\sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} \phi_f(z_i) - \frac{1}{n}\sum_{i=n+1}^{2n} \phi_f(z_i)\right| \geq \epsilon\right) \\
\leq &2\mathbb{P}\left(\sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} \phi_f(z_i) - \mathbb{E}[\phi_f(z)]\right| \geq \frac{\epsilon}{2}\right)
\end{aligned} \tag{22}$$

Proof. We know that $\phi_f(z_1), ..., \phi_f(z_n), \phi_f(z_{n+1}), ..., \phi_f(z_{2n})$ are i.i.d Bernoulli R.V.. $\mathbb{E}(\phi_f(z)) = p$. Denote $\nu_1 = \frac{1}{n}\sum_{i=1}^{n} \phi_f(z_i)$, $\nu_2 = \frac{1}{n}\sum_{i=n+1}^{2n} \phi_f(z_i)$.

Consider the right inequality, we have

$$\sup_{f\in\mathcal{H}} |\nu_1 - \nu_2| \geq \epsilon \Longrightarrow \sup_{f\in\mathcal{H}} |\nu_1 - p| \geq \frac{\epsilon}{2} \text{ or } \sup_{f\in\mathcal{H}} |\nu_2 - p| \geq \frac{\epsilon}{2}$$

Therefore,

$$\{\sup_{f\in\mathcal{H}}|\nu_1-\nu_2|\geq\epsilon\}\subset\{\sup_{f\in\mathcal{H}}|\nu_1-p|\geq\frac{\epsilon}{2}\}\cup\{\sup_{f\in\mathcal{H}}|\nu_2-p|\geq\frac{\epsilon}{2}\}$$

$$\begin{aligned}\therefore\ \mathbb{P}\left(\sup_{f\in\mathcal{H}}|\nu_1-\nu_2|\geq\epsilon\right)&\leq\mathbb{P}\left(\sup_{f\in\mathcal{H}}|\nu_1-p|\geq\frac{\epsilon}{2}\cup\sup_{f\in\mathcal{H}}|\nu 2-p|\geq\frac{\epsilon}{2}\right)\\&\leq\mathbb{P}\left(\sup_{f\in\mathcal{H}}|\nu_1-p|\geq\frac{\epsilon}{2}\right)+\mathbb{P}\left(\sup_{f\in\mathcal{H}}|\nu_2-p|\geq\frac{\epsilon}{2}\right)\\&=2\mathbb{P}\left(\sup_{f\in\mathcal{H}}|\nu_1-p|\geq\frac{\epsilon}{2}\right)\end{aligned}$$

Similarily, assume $f\in\mathcal{H}$ such that $\sup_{f\in\mathcal{H}}|\nu_1-p|=|\nu_1'-p'|$, then

$$\sup_{f\in\mathcal{H}}|\nu_1-p|\geq 2\epsilon \text{ and } |\nu_2'-p'|\leq\epsilon\Longrightarrow\sup_{f\in\mathcal{H}}|\nu_1-\nu_2|\geq\epsilon$$

$$\therefore\left\{\sup_{f\in\mathcal{H}}|\nu_1-p|\geq 2\epsilon\right\}\cap\{|\nu_2'-p'|\leq\epsilon\}\subset\left\{\sup_{f\in\mathcal{H}}|\nu_1-\nu_2|\geq\epsilon\right\}$$

$$\therefore\mathbb{P}\left(\left\{\sup_{f\in\mathcal{H}}|\nu_1-\nu_2|\geq\epsilon\right\}\right)\geq\mathbb{P}\left(\left\{\sup_{f\in\mathcal{H}}|\nu_1-p|\geq 2\epsilon\right\}\right)\mathbb{P}\left(\{|\nu_2'-p'|\leq\epsilon\}\right)$$

We have an upper bound on $\mathbb{P}\left(\{|\nu_2'-p'|\geq\epsilon\}\right)$. When $n\geq\frac{\ln 2}{2\epsilon^2}$,

$$\begin{aligned}\mathbb{P}\left(\{|\nu_2'-p'|\leq\epsilon\}\right)&=1-\mathbb{P}\left(\{|\nu_2'-p'|\geq\epsilon\}\right)\\&\geq 1-\exp\{-2n\epsilon^2\}\\&\geq\frac{1}{2}\end{aligned}$$

$$\begin{aligned}\therefore\ \mathbb{P}\left(\left\{\sup_{f\in\mathcal{H}}|\nu_1-\nu_2|\geq\epsilon\right\}\right)&\geq\mathbb{P}\left(\left\{\sup_{f\in\mathcal{H}}|\nu_1-p|\geq 2\epsilon\right\}\right)\mathbb{P}\left(\{|\nu_2'-p'|\leq\epsilon\}\right)\\&\geq\frac{1}{2}\mathbb{P}\left(\left\{\sup_{f\in\mathcal{H}}|\nu_1-p|\geq 2\epsilon\right\}\right)\end{aligned}$$

□

Step II (Symmetrization)

Definition. *Denote* $N^{\mathcal{H}}(z_1,...,z_n)$*:*

$$N^{\mathcal{H}}(z_1,...,z_n):=\#\left\{(\phi_f(z_1),...,\phi_f(z_n)),f\in\mathcal{H}\right\}$$

If $N^{\mathcal{H}}(z_1,...,z_n)=2^n$*, then we call* $\mathcal{H}$ ***shatters*** $\{z_1,...,z_n\}$*.*

Also $N^{\mathcal{H}}(n)$ *is* ***growth function***

$$N^{\mathcal{H}}(n):=\max_{z_1...z_n}N^{\mathcal{H}}(z_1...z_n)$$

Lemma.

$$\mathbb{P}\left(\sup_{f\in\mathcal{H}}|\nu_1(z)-\nu_2(z)|\geq\epsilon\right)\leq\mathbb{E}\left[N^{\mathcal{H}}(z_1,...,z_n)\right]2e^{-2n\epsilon^2}\leq N^{\mathcal{H}}(2n)\,2e^{-2n\epsilon^2}\tag{23}$$

where $\nu_1(z)=\frac{1}{n}\sum_{i=1}^{n}\phi_f(z_i)$*,* $\nu_2(z)=\frac{1}{n}\sum_{i=n+1}^{2n}\phi_f(z_i)$*.*

Proof. Draw a set with permutation, fix set (draw without replacement). Then,

$$\mathop{\mathbb{P}}_{\text{permutation}}\left(\sup_{f\in\mathcal{H}}|\nu_1(z)-\nu_2(z)|\geq\epsilon\right)\leq N^{\mathcal{H}}(z_1,...,z_n)\,2e^{-2n\epsilon^2}$$

Take expectation,

$$\mathop{\mathbb{E}}_{\text{set}}\left\{\mathop{\mathbb{P}}_{\text{permutation}}\left(\sup_{f\in\mathcal{H}}|\nu_1(z)-\nu_2(z)|\geq\epsilon\right)\right\}\leq\mathbb{E}\left[N^{\mathcal{H}}(z_1,...,z_n)\right]\,2e^{-2n\epsilon^2}$$

□

Step III (VC-Dimension)

Definition. ***VC Dimension*** *of hypothesis space $\mathcal{H}$ is defined as the maximum value of d s.t.*

$$N^{\mathcal{H}}(d)=2^d$$

In other words, $\exists\{z_1,...,z_d\}$, can be shattered by $\mathcal{H}$ while $\forall\{z_1,...,z_{d+1}\}$, cannot be shattered by $\mathcal{H}$.

For growth function $N^{\mathcal{H}(n)}$, we only know that $N^{\mathcal{H}}(n)\leq 2^n$. However, we don't know how it grows actually, maybe exponential or polynomial. When n is small can be exponential but when n is large it is relatively smaller.

Sauer's Lemma. *For a set of n elements, $\mathcal{N}=\{1,...,n\}$, set $\mathcal{H}\subset 2^{\mathcal{N}}$. Then,*

$$|\mathcal{H}|\leq\#\{x\subset\mathcal{N}\mid x\text{ is shattered by }\mathcal{H}\}\tag{24}$$

Proof. We first prove the Sauer's Lemma holds for a set of 1 elements. Note that empty set can be shattered by any set. Given this, $\#\{x\subset\mathcal{N}\mid x\text{ is shattered by }\mathcal{H}\}\geq 1$. And $\#\{x\subset\mathcal{N}\mid x\text{ is shattered by }\mathcal{H}\}=2$ here if and only if $\mathcal{H}=2^{\mathcal{N}}$. Therefore, $|\mathcal{H}|\leq\#\{x\subset\mathcal{N}\mid x\text{ is shattered by }\mathcal{H}\}$.

