Statistical and Computational Foundations of Machine Learning

Lecture 02

Dávid Pál

New York University

February 5, 2023

Last modified: 2023/01/07 10:43:37 -05'00'

Overview

Reading: Chapters 2,3,4 and Appendix B

Elementary tools

Union bound

Lemma (Union bound)

Let $A_1, A_2, \ldots, A_n$ be probability events. Then,

$$\Pr[A_1 \cup A_2 \cup \cdots \cup A_n] \leq \Pr[A_1] + \Pr[A_2] + \cdots + \Pr[A_n] \,.$$

- For $\Pr[\cdot]$ to make sense, we implicitly assume there is an underlying probability space $(\mathcal{X}, \mathcal{F}, P)$.
- $\Pr[\cdot]$ is another notation for $P(\cdot)$. This notation $\Pr[\cdot]$ used when $(\mathcal{X}, \mathcal{F}, P)$ is only implicitly defined.
- A *(probability) event* A is just fancy another word for $A \in \mathcal{F}$.

Union bound

Proof.

If $A \subseteq B$ then

$$\Pr[A] \leq \Pr[A] + \Pr[B \setminus A] = \Pr[B] \,.$$

If $n = 1$, the inequality is trivially true. If $n = 2$,

$$\begin{aligned}\Pr[A_1 \cup A_2] &= \Pr[A_1 \cup (A_2 \setminus A_1)] \\ &= \Pr[A_1] + \Pr[A_2 \setminus A_1] \leq \Pr[A_1] + \Pr[A_2] \,.\end{aligned}$$

For $n \geq 3$, using the $n = 2$ case and induction hypothesis,

$$\begin{aligned}\Pr[A_1 \cup A_2 \cup \cdots \cup A_n] &\leq \Pr[A_1 \cup A_2 \cup \cdots \cup A_{n-1}] + \Pr[A_n] \\ &\leq \Pr[A_1] + \Pr[A_2] + \cdots + \Pr[A_n] \,.\end{aligned}$$

■

Properties of expectation

Lemma (Linearity of expectation)

Let X, Y be any random variables (possibly correlated). Then,

$$\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]\,.$$

Lemma (Expectation of product)

Let X, Y be **independent** *random variables. Then,*

$$\mathbf{E}[XY] = \mathbf{E}[X] \cdot \mathbf{E}[Y]\,.$$

Lemma (Tower rule)

Let X, Y be any random variables (possibly correlated). Then,

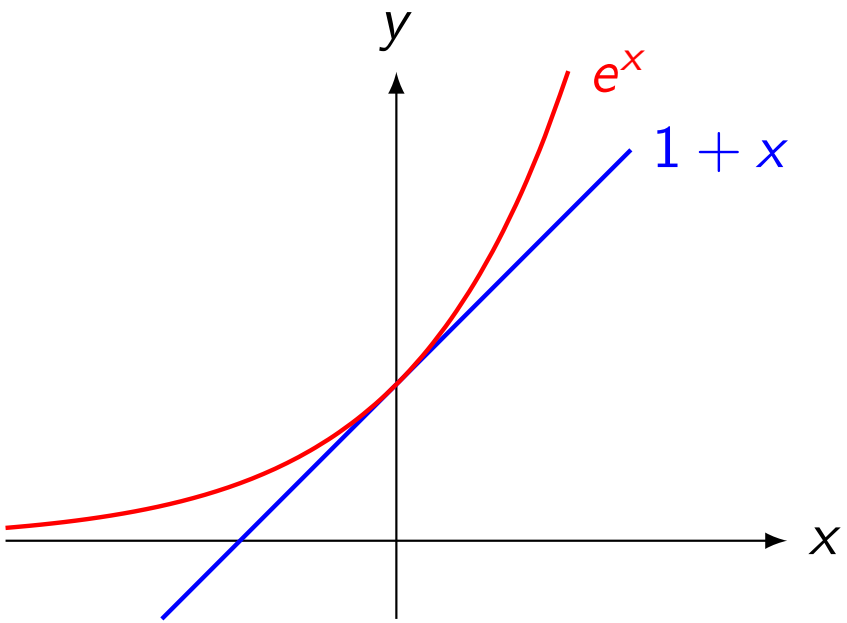
$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|Y]]\,.$$

Useful inequality

Lemma (Useful inequality)

For any $x \in \mathbb{R}$,

$$1 + x \leq e^x \,.$$



- The two sides of the inequality are close when x is close to 0.
- Very often used in the form $1 - x \leq e^{-x}$.