Assume the lemma holds for any set with k elements. We add x into such set. For a $\mathcal{H}$, we divide it into two subsets where $\mathcal{H}_1$ contains sets that contains x and $\mathcal{H}_2$ contains sets that does not contain x. Denote $\alpha(\mathcal{H}_i):=\#\{x\subset\mathcal{N}\mid x\text{ is shattered by }\mathcal{H}_i\}$. From the assumption,

$$\begin{aligned}|\mathcal{H}_1|&\leq\alpha(\mathcal{H}_1)\\|\mathcal{H}_2|&\leq\alpha(\mathcal{H}_2)\end{aligned}$$

Denote subsets of $\mathcal{N}$ that can be shattered by both $\mathcal{H}_1$ and $\mathcal{H}_2$ as P. For $\mathcal{H}=\mathcal{H}_1\cup\mathcal{H}_2$, the union of $\mathcal{H}_1$ and $\mathcal{H}_1$ can shatter new subsets of $\mathcal{N}\cup\{x\}$ that has the following form: $P\cup\{x\}$. Therefore,

$$\begin{aligned}\alpha(\mathcal{H})&=\alpha(\mathcal{H}_1\cup\mathcal{H}_2)\\&=\alpha(\mathcal{H}_1)+\alpha(\mathcal{H}_2)-\alpha(\mathcal{H}_1\cup\mathcal{H}_2)+\alpha(\mathcal{H}_1\cup\mathcal{H}_2)\\&\geq|\mathcal{H}_1|+|\mathcal{H}|\\&=|\mathcal{H}|\end{aligned}$$

□

Corollary. *For a hypothesis space $\mathcal{H}$, if $VC(\mathcal{H})=d$, then*

$$N^{\mathcal{H}}(n)\begin{cases}=2^n & \text{if } n\leq d\\ \leq\sum_{k=0}^{d}\binom{n}{k}\leq\left(\frac{ne}{d}\right)^d & \text{if } n>d\end{cases}\tag{25}$$

Proof. By the definition of VC Dimension, $N^{\mathcal{H}}(n) = 2^n$ when $n \leq d$. When $n > d$,

$$\begin{aligned}
\sum_{k=0}^{d} C_n^k &= \sum_{k=0}^{d} \frac{n...(n-k+1)}{k!} \\
&\leq \sum_{k=0}^{d} \frac{n^k}{k!} \\
&= \sum_{k=0}^{d} \frac{d^k}{k!} \frac{n^k}{d^k} \\
&\leq \left(\frac{n}{d}\right)^d \sum_{k=0}^{d} \frac{d^k}{k!} \\
&\leq \left(\frac{n}{d}\right)^d \sum_{k=0}^{+\infty} \frac{d^k}{k!} \\
&= \left(\frac{ne}{d}\right)^d
\end{aligned}$$

□

Theorem.

$$\mathbb{P}\left(\sup_{f\in\mathcal{H}} \left| \frac{1}{n}\sum_i \phi_f(z_i) - \mathbb{E}[\phi_f(z)] \right| \geq \epsilon \right) \leq \exp\left\{-O(n\epsilon^2)\right\} \left(\frac{ne}{d}\right)^d \tag{26}$$

Theorem. *$\forall \delta > 0$, with probability $1-\delta$ over the random draw of training data $\{(x_i, y_i)\}$, generalization error can be controled by empirical error and VC Dimension.*

$$\mathbb{P}_D(y \neq f(x)) \leq \mathbb{P}_S(y \neq f(x)) + O(\sqrt{\frac{d\ln n + \ln\frac{1}{\delta}}{n}}) \tag{27}$$

where d is the VC Dimension of hypothesis space. For ERM, denote $\hat{f}$ the optimal function $\hat{f} = \arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq f(x_i))$, and $f^* = \arg\min_{f\in\mathcal{H}} \mathbb{P}_D(y \neq f(x))$, then with probability $1-\delta$,

$$\mathbb{P}_D(y \neq f^*(x)) \leq \mathbb{P}_S(y \neq \hat{f}(x)) + O(\sqrt{\frac{d\ln n + \ln\frac{1}{\delta}}{n}}) \tag{28}$$

Note that for linear classifier, its VC Dimension $= r + 1$ where $x \in \mathbb{R}^r$.

Chapter 2. Supported Vector Machine

In previous chapter, we use ERM which finds

$$\hat{f} = \arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq f(x_i)) \tag{29}$$

However, indicator function is hard to minimize. So we change classification error from 0-1 loss to other functions.

2.0 Appendix Game Theory

Game Theory. *In two player zero-sum game, We have minmax = maxmin in a Nash equilibrium.*[2]

$$\max_{y\in A_2}\min_{x\in A_1} u_1(x,y) = \min_{x\in A_1}\max_{y\in A_2} u_1(x,y) \tag{30}$$

Proof. Denote $\min_{i\le n}\max_{j\le m} a_{ij} = a_{pq}$, $\max_{i\le m}\min_{j\le n} a_{ji} = a_{rt}$. Then

$$a_{pq} \ge a_{rq} \ge a_{rt}$$

Therefore,

$$\min_{i\le n}\max_{j\le m} a_{ij} \ge \max_{i\le m}\min_{j\le n} a_{ji}$$

□

Theorem. *For a matrix M, $\min_i\max_j M_{ij} \ge \max_j\min_i M_{ij}$ which means that if two player choose pure strategy and to act sequentially, second one gets an advantage.*

Note that if two players can choose mixed strategy, a Nash equilibrium exists, and the equation above holds.

Saddle Point Theorem. *Consider function $f(x,y)$, if fix y, $f(\cdot,y)$ is convex; if fix x, $f(x,\cdot)$ is concave. Then,*

$$\max_y\min_x f(x,y) = \min_x\max_y f(x,y) \tag{31}$$

2.1 KKT Conditions

Lagrange Duality. *Our **primal problem** is*

$$\begin{cases} & \min f(x) \\ \text{s.t.} & g_i(x) \le 0 \\ & h_i(x) = 0 \end{cases} \tag{32}$$

$$\iff \begin{cases} & \min_x\max_{\mu,\lambda} f(x) + \sum \mu_i h_i(x) + \sum \lambda_i g_i(x) \\ \text{s.t.} & \lambda_i \ge 0 \end{cases} \tag{33}$$

*And our **dual problem** is*

$$\begin{cases} & \max_{\mu,\lambda}\min_x f(x) + \sum \mu_i h_i(x) + \sum \lambda_i g_i(x) \\ \text{s.t.} & \lambda_i \ge 0 \end{cases} \tag{34}$$

[2] A course in Game Theory, P25

Denote $L(x,\mu,\lambda) = f(x) + \sum \mu_i h_i(x) + \sum \lambda_i g_i(x)$, (x_1,μ_1,λ_1) the solution of primal problem, (x_2,μ_2,λ_2) the solution of dual problem. Then we have

$$L(x_1,\mu_1,\lambda_1) \geq L(x_2,\mu_2,\lambda_2)$$

KKT-conditions. *The following 4 conditions are KKT conditions:*

(1) Stationary $\nabla L(x,\lambda,\mu)|_{x^*,\lambda^*,\mu^*} = 0$
(2) Primal Feasibility $h_i(x^*) = 0, g_i(x^*) \leq 0$
(3) Dual Feasible $\lambda^* \geq 0$
(4)Complementary slackness $\lambda_i g_i(x^*) = 0$

Theorem. *KKT conditions are necessary condition. If primal and duality problem have a same solution, then this solution satisfies KKT conditions.*

Proof. Denote (x^*,λ^*,μ^*) the solution of primal and dual problems. Obviously, it satisfies (1), (2) and (3) of KKT conditions.

For (4), consider the primal problem, if $g_i(x^*) < 0$, then $\lambda_i(x^*)$ should equal to 0 in order to max the primal function. Therefore, $\lambda_i g_i(x^*) = 0$. □

Theorem. *If $f(x), g_i(x)$ are convex and $h_i(x)$ is linear, then KKT conditions are sufficient condition, which means*

$$\begin{cases} \min\limits_x \max\limits_{\mu,\lambda} f(x) + \sum \mu_i h_i(x) + \sum \lambda_i g_i(x) \\ \text{s.t.} \quad \lambda_i \geq 0 \end{cases} \iff \begin{cases} \max_{\mu,\lambda} \min_x f(x) + \sum \mu_i h_i(x) + \sum \lambda_i g_i(x) \\ \text{s.t.} \quad \lambda_i \geq 0 \end{cases}$$

Proof. First, a x_0 satisties the above conditions also satisfies the constraints in primal and dual problems.

Denote $x1$ the solution of primal problem and x_2 the solution of dual problem, deonte $L(x,\mu,\lambda) = f(x) + \sum\limits_i \mu_i h_i(x) + \sum\limits_i \lambda_i g_i(x)$.

With the additional condition, $L(x,\mu,\lambda)$ is a convex function w.r.t x, then $x_0 = \arg\min\limits_x L(x,\mu,\lambda)$. Therefore, $L(x_0,\mu,\lambda) \leq L(x_2,\mu,\lambda) \leq L(x_1,\mu,\lambda)$.

Note that x_0 satisfies primal constraints and x_2 is the argmin of primal problem. Therefore, $L(x_1,\mu,\lambda) \leq L(x_0,\mu,\lambda)$. Therefore, we have our conclusion that

$$L(x_0,\mu,\lambda) = L(x_2,\mu,\lambda) = L(x_1,\mu,\lambda)$$

□

Let $L(x,\mu,\lambda) = f(x) + \sum \mu_i h_i(x) + \sum \lambda_i g_i(x)$, solve $\frac{\partial L}{\partial x} = 0$ we can get $x^* = \phi(\mu,\lambda)$. It suffices to solve $max_{\mu,\lambda} f(\phi(\mu,\lambda)) + \sum \mu_i h_i(\phi(\mu,\lambda)) + \sum \lambda_i g_i(\phi(\mu,\lambda))$.

2.2 Supported Vector Machine

Linear Classifier. *Note that $x \in \mathbb{R}^d$, $y \in \{\pm 1\}$, $\mathcal{H} = \{f(x) | f(x) = \text{sgn}(w^T x + b)\}$. Here we want to solve a constrained optimization problem,*

$$\begin{aligned} &\max_{w,b} t \quad \text{s.t.} \\ &\begin{cases} y_i(w^T x_i + b) \geq t \\ \|w\| = 1 \end{cases} \end{aligned} \tag{35}$$

The above can also be rewirtten as

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad &y_i(w^T x_i + b) \geq 1 \end{aligned} \tag{36}$$

The linear classifier problem is

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad y_i(w^T x_i + b) \geq 1 \end{cases} \iff \begin{cases} \min\limits_{w,b,\lambda_i} L(w,b) = \frac{1}{2}\|w\|^2 - \sum \lambda_i [y_i(w^T x_i + b) - 1] \\ \text{s.t.} \quad \lambda_i \geq 0 \end{cases} \tag{37}$$

$L(w,b) = \frac{1}{2}\|w\|^2 - \sum \lambda_i [y_i(w^T x_i + b) - 1]$, $w^* = \sum \lambda_i y_i x_i$, $\sum \lambda_i y_i = 0$. By KKT condition, $\lambda_i^* [y_i(w^{*T} x_i + b^*) - 1] = 0$. $\lambda_i^* = 0$ for all (x_i, y_i) that are not closest to the hyperspace. $\lambda_i^* \neq 0$ for all support vector. It is the **Support Vector Machine (SVM)**.

2.3 Soft-margin SVM

Soft-margin SVM. *How to find a linear classifier when the data set is not seperable? The soft-margin SVM can be defined as:*

$$\begin{aligned} &\min_{w,b,\xi_i} \frac{1}{2}\|\omega\|^2 + C\sum_i \xi_i \\ &s.t. \quad \begin{cases} y_i(\omega^T x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \end{aligned} \tag{38}$$

The soft-margin SVM can be rewritten as

$$\begin{aligned} &\max_{\lambda_i} \sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j (x_i^T x_j) \\ &s.t. \quad \begin{cases} 0 \leq \lambda_i \leq C \\ \sum\limits_i \lambda_i y_i = 0 \end{cases} \end{aligned} \tag{39}$$

Hinge Loss. *The above sotf-margin SVM can also be rewritten as a optimization problem without constraint. Using hinge loss, the above problem is as same as:*

$$\min_{\omega,b} \frac{1}{2}\|\omega\|^2 + C\sum_i \left[1 - y_i(\omega^T x_i + b)\right]_+ \tag{40}$$

where

$$x_+ = \begin{cases} 0 & x \leq 0 \\ x & x > 0 \end{cases} \tag{41}$$

Here we use hinge loss as a surrogate loss of 0-1 loss, which has the following two good properties:

- hinge loss is the upper bound of 0-1 loss.
- hinge loss is computationally efficient.
- although hinge loss is not differentiable everywhere, it is convex.

2.4 Kernel Method

Sometimes we want to do some mapping on the original space.

$$\begin{aligned} &\max_{\lambda_i} \sum_i \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j (\phi(x_i)^T \phi(x_j)) \\ &s.t. \quad \begin{cases} 0 \le \lambda_i \le C \\ \sum_i \lambda_i y_i = 0 \end{cases} \end{aligned} \tag{42}$$

where $x = (x^{(1),\dots,x^{(d)}})$, and for example

$$x :\mapsto \phi(x) = (x^{(1)}, ..., x^{(d)}, [x^{(1)}]^2, [x^{(1)}x^{(2)}], ..., [x^{(d)}]^2)$$

However, sometimes we cannot have a explicit form of $\phi(\cdot)$. We have **kernel trick**.

Kernel Trick. *We define a binary function $K(\cdot, \cdot)$,*

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \tag{43}$$

For example, Gaussian Kernel is

$$K(x, x') = \exp\left\{ -\frac{\|x - x'\|^2}{2\sigma^2} \right\}$$

Reproducing kernel Hilbert space**???**

Chapter 3. Ensemble Learning

3.1 Boosting(Meta Learning)

Idea Combine base classifier

1. generate
2. combine

Algorithm 1 AdaBoost

AdaBoost. Require: Input $S = \{(x_1, y_1), ..., (x_n, y_n)\}$, $y_i \in \{\pm 1\}$
Require: $\mathcal{A}$ a base learning algorithm
  Initialize $D_1(i) = \frac{1}{n},\ i \in \{1, ..., n\}$
  for $t = 1, 2, ..., T$ **do**
    Learn a base classifier $h_t(\cdot)$ using $\mathcal{A}$ with $D_t(\cdot)$ on S
    $\epsilon_t := \sum_{i=1}^{n} D_t(i) I[y_i \neq h_t(x_i)]$
    $\gamma_t := 1 - 2\epsilon_t$
    $\alpha_t := \frac{1}{2} \ln \frac{1+\gamma_t}{1-\gamma_t}$
    $z_t := \sum_i D_t(i) \exp\{-y_i \alpha_t h_t(x_i)\}$
    $D_{t+1}(i) = \frac{D_t(i) \exp\{-y_i \alpha_t h_t(x_i)\}}{z_t}$
  end for
  $f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$
  return $F(x) = \operatorname{sgn}[f(x)]$

Proposition(Homework). *AdaBoost is a greedy exponential loss with the following two properties:*

$$\alpha_t = \arg\min_{\alpha} \sum_{i=1}^{n} D_t(i) \exp\{-y_i \alpha_t h_t(x_i)\} \tag{44}$$

$$\prod_{t=1}^{T} z_t = \frac{1}{n} \sum_{i=1}^{n} \exp\left\{-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i)\right\} = \frac{1}{n} \sum_i \exp\{-y_i f(x_i)\} \tag{45}$$

Proof. We first prove (44). Denote $F_t(\alpha) = \sum_{i=1}^{n} D_t(i) \exp\{-y_i \alpha_t h_t(x_i)\}$, when $F_t(\alpha)$ reaches its minimum,

$$\frac{\partial F_t(\alpha)}{\partial \alpha} = \sum_{i=1}^{n} -y_i h_t(x_i) D_t(i) \exp\{-y_i \alpha_t h_t(x_i)\} = 0$$

$$\implies \sum_{i=1}^{n} I[y_i \neq h_t(x_i)] D_t(i) \exp\{I[y_i \neq h_t(x_i)]\alpha_t\} = \sum_{i=1}^{n} I[y_i = h_t(x_i)] D_t(i) \exp\{-I[y_i = h_t(x_i)]\alpha_t\}$$

Given the fact that

$$\sum_{i=1}^{n} D_t(i) I[y_i \neq h_t(x_i)] + \sum_{i=1}^{n} D_t(i) I[y_i = h_t(x_i)] = \sum_{i=1}^{n} D_t(i) = 1$$

We have

$$\epsilon_t \exp\{\alpha_t\} = (1 - \epsilon_t) \exp\{-\alpha_t\}$$

Therefore,

$$\alpha_t = \frac{1}{2} \ln \frac{1 - \epsilon_t}{\epsilon_t} = \frac{1}{2} \ln \frac{1 + \gamma_t}{1 - \gamma_t}$$