Empirical Risk Minimization

Empirical Risk Minimization

Definition (Empirical Risk Minimization)

Suppose that $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is a non-empty set of predictors. *Empirical Risk Minimization* (ERM) is a learning algorithm that for a labeled sample $S \in (\mathcal{X} \times \mathcal{Y})^*$ outputs

$$\widehat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} L_S(h)$$

- $\mathcal{H}$ has many different names:
  - hypothesis class, concept class, function class, model class
  - class of predictors, class of classifiers
  - set system: a classifier $h \in \mathcal{H}$ corresponds to a set $h^{-1}(1)$
- Different ERMs exist for different tie-breaking rules.

ERM can sometimes have a high generalization error

- Suppose $\mathcal{X} = [0,1]$ and $\mathcal{Y} = \{0,1\}$.
- Let D be a distribution over $\mathcal{X} \times \mathcal{Y}$ such that if $(X,Y) \sim D$ then X is uniform over $[0,1]$ and

$$Y = \begin{cases} 0 & \text{if } X < 1/2, \\ 1 & \text{if } X \geq 1/2. \end{cases}$$

- Note that Bayes optimal classifier h^* has error $L_D(h^*) = 0$.
- Let $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$ and $A : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$ be an ERM algorithm over $\mathcal{H}$ such that

$$\widehat{h} = A(S)$$

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } (x,1) \in S, \\ 0 & \text{otherwise.} \end{cases}$$

  That is, algorithm A "memorizes" the positive examples.
- With probability one, $L_D(\widehat{h}) = 1/2$.

ERM can sometimes have a low generalization error

Theorem (ERM over finite hypothesis classes)

Let $\mathcal{X}$ be arbitrary and $\mathcal{Y} = \{0, 1\}$. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be finite. Let $h^* \in \mathcal{H}$. Let D be a distribution over $\mathcal{X} \times \mathcal{Y}$ such that if $(X, Y) \sim D$ then $Y = h^*(X)$. Consider any ERM algorithm A over $\mathcal{H}$. Let $\epsilon, \delta \in (0, 1)$. Suppose m is an integer that satisfies

$$m \geq \frac{\ln(|\mathcal{H}|) + \ln(1/\delta)}{\epsilon} .$$

If $S \sim D^m$ then

$$\Pr\left[L_D(A(S)) < \epsilon\right] \geq 1 - \delta .$$

ERM over finite hypothesis classes

Proof (part 1)

- Note that $L_S(h^*) = L_D(h^*) = 0$.
- Let $\mathcal{H}_{\text{bad}} = \{h \in \mathcal{H} \ : \ L_D(h) \geq \epsilon\}$.
- If $L_S(h) \neq 0$ for all $h \in \mathcal{H}_{\text{bad}}$ then $L_D(A(S)) < \epsilon$.
- $L_D(A(S)) \geq \epsilon$ can happen only if $\exists h \in \mathcal{H}_{\text{bad}}$ s.t. $L_S(h) = 0$.
- Let us upper bound the probability $\exists h \in \mathcal{H}_{\text{bad}}$ s.t. $L_S(h) = 0$.

ERM over finite hypothesis classes

Proof (part 2)

Consider a predictor $h \in \mathcal{H}_{\text{bad}}$. Suppose $L_D(h) = p$.
Since $h \in \mathcal{H}_{\text{bad}}$,

$$p \geq \epsilon \,.$$

Let $S \sim D^m$ where $S = ((x_1, y_1), (x_2, y_2) \ldots, (x_m, y_m))$. Then,

$$\begin{aligned}
\Pr\left[L_S(h) = 0\right] &= \Pr\left[h(x_1) = y_1, h(x_2) = y_2, \ldots, h(x_m) = y_m\right] \\
&= \prod_{i=1}^{m} \Pr\left[h(x_i) = y_i\right] \\
&= (1-p)^m \\
&\leq (1-\epsilon)^m \\
&\leq e^{-\epsilon m} \,.
\end{aligned}$$

ERM over finite hypothesis classes

Proof (part 3)

Let E_h be the event that $L_S(h) = 0$.

$$
\begin{aligned}
\Pr[\overbrace{L_D(A(S)) \geq \epsilon}^{\text{“bad event”}}] &= \Pr[A(S) \in \mathcal{H}_{\text{bad}}] \\
&\leq \Pr\left[\exists h \in \mathcal{H}_{\text{bad}} \text{ s.t. } L_S(h) = 0\right] \\
&= \Pr\left[\bigcup_{h \in \mathcal{H}_{\text{bad}}} E_h\right] \\
&\leq \sum_{h \in \mathcal{H}_{\text{bad}}} \Pr\left[L_S(h) = 0\right] \qquad \text{(union bound)} \\
&\leq \sum_{h \in \mathcal{H}_{\text{bad}}} e^{-\epsilon m} = |\mathcal{H}_{\text{bad}}| e^{-\epsilon m} \leq |\mathcal{H}| e^{-\epsilon m}
\end{aligned}
$$