We then prove (45). Given that

$$z_t := \sum_{i=1}^{n} D_t(i) \exp\{-y_i \alpha_t h_t(x_i)\}$$
$$D_t(i) := \frac{D_{t-1}(i) \exp\{-y_i \alpha_{t-1} h_{t-1}(x_i)\}}{z_{t-1}}$$

Obviously,

$$\begin{aligned} z_T &= \sum_{i=1}^{n} D_T(i) \exp\{-y_i \alpha_T h_T(x_i)\} \\ &= \sum_{i=1}^{n} \frac{D_{T-1}(i) \exp\{-y_i \alpha_{T-1} h_{T-1}(x_i)\}}{z_{T-1}} \exp\{-y_i \alpha_T h_T(x_i)\} \\ &..... \\ &= \sum_{i=1}^{n} D_1(i) \frac{\exp\left\{-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i)\right\}}{\prod_{t=1}^{T-1} z_t} \end{aligned}$$

Therefore, we have our conclusion.

$$\prod_{t=1}^{T} z_t = \frac{1}{n} \sum_{i=1}^{n} \exp\left\{-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i)\right\} = \frac{1}{n} \sum_{i=1}^{n} \exp\{-y_i f(x_i)\}$$

□

Note that $\exp\{-y_i f(x_i)\}$ is also a surrogate loss of 0-1 loss, differentiable as well as convex.

Proposition(Homework). *Suppose* $\gamma_t \geq \gamma \geq 0$ *for* $t \in [1, ..., T]$. *Then*

$$\begin{aligned} P_s\left(yf(x) \leq 0\right) &= \frac{1}{n} I\left[y_i f(x_i) \leq 0\right] \\ &\leq \frac{1}{n} \sum_{i=1}^{n} \exp\{-y_i f(x_i)\} \\ &\leq (1-\gamma^2)^{\frac{T}{2}} \end{aligned} \tag{46}$$

Proof. We first prove that

$$\frac{1}{n} I\left[y_i f(x_i) \leq 0\right] \leq \frac{1}{n} \sum_{i=1}^{n} \exp\{-y_i f(x_i)\}$$

However, this is quite obvious given the fact that $I\left[y_i f(x_i) \leq 0\right] \leq \exp\{-y_i f(x_i)\}$ everywhere(surrogate loss of 0-1 loss).

We then prove that

$$\frac{1}{n} \sum_{i=1}^{n} \exp\{-y_i f(x_i)\} \leq (1-\gamma^2)^{\frac{T}{2}}$$

We already know that

$$\frac{1}{n} \sum_{i=1}^{n} \exp\{-y_i f(x_i)\} = \prod_{t=1}^{T} z_t$$

Therefore,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\exp\left\{-y_i f(x_i)\right\} &= \prod_{t=1}^{T}\left\{\sum_{i=1}^{n}\exp\left\{\alpha_t\right\}D_t(i)I[y_i \neq h_t(x_i) + \exp\left\{-\alpha_t\right\}D_t(i)I[y_i = h_t(x_i)]\right\} \\
&= \prod_{t=1}^{T}\left\{\exp\{\alpha_t\}\epsilon_t + \exp\{-\alpha_t\}(1-\epsilon_t)\right\} \\
&= \prod_{t=1}^{T}\left\{\sqrt{\frac{1+\gamma_t}{1-\gamma_t}}\frac{1-\gamma_t}{2} + \sqrt{\frac{1-\gamma_t}{1+\gamma_t}}\frac{1+\gamma_t}{2}\right\} \\
&= \prod_{t=1}^{T}\sqrt{1-\gamma_t^2} \\
&\leq (1-\gamma^2)^{\frac{T}{2}}
\end{aligned}$$

□

Proposition(Homework). *Calculate the following function*

$$\sum_{i=1}^{n} D_{t+1}(i)I[y_i \neq h_t(x_i)] \tag{47}$$

Proof.

$$\begin{aligned}
\sum_{i=1}^{n} D_{t+1}(i)I[y_i \neq h_t(x_i)] &= \sum_{i=1}^{n}\frac{D_t(i)\exp\left\{-y_i\alpha_t h_t(x_i)\right\}}{\sum_i D_t(i)\exp\left\{-y_i\alpha_t h_t(x_i)\right\}}I[y_i \neq h_t(x_i)] \\
&= \frac{\sum_{i=1}^{n}\exp\left\{\alpha_t\right\}D_t(i)I[y_i \neq h_t(x_i)]}{\sum_{i=1}^{n}\exp\left\{\alpha_t\right\}D_t(i)I[y_i \neq h_t(x_i) + \exp\left\{-\alpha_t\right\}D_t(i)I[y_i = h_t(x_i)]} \\
&= \frac{\exp\left\{\alpha_t\right\}\epsilon_t}{\exp\left\{\alpha_t\right\}\epsilon_t + \exp\left\{-\alpha_t\right\}(1-\epsilon_t)}
\end{aligned}$$

Given that

$$\alpha_t = \frac{1}{2}\ln\frac{1+\gamma_t}{1-\gamma_t} = \frac{1}{2}\ln\frac{1-\epsilon_t}{\epsilon_t}$$

We have

$$\begin{aligned}
\sum_{i=1}^{n} D_{t+1}(i)I[y_i \neq h_t(x_i)] &= \frac{\frac{1-\epsilon_t}{\epsilon_t}\epsilon_t}{\frac{1-\epsilon_t}{\epsilon_t}\epsilon_t + (1-\epsilon_t)} \\
&= \frac{1}{2}
\end{aligned}$$

□

Note that

$$\begin{aligned}
f(x) &= \sum_{t=1}^{T}\alpha_t h_t(x) \\
\tilde{f}(x) &= \frac{\sum_{t=1}^{T}\alpha_t h_t(x)}{\sum_{t=1}^{T}\alpha_t}
\end{aligned}$$

which is a convex combination of $h_t(x)$ and $y\tilde{f}(x) \in [-1, 1]$. We can see this as a margin. In SVM margin represents Euclidean distance yet here margin denotes confidence.

Here we can see AdaBoost as multiplicative weight updating which greedily optimize its exponential loss. Exponential loss $= \frac{1}{n}\sum_{i=1}^{n} exp\{-y_i f(x_i)\}$. And,

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$
$$F(x) = \text{sign}[f(x)]$$

For $F \in CH_\tau(\mathcal{H})$, Obviously VC dimension grows at leat linearly by the sample size, $VC[CH_\tau(\mathcal{H})] > n$ so when the sample size is large the VC-dimension is too large. By practice, even when the training error is 0 when we continue to add samples the test error will still goes down, no overfitting problems.

3.2 Margin Theory for Boosting

Function:

$$yf(x) = y\sum_t \alpha_t h_t(x)$$

is another distance but not Euclid of course.

$$x \mapsto (h_1(x), ...h_T(x)), h_t(x) = \pm 1$$
$yf(x)$ is a distance for (x, y) defined by $(\alpha_1, ..., \alpha_T)$

Homework. *Prove that $yf(x) = y\sum_t \alpha_t h_t(x)$ is a distance of (x, y).*

Proof. We can rewrite the expression as:

$$\frac{\sum_{t=1}^{T} \alpha_t h_t(x)}{\sum_{t=1}^{T} \alpha_t} = \frac{< \alpha, h(x) >}{|\alpha|_1}$$

where $\alpha = (\alpha_1, ..., \alpha_T)$ and $h(x) = (h_1(x), ..., h_T(x))$. Therefore, it can be seen as the distance from $h(x)$ to the line α.

Here margin can be seen as the distance between the predict vector to the linear combination of classifiers.

□

If for most (x_i, y_i), $y_i f(x_i)$ is large then f has good generalization ability, which is called data dependent generalization.

For $CH(\mathcal{H})$, key idea: Approximation, to find a set that has a small VC-Dimension and is close to convex τ, that is: $CH_N(\mathcal{H}) \approx CH(\mathcal{H})$. Notice that: $CH(\mathcal{H}) = \sum \alpha_t h_t, \alpha_t \geq 0, \sum \alpha_t = 1$. Then we have:

$$CH_N(\mathcal{H}) = \frac{1}{N}\sum_{i=1}^{N} h_{t_i}, h_{t_i} \in \mathcal{H}$$

For given x, $\forall f \in CH(\mathcal{H}), \exists g \in CH_N(\mathcal{H})$,

$$f(x) \approx g(x)$$

Theorem. *Assume $|\mathcal{H}| < \infty$, then with probability $1 - d$ over the following inequalities holds simultaneously for all $f \in CH(\mathcal{H})$ and all $\theta \in (0, 1]$*

$$P_D(yf(x) \leq 0) \leq P_S(yf(x) \leq \theta) + O(\frac{1}{\sqrt{n}}(\frac{log(n)log|\mathcal{H}|}{\theta^2} + log\frac{1}{\delta})^{\frac{1}{2}}) \tag{48}$$

which means that if for most data the margin is small then this classifier is good.

Proof. There are 3 steps:

1. We want to show that $f \in CH(\mathcal{H}), \text{w.h.p.} g \in CH_N(\mathcal{H})$

$$f(x) \approx g(x)$$

which suffices to prove

$$P_D(yf(x) \leq 0) \leq P_D(yg(x) \leq \frac{\theta}{2}) + \text{small}$$

2. Then we can show that:

$$P_D(yg(x) \leq \frac{\theta}{2}) \leq P_S(yg(x) \leq \frac{\theta}{2}) + \text{Complexity of } CH_N(\mathcal{H})$$

3.

$$P_S(yg(x) \leq \frac{\theta}{2}) \leq P_S(yf(x) \leq \theta) + \text{small}$$

Rob Schapire in Microsoft, 1990, The strength of Weak Learnable. Take the prove in xitike □

3.3 Bagging

Bootstrap. *Given dataset $D = \{x_1, ..., x_n\}$, draw with replcement, we can get many dadaset with the same sample size.* $\{x_1^1, ..., x_n^1\}, \{x_1^2, ..., x_n^2\}, ..., \{x_1^k, ..., x_n^k\}$.

Algorithm 2 Bagging

Bagging. Require: Input $S = \{(x_1, y_1), ..., (x_n, y_n)\}$
Require: $\mathcal{A}$ a base learning algorithm
  Bootstrap on the original dataset S and get $S_1, ..., S_k$
  for $t = 1, 2, ..., k$ **do**
    Learn a base classifier $h_t(\cdot)$ using $\mathcal{A}$ on S_t
  end for
  return $F(x) = \frac{1}{k} h_t(x)$

3.4 Algorithmic Stability and Generalization

SVM and Boosting try to improve the margin. Now we try to analyse the property of a algorithm and its error. For Boosting:

$$l = \frac{1}{n} \sum_{i=1}^{n} \exp(-y_i \sum_{t=1}^{T} \alpha_t h_t(x))$$

and for SVM:

$$l = \frac{1}{n} \sum_{i=1}^{n} \left[1 - y_i(w^T x_i + b)\right] + \frac{\lambda}{2} ||w||^2$$

Denote $\mathcal{A}$ the learning algorithm. $S = \{(x_i, y_i)\}$ the training data set, $l(\mathcal{A}(S), z)$ the loss function, $\mathcal{A}(S)$ the result of the learning algorithm, z the test data. Then risk function is

$$R(\mathcal{A}(S)) = E_z[l(\mathcal{A}(S), z)]$$

and empirical risk is

$$R_{\text{emp}}(\mathcal{A}(S)) = \frac{1}{n} \sum_{i=1}^{n} l(\mathcal{A}(S), z_i)$$

Definition. *A learning algorithm $\mathcal{A}$ is said to have* ***uniform stability*** β *with respect to loss l, if for* $\forall S = (z_1, ..., z_n), S^i = (z_{-i}, z_i')$,

$$\left| l\left(\mathcal{A}(S), z\right) - l\left(\mathcal{A}(S^i), z\right) \right| \leq \beta$$

Theorem. *Suppose $\mathcal{A}$ has uniform stability β with respect to loss l and $l \leq M$, then with probability $1 - \delta$,*

$$R(\mathcal{A}(S)) \leq R_{\text{emp}}(\mathcal{A}(S)) + \beta + (n\beta + M)\sqrt{\frac{2\log\frac{1}{\delta}}{n}} \tag{49}$$

Proof. Denote $f(S) = R(\mathcal{A}(S)) - R_{\text{emp}}(\mathcal{A}(S))$. The theorem is equivalent to

$$\mathbb{P}[f(S) \geq \beta + \epsilon] \leq \exp(-\frac{n\epsilon^2}{2(n\beta + M)^2})$$

$$\begin{aligned} E_S[R_{\text{emp}}(\mathcal{A}(S))] &= E_S\left[\frac{1}{n}\sum_{i=1}^{n} l(\mathcal{A}(S), z_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E_{S,z_i'}\left[l(\mathcal{A}(S^i), z_i')\right] \end{aligned}$$

also,

$$E_S[R(\mathcal{A}(S))] = E_{S,z_i'} l(\mathcal{A}(S), z_i')$$

By the definition of uniform stability, we know that

$$|l(\mathcal{A}(S), z_i') - l(\mathcal{A}(S^i), z_i')| \leq \beta$$

Therefore,

$$E_S[f(S)] \leq \beta$$

Also,

$$\begin{aligned} |f(S) - f(S^i)| &= |R(\mathcal{A}(S)) - R_{\text{emp}}(\mathcal{A}(S)) - R(\mathcal{A}(S^i)) + R_{\text{emp}}(\mathcal{A}(S^i))| \\ &\leq |R(\mathcal{A}(S)) - R(\mathcal{A}(S^i))| + \frac{1}{n}\sum_{i=1}^{n} |l(\mathcal{A}(S), z_i) - l(\mathcal{A}(S^i), z_i)| \\ &\leq \beta + \frac{n-1}{n}\beta + \frac{2M}{n} \\ &\leq 2(\beta + \frac{M}{n}) \end{aligned}$$

Then, by Mcdiarmid's Lemma,

$$\mathbb{P}[f(S) \geq E_S[f(S)] + \epsilon] \leq \exp(-\frac{n\epsilon^2}{2(n\beta + M)^2})$$

Therefore, by the fact that $E_S[f(S)] \leq \beta$

$$\mathbb{P}[f(S) \geq \beta + \epsilon] \leq \exp(-\frac{n\epsilon^2}{2(n\beta + M)^2})$$

□

3.5 Online Learning

Algorithm 3 Weighted Majority Vote

Require: Parameter $\beta \in (0,1)$
  Initialize weights $w_{1,i} = 1, i \in \{1, ..., N\}$
  for $t = 1, 2, ..., T$ **do**
    Every expert $i \in \{1, ..., N\}$ makes a prediction $\tilde{y}_{t,i}$
    Majority Vote
    if $\sum_{y_{i,t}=0} w_{t,i} \leq \sum_{y_{i,t}=1} w_{t,i}$ **then**
      $\tilde{y}_t = 1$
    else
      $\tilde{y}_t = 0$
    end if
    Observe the real value y_t
    if $\tilde{y}_t = y_t$ **then**
      $w_{t+1,i} \leftarrow w_{t,i}$
    else
      $w_{t+1,i} \leftarrow \beta w_{t,i}$ for all i such that $\tilde{y}_{t,i} \neq y_t$
    end if
  end for

Theorem. *Let* $L_T = \sum_{t=1}^{T} |\tilde{y}_t - y_t|$, $m_T^{(i)} = \sum_{t=1}^{T} |\tilde{y}_{t,i} - y_t|$, $m_T^* = \min_{i \in [1...N]} m_T^{(i)}$. *Then*

$$L_T \leq \frac{\ln \frac{1}{\beta}}{\ln \frac{2}{1+\beta}} m_T^* + \frac{\ln N}{\ln \frac{2}{1+\beta}} \tag{50}$$

Proof. **Potential Function Method**. Denote

$$W_t := \sum_{i=1}^{N} w_{t,i}$$

Now when $\tilde{y}_t \neq y_t$ (Algorithm makes wrong prediction),

$$W_{t+1} \leq \left(\frac{1+\beta}{2}\right) W_t$$

Therefore,

$$W_T \leq \left(\frac{1+\beta}{2}\right)^{L_T} N \tag{51}$$

$$\forall i,\ W_T \geq w_{T,i} = \beta^{m_T^{(i)}} \tag{52}$$

Combine (51) and (52) together, we get our conclusion. □

Algorithm 4 Randomized Weighted Majority Vote

Require: Parameter $\beta \in (\frac{1}{2}, 1)$
  Initialize weights $w_{1,i} = 1, i \in \{1, ..., N\}$
  for $t = 1, 2, ..., T$ **do**
    Every expert $i \in \{1, ..., N\}$ makes a prediction $\tilde{y}_{t,i}$
    Randomized Majority Vote: sample a prediction according to $p_{t,i} = \frac{w_{t,i}}{\sum_{i=1}^{N} w_{t,i}}$
    Observe the real value y_t
    $w_{t+1,i} \leftarrow \beta w_{t,i}$ for all i such that $\tilde{y}_{t,i} \neq y_t$
  end for

Theorem(Homework). *For the Randomized Weighted Majority Vote Algorithm, define expected loss:*

$$L_T = \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{w_{t,i}}{\sum_{j=1}^{N} w_{t,j}} |\tilde{y}_{t,i} - y_t|$$

Then $\forall \beta \in (\frac{1}{2}, 1)$ *we have*

$$L_T \leq (2-\beta) m_T^* + \frac{\ln N}{1-\beta} \tag{53}$$

Further, if T *is known, let* $\beta = 1 - \sqrt{\frac{\ln N}{T}}$*, then we have*

$$L_T \leq m_T^* + 2\sqrt{T * \ln N} \tag{54}$$

which is usually written as:

$$\frac{L_T}{T} \leq \frac{m_T^*}{T} + O\left(\sqrt{\frac{\ln N}{T}}\right) \tag{55}$$

Proof. Define $W_t = \sum_{j=1}^{N} w_{t,j}$. Therefore, $W_1 = N$ and

$$L_T = \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{w_{t,i}}{W_t} |\tilde{y}_{t,i} - y_t|$$

Denote

$$l_t = \frac{\sum_{\tilde{y}_{t,i} \neq y_t} w_{t,i}}{W_t}$$

Therefore,

$$\begin{aligned} W_{t+1} &= (1 - l_t) W_t + \beta l_t W_t \\ &= W_t \left(1 - l_t + \beta l_t\right) \end{aligned}$$

And we have

$$\begin{aligned} W_{\text{final}} &= W_1 \prod_{t=1}^{t} (1 - (1-\beta) l_t) \\ &\leq N \prod_{t=1}^{T} \exp\left\{-(1-\beta) l_t\right\} \\ &= N \exp\left\{-(1-\beta) \sum_{t=1}^{T} l_t\right\} \end{aligned}$$

Note that $\forall i$,

$$W_{\text{final}} \geq w_{T,i} = \beta^{m_T^{(i)}}$$

Therefore, $W_{\text{final}} \geq \beta^{m_T^*}$

$$N \exp\left\{-(1-\beta) \sum_{t=1}^{T} l_t\right\} \geq \beta^{m_T^*}$$

Therefore,

$$\sum_{t=1}^{T} l_t \leq \frac{\ln \frac{1}{\beta}}{1-\beta} m_T^* + \frac{\ln N}{1-\beta}$$

Note that $L_T = \sum_{t=1}^{T} l_t$ and when $\beta \in (\frac{1}{2}, 1)$

$$\frac{\ln \frac{1}{\beta}}{1-\beta} \leq 2 - \beta$$

So we have the conclusion

$$L_T \leq (2-\beta) m_T^* + \frac{\ln N}{1-\beta}$$

□

Note: for online learning algorithms, T is usually unknown. We instead use doubling trick to solve it: we can guess a T first, then if we want to continue then we double T. It is easy to prove that with this trick we can get a similar result.

Von Neumann Minmax Theorem.

$$\min_p \max_q p^T M q = \max_q \min_p p^T M q \tag{56}$$

Proof. Repeated Game, zero-sum matrix game.

Consider each row an expert, row player combines experts and chooses p_t. Consider column player the adversarial, chooses q_t. At time t, expert i suffers loss $(Mq_t)_i$ and row player loss $p_t^T M q_t$

□

Algorithm 5 Von Neumann Minmax

Require: Parameter $\beta \in (\frac{1}{2}, 1)$
Initialize $p_1 = (\frac{1}{N}, ..., \frac{1}{N})$
We want to make $q_t = \operatorname{argmax}_q p_t^T M q$
for $t = 1, 2, ..., T$ **do**
1. Row player chooses p_t
2. Column player chooses q_t which may depend on p_t
3. Row player observes the loss of each row (Mq_t)
4. $p_{t+1}^{(i)} = \frac{p_t^{(i)} \beta (Mq_t)^{(i)}}{z_t}$ where z_t is a normalization factor *s.t.* $\sum p_{t+1}^{(i)} = 1$

end for

Theorem. *Assume $M_{ij} \in [0, 1]$, then*

$$\sum_{t=1}^{T} p_t^T M q_t \leq (2-\beta) \min_i (\sum_{t=1}^{T} M q_t)_i + \frac{\ln N}{1-\beta} \tag{57}$$

$$\frac{1}{T} \sum_{t=1}^{T} p_t^T M q_t \leq \frac{1}{T} \min_i (\sum_{t=1}^{T} M q_t)_i + O\left(\sqrt{\frac{\ln N}{T}}\right) \tag{58}$$

where the left side is $\min_p \max_q p^T M q$ while the one on the right side is another

Algorithm 6 Multiplicative Weight Updating

Require: pmf $x = (x(1), ..., x(n))$ $x_i \geq 0, \sum x_i = 1$ where x is unknown to the learner
Require: Parameters δ, ϵ
  Initialize $x_0 = (\frac{1}{N}, ..., \frac{1}{N})$
  for $t = 1, 2, ..., T$ **do**
    Adversary chooses $f_t \in \{0, 1\}^N$, calculates $\langle f_t, x\rangle = \sum_{i=1}^N f_t(i)x_i$
    Adversary releases f_t and $\langle f_t, x\rangle$ to the learner
    if $\langle f_t, x\rangle - \langle f_t, x_t\rangle > \delta$ **then**
      if $f_t(i) = 1$ **then**
        $x_t(i) \leftarrow (1 + \epsilon)x_{t-1}(i)$
      end if
      Normalize x_t
    else if $\langle f_t, x\rangle - \langle f_t, x_t\rangle < -\delta$ **then**
      if $f_t(i) = 0$ **then**
        $x_t(i) \leftarrow (1 + \epsilon)x_{t-1}(i)$
      end if
      Normalize x_t
    end if
  end for

Theorem(Homework). *For every choice of $f_1, f_2, ...$ the algorithm goes into these two situations for at most $\frac{2 \log N}{\epsilon\delta}$ $(0 < \epsilon < \delta)$*

Proof. Use Kullback–Leibler divergence $D(x\|x_t) = \sum_{i=1}^N x(i) \log \frac{x(i)}{x_t(i)}$ as potential function. Therefore, when $\langle f_t, x\rangle - \langle f_t, x_t\rangle > \delta$

$$\begin{aligned}
D(x\|x_{t+1}) &= \sum_{i=1}^N x(i) \log \frac{x(i)}{x_{t+1}(i)} \\
&= D(x\|x_t) + \sum_{i=1}^N x(i) \log \frac{x_t(i)}{x_{t+1}(i)} \\
&= D(x\|x_t) + \sum_{f_t(i)=1} x(i) \log \frac{x_t(i)}{\frac{(1+\epsilon)x_t(i)}{1+\epsilon\langle f_t, x_t\rangle}} \\
&= D(x\|x_t) + \log(1 + \epsilon\langle f_t, x_t\rangle) - \langle f_t, x\rangle \log(1 + \epsilon) \\
&\leq D(x\|x_t) + \log(1 + \epsilon\langle f_t, x_t\rangle) - (\langle f_t, x_t\rangle + \delta) \log(1 + \epsilon) \\
&\leq D(x\|x_t) - \delta \log(1 + \epsilon)
\end{aligned}$$

Given the ground truth that

$$(1 + \epsilon)^k \geq 1 + \epsilon k$$

Also, when $\langle f_t, x\rangle - \langle f_t, x_t\rangle < -\delta$,

$$\begin{aligned}
D(x\|x_{t+1}) &= \sum_{i=1}^N x(i) \log \frac{x(i)}{x_{t+1}(i)} \\
&= D(x\|x_t) + \log(1 + \epsilon\langle f_t, x_t\rangle) - \langle f_t, x\rangle \log(1 + \epsilon) \\
&\leq D(x\|x_t) + \log\left[1 + \epsilon(\langle f_t, x\rangle + \delta)\right] - \langle f_t, x\rangle \log(1 + \epsilon) \\
&\leq D(x\|x_t) - \delta \log(1 + \epsilon)
\end{aligned}$$

Therefore, denote n_t as the update times till time t, which is the times that $|\langle f_t, x\rangle - \langle f_t, x_t\rangle| > \delta$ till time t.

Then,

$$\begin{aligned} 0 &\leq D(x\|x_T) \\ &\leq D(x\|x_1) - n_T\delta\log(1+\epsilon) \\ &\leq \log N - n_T\delta\log(1+\epsilon) \end{aligned}$$

Therefore,

$$n_T \leq \frac{\log N}{\delta\log(1+\epsilon)} \leq \frac{2\log N}{\epsilon\delta}$$

Given the fact that $0 < \epsilon < \delta < 1$ and $\log(1+\epsilon) \geq \frac{\epsilon}{2}$ when $\epsilon \in (0,1)$. □

Chapter 4. PAC Learning

4.1 Bayesian and Frequentist

Bayesian: learn a distribution of classifier so it is a stochastic classifier. Then the function of it can be valued by its expectation.

Homework. *Let Q be a distribution of classifiers, stochastic classifier,* $\text{error}_D(f_Q)$ *and voting classifier,* $\text{error}_D(v_Q)$*. Find the relationship between these two errors.*

Proof. We know that if $f_v(x) \neq y$, then

$$\mathbb{P}_Q(f(x) \neq y) > 1/2$$

Let $f(x, y) = \mathbb{P}_Q(f(x) \neq y)$,

$$\text{error}_D(v_Q) = \mathbb{P}_{(x,y)}(f_v(x) \neq y) = \mathbb{E}_{(x,y)}(I[f(x, y) > \frac{1}{2}])$$

also,

$$\begin{aligned}
\text{error}_D(f_Q) &= \mathbb{P}_{Q,(x,y)}(f(x) \neq y) \\
&= \mathbb{E}_{(x,y)}(f(x, y)) \\
&= \frac{1}{2}\mathbb{E}_{(x,y)}(2f(x, y)) \\
&\geq \frac{1}{2}\mathbb{E}_{(x,y)}(I[f(x, y) > 0.5]) \\
&= \frac{1}{2}\text{error}_D(v_Q)
\end{aligned}$$

Therefore,

$$\text{error}_D(f_Q) \geq \frac{1}{2}\text{error}_D(v_Q)$$

□

As VC theory is some kind of Frequentist theory, so now by the view of Bayesion we want a uniform convergence, but it is a stochastic classifier, we should uniform all the distribution of the stochastic classifiers. For fixed prior distribution $\mathcal{P}$ (w.h.p over random draw of training data) for all distribution $\mathcal{Q}$. Recall that for VC-theory or Margin-theory we want to get:

$$\text{For all classifier } f \in \mathcal{H} \text{ , } \text{err}_D(f) \leq \text{err}_S(f) + \text{Complexity}$$

So for this case it should be:

$$\text{For all distribution } Q \text{ , } \text{err}_D(Q) \leq \text{err}_S(Q) + D(Q||P)$$

4.2 PAC Bayes Theorem

For any fixed prior distribution$\mathcal{P}$, with probability $1 - \delta$, we have

$$\text{err}_D(Q) \leq \text{err}_S(Q) + \sqrt{\frac{D(Q||P) + \log(\frac{3}{\delta})}{n}} \tag{59}$$

Holds for all α simultaneously.

Lemma(Homework). *For any functional f of classifier $\hbar$*

$$E_{\hbar\sim Q}[f(\hbar)] \leq \ln E_{\hbar\sim P}[e^{f(\hbar)}] + D(Q||P) \tag{60}$$

Proof.

$$\begin{aligned}
E_{\hbar\sim Q}[f(\hbar)] &= E_{\hbar\sim Q}[\ln e^{f(\hbar)}] \\
&= E_{\hbar\sim Q}\left[\ln \frac{dP(x)}{dQ(x)} e^{f(\hbar)} + \ln \frac{dQ(x)}{dP(x)}\right] \\
&= E_{\hbar\sim Q}\left[\ln \frac{dP(x)}{dQ(x)} e^{f(\hbar)}\right] + D(Q\|P) \\
&\leq \ln E_{\hbar\sim Q}\left[\frac{dP(x)}{dQ(x)} e^{f(\hbar)}\right] + D(Q\|P) \\
&= \ln E_{\hbar\sim P}[e^{f(\hbar)}] + D(Q||P)
\end{aligned}$$

□

Lemma(Homework). *Let $f(\hbar) = n[\mathrm{err}_D(\hbar) - \mathrm{err}_S(\hbar)]^2$. Then*

$$\mathbb{P}\left[\mathbb{E}_{\hbar\sim P} \exp\{f(\hbar)\} \geq \frac{3}{\delta}\right] \leq \delta \tag{61}$$

Prove that for all fixed $\hbar$,

$$\mathbb{P}\left(|\mathrm{err}_D(\hbar) - \mathrm{err}_S(\hbar)| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2} \tag{62}$$

Proof. According to chernoff bound,

$$\mathbb{P}\left[|\mathrm{err}_D(h) - \mathrm{err}_S(h)| \geq \epsilon\right] \leq 2e^{-2n\epsilon^2}$$

Therefore,

$$\mathbb{P}[e^{f(h)} \geq t] = \mathbb{P}\left[|\mathrm{err}_D(h) - \mathrm{err}_S(h)| \geq \sqrt{\frac{\ln t}{n}}\right] \leq \frac{2}{t^2}$$

$$\begin{aligned}
\therefore E_{h\sim P} \exp\{f(h)\} &= \int_0^{+\infty} \mathbb{P}[e^{f(h)} \geq t]dt \\
&\leq \int_0^1 1dt + \int_1^{+\infty} \frac{2}{t^2}dt \\
&= 3
\end{aligned}$$

By Markov's inequality, the result holds. □

Lemma. *Improved PAC Bayes Thm:*

$$\mathbb{P}\left(D_B(\mathrm{err}_S(Q)||\mathrm{err}_D(Q)) \geq \delta\right) \geq \frac{D(Q||P) + \log\frac{n+1}{\delta}}{n} \tag{63}$$

4.3 PAC-Bayes implies Margin theory for SVM

We have a distribution Q and it derives a voting classifier and a stochastic classifier, but the error of the first one cannot exceed the 2-times of the second one. Because if the voting classifier is wrong then the stochastic is only half right.

Assume the linear classifier goes from the origin (because we can add 1 more dimension to make it), then the unit normal vector of the classifier is uniformly distributed on the surface of unit ball centering at origin, this is our prior distribution.

But it is hard to calculate so we use $P \sim \mathcal{N}(0, I)$ instead.???

Gaussian distribution for posteriors distribution is easy to calculate the KL-distance, thus we suppose $Q \sim \mathcal{N}(\mu, I)$. Actually $Q \sim \mathcal{N}(u * \overrightarrow{w}, I)$, where $u \in \mathcal{R}$.

Theorem. *Assume $Q \sim \mathcal{N}(u * \overrightarrow{w}, I)$, where $u \in \mathcal{R}$, then*

$$\mathrm{err}_D(Q) \leq \mathrm{err}_S(Q) + \sqrt{\frac{D(Q||P) + \log\frac{3}{\delta}}{n}} \tag{64}$$

Proof. First we have:

$$D(Q||P) = D(\mathcal{N}(u * \overrightarrow{w}, I)||\mathcal{N}(0, I)) = \frac{u^2}{2}$$

Here only 1-dimension is needed to integral and all others are equivalent.

Now consider $\mathrm{err}_S(Q)$ where $Q \sim \mathcal{N}(u * \overrightarrow{w}, I)$. For any given (x, y) the probability for wrong classified is:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt$$

where $t = u * y * \frac{w*x}{||x||}$. So we have:

$$\mathrm{err}_S(Q) = \frac{1}{n} \sum \Phi(u * y_i * \frac{w * x_i}{||x_i||})$$

where $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt$. Then when $u \to 0$ obviously the margin is small. □

Chapter 5. Other Algorithms

5.1 K-Clustering

Criterion for k-cluster, learn $c = (c_1, ...c_k)$ (cluster center) according to data $x_1, ..., x_n \in R^d$ and $x_i \xrightarrow{\text{nearest}} c_{j(i)}$. Objective:

$$\min_{c_1,...,c_k} \sum_{i=1}^{n} ||x_i - c_{j(i)}||^2$$

Algorithm 7 K-Means

Require: k, the number of different centroids
  Initialize k different centroids $c = \{c_1, ..., c_k\}$ to k data points
  while some conditions **do**
    Assign each training example to cluster, with centroid c_j
    Each centroid c_j is updated to the mean of all training examples x_i assigned to cluster j
  end while

Algorithm 8 K-Means ++

Require: k, the number of different centroids
  Initialize k different centroids $c = \{c_1, ..., c_k\}$ to k data points
  while some conditions **do**
    1. $\phi_{\text{OPT}} = \sum_{i=1}^{N} ||x_i - c_{j(i)}||^2$
    2. $\phi_{\text{K-Means++}} =$??? s.t. $E[\phi_{\text{K-Means++}}] \leq 8 (\log k + 2) \phi_{\text{OPT}}$
    3. $c_i \sim \frac{D(x)^2}{\sum_x D(x)^2}$ where $D(x) = ||x - C(x)||$ and $C(x) = \arg\min ||x - c_j||^2, j < i$
  end while

5.2 Reinforcement Learning

Definition. ***Markov Decision Process*** *S(State), P(Probability), A(Action), R(Reward). But actually it should be written as* $P^S_{S,A}$, $R_{S,A}$ *or Transition prob* $P(S_{t+1}|S_t, a_t)$ *and Reward* $R_{t+1} = E[R(S_t, a_t)]$

Our goal is to maximize long-term reward. At every time t reward, $\gamma \in (0, 1]$, then

$$\max G_t := R_{t+1} + \gamma R_{t+2} + ...$$

Policy $S \mapsto a$ and $\pi : S \mapsto A$, Given policy π, value function

$$v_\pi(s) = E[R_{t+1} + \gamma R_{t+1} + ...]$$

Action value function

$$q_\pi(s, a) := E[R_{t+1} + \gamma R_{t+2} + ...|S_t = s, A_t = a]$$

At time t, take action a, then follow π at time $t + 1$ and so on.

Definition. ***Bellman Equation****, Given policy* π*,*

$$v_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) v_\pi(s')$$

then by vector and matrix it is

$$v_\pi = R^{(\pi)} + \gamma P^{(\pi)} v_\pi$$

Now denote

$$\phi_\pi(v) := R^{(\pi)} + \gamma P^{(\pi)} v$$

which is called ***Bellman Expectation Operator****.*

Theorem. *For $\gamma \in (0,1)$, define*

$$d(v, v') = \max_{s \in S} |v(s) - v'(s)|$$

then Bellman Expectation Operator is a Contraction mapping. So by iteration we can simply calculate $v_\pi(s)$ by Bellman Operator. Also,

$$q_\pi(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a) + q_\pi(s', \pi(s'))$$

so it is an evaluation for policy π

Appendix A. Term Project

Term Project 1. *Denote Data space* $= [\mathcal{N}]$, $\mathcal{F} \subseteq \{0,1\}^{\mathcal{N}}$, *VC dmension* $VC(\mathcal{F}) = d$ *if* $\exists i_1, ..., i_d$, $\mathcal{F}$*'s projection onto* $i_1, ..., i_d$ *contains* $\{0,1\}^d$ *and* $\forall i_1, ..., i_{d+1}$*'s projection onto* $i_1, ..., i_{d+1} \neq \{0,1\}^{d+1}$

$f \in \mathcal{F}$, $\exists X_f \subseteq X$, $f|_{X_f} \neq f'|_{X_f}, \forall f' \in \mathcal{F}$

Teaching dimension of f *is* $TD(f, \mathcal{F}) = \min |X_f|$, *best case teaching dimension is* $TD_{\min}(\mathcal{F}) = \min_{f \in \mathcal{F}} TD(f, \mathcal{F})$.

Let $\mathcal{F}_0 = \mathcal{F}$ *and assume* $f_1, s.t. TD_{\min}(\mathcal{F}) = TD(f_1, \mathcal{F})$, $\mathcal{F}_1 = \mathcal{F}_0 \backslash \{f_1\}$ *and so on.*

Define Recursive Teaching Dimension: $RTD(\mathcal{F}) = max_t TD_{min}(\mathcal{F}_t)$, *then* $RTD(\mathcal{F}) = O(VC(\mathcal{F})^2)$

Term Project 2. *Fundamental background: Neural network in our brain is not too deep and calculation efficiency is far better, and is highlight distributed instead of central control system. So I want to compare human brain with computer, although we may reach the complexity of brain but our efficiency and control system is still not comparative to brain. So there are some advance system or algorithm in brain we can explore.*

We have many concepts in our mind, whether concrete or abstract, and how these concepts performs is still a mystery. Then how to put concepts into our algorithm is an interesting question. Now deep learning performs well because it is quite the same way human brain used to deal with imagines. But to go further we need to explore concepts.(Les Valiant, The Circuit of the Mind).Concepts are connected, and can be classified into deep concepts and surface concepts and so on.

Hinton, Capsule. Reference papers:
① Dynamic Routing Between Capsules. NIPS'17.
② Matrix Capsules with EM Routing, ICLR'18 submission.

Our work is to assume that concepts is formed by the connection by neural networks, and then explore how to put concepts into our algorithm. But our methods are not limited on these two papers, so we can design new algorithms for learning network with capsules. Furthermore there work are focused on simple training data, so we can find experimental results at bench mark datasets.

Term Project 3. *By hardware human brain is far stronger than any computer but now or in the near future the size will be comparable, at least the speed of computer is faster. At the same time the performance of our computer is far below the human brain performance. So may be human brain is not work by concurrent algorithm but by a global control system and worked like distributed computation, so we should explore it.*

1. Decoupled Neural Interfaces using Synthetic Gradients

2. Understanding Synthetic Gradient and Decoupled Neural Interfaces.

Appendix B. Reading List

Chapter 3 Reading List

- Schapire, R. E. (1990). The strength of weak learnability. Machine learning, 5(2), 197-227
- Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. The annals of statistics, 26(5), 1651-1686
- Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. Machine learning, 37(3), 297-336
- Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. Journal of Machine Learning Research, 2(Mar), 499-526
- Littlestone, N., & Warmuth, M. K. (1994). The weighted majority algorithm. Information and computation, 108(2), 212-261
- Arora, S., Hazan, E., & Kale, S. (2012). The Multiplicative Weights Update Method: a Meta-Algorithm and Applications. Theory of Computing, 8(1), 121-164

Chapter 4 Reading List

- McAllester, D. (2003). Simplified PAC-Bayesian margin bounds. Lecture notes in computer science, 203-215

Term Project 1 Reading List

- Doliwa, T., Fan, G., Simon, H. U., & Zilles, S. (2014). Recursive teaching dimension, VC-dimension and sample compression. Journal of Machine Learning Research, 15(1), 3107-3131
- Simon, H. U., & Zilles, S. (2015, June). Open problem: Recursive teaching dimension versus VC dimension. In Conference on Learning Theory (pp. 1770-1772)
- Chen, X., Cheng, Y., & Tang, B. (2016). On the recursive teaching dimension of vc classes. In Advances in Neural Information Processing Systems (pp. 2164-2171)
- Hu, L., Wu, R., Li, T., & Wang, L. (2017). Quadratic Upper Bound for Recursive Teaching Dimension of Finite VC Classes. arXiv preprint arXiv:1702.05677

Term Project 2 Reading List

- Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic Routing Between Capsules. arXiv preprint arXiv:1710.09829
- Matrix Capsules with Em Routing

Appendix C. Deep Learning Reading List

a. Generalization/Understanding

- Mou, W., Wang, L., Zhai, X., & Zheng, K. (2017). Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints. arXiv preprint arXiv:1707.05947
- Telgarsky, M. (2016). Benefits of depth in neural networks. arXiv preprint arXiv:1602.04485
- Montufar, G. F., Pascanu, R., Cho, K., & Bengio, Y. (2014). On the number of linear regions of deep neural networks. In Advances in neural information processing systems (pp. 2924-2932)
- Eldan, R., & Shamir, O. (2016, June). The power of depth for feedforward neural networks. In Conference on Learning Theory (pp. 907-940)
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199
- Hardt, M., Recht, B., & Singer, Y. (2015). Train faster, generalize better: Stability of stochastic gradient descent. arXiv preprint arXiv:1509.01240
- Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. arXiv preprint arXiv:1703.04730
- Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530
- Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. In Advances in neural information processing systems (pp. 3320-3328)
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572

b. Computer Vision(CNN)

- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105)
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9)
- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778)

c. Training Techniques

- Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
- Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International Conference on Machine Learning (pp. 448-456)

- Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. Journal of machine learning research, 15(1), 1929-1958
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580
- Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., & Bengio, Y. (2013). Maxout networks. arXiv preprint arXiv:1302.4389

d. Reinforcement Learning

- Van Hasselt, H., Guez, A., & Silver, D. (2016, February). Deep Reinforcement Learning with Double Q-Learning. In AAAI (pp. 2094-2100)
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533

e. Generative Models

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. In Advances in Neural Information Processing Systems (pp. 2234-2242)
- Arjovsky, M., & Bottou, L. (2017). Towards principled methods for training generative adversarial networks. arXiv preprint arXiv:1701.04862
- Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein gan. arXiv preprint arXiv:1701.07875
- Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114

f. Natural Language Process(RNN)

- Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., ... & Torr, P. H. (2015). Conditional random fields as recurrent neural networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1529-1537)
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112)
- Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. In Advances in Neural Information Processing Systems (pp. 1693-1701)
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780