ERM over finite hypothesis classes

Proof (part 4)

Recall that

$$\Pr[L_D(A(S)) \geq \epsilon] \leq |\mathcal{H}| e^{-\epsilon m} .$$

Note that $|\mathcal{H}| e^{-\epsilon m} \leq \delta$ if and only if

$$m \geq \frac{\ln(|\mathcal{H}|) + \ln(1/\delta)}{\epsilon} .$$

Thus,

$$\Pr[L_D(A(S)) \geq \epsilon] \leq |\mathcal{H}| e^{-\epsilon m} \leq \delta .$$

Thus,

$$\Pr[L_D(A(S)) < \epsilon] \geq 1 - \delta .$$

■

Concentration inequalities

Markov's inequality

Theorem (Markov's inequality)

Let X be a non-negative random variable. For any $t > 0$,

$$\Pr[X \geq t] \leq \frac{\mathbf{E}[X]}{t} .$$

Markov's inequality

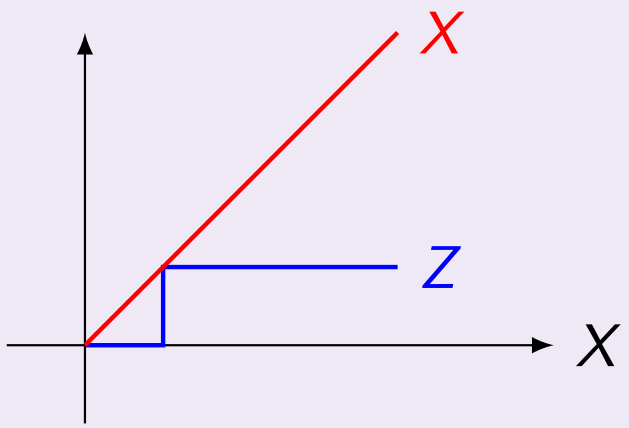
Proof (part 1)

Define a new random variable

$$Z = \begin{cases} 0 & \text{if } X < t, \\ t & \text{if } X \geq t. \end{cases}$$

With probability one,

$$Z \leq X\,.$$



Markov's inequality

Proof (part 2)

Take expectation of both sides of $Z \leq X$:

$$\mathbf{E}[Z] \leq \mathbf{E}[X]$$

By definition of Z

$$\mathbf{E}[Z] = 0 \cdot \Pr[X < t] + t \cdot \Pr[X \geq t] = t \cdot \Pr[X \geq t] \,.$$

Thus,

$$t \cdot \Pr[X \geq t] \leq \mathbf{E}[X] \,.$$

The inequality follows by dividing through by t.

Chebyshev's inequality

Theorem (Chebyshev's inequality)

Let X be a random variable such that $\mathbf{E}[X]$ and $\mathbf{Var}[X]$ exist. For any $t > 0$,

$$\Pr[|X - \mathbf{E}[X]| \geq t] \leq \frac{\mathbf{Var}[X]}{t^2} .$$

Recall that $\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$.

Proof.

Let $Z = (X - \mathbf{E}[X])^2$. Since $Z \geq 0$, by Markov's inequality,

$$\begin{aligned}\Pr[|X - \mathbf{E}[X]| \geq t] &= \Pr[(X - \mathbf{E}[X])^2 \geq t^2] \\ &= \Pr[Z \geq t^2] \\ &\leq \frac{\mathbf{E}[Z]}{t^2} = \frac{\mathbf{Var}[X]}{t^2} .\end{aligned}$$

■

An application of Chebyshev's inequality: Binomial tails

- Let $X_1, X_2, \ldots$ be i.i.d. Bernoullli(p) for some $p \in (0,1)$.
- Let $Z_n = \frac{1}{n}\sum_{i=1}^n X_i$.
- $\mathbf{E}[Z_n] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}[X_i] = p$
- $\mathbf{Var}[Z_n] = \frac{1}{n^2}\sum_{i=1}^n \mathbf{Var}[X_i] = \frac{p(1-p)}{n}$
- Chebyshev's inequality implies that for any $t > 0$,

$$\Pr[|Z_n - p| \geq t] \leq \frac{p(1-p)}{nt^2} \,.$$

Central limit theorem

Theorem (Central limit theorem)

Let $\mu \in \mathbb{R}$ and $\sigma > 0$. Let $X_1, X_2, \ldots$ be i.i.d. random variables such that $\mathbf{E}[X_i] = \mu$ and $\mathbf{Var}[X_i] = \sigma^2$ for all $i = 1, 2, \ldots$. Then, for any $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \Pr\left[\frac{(\sum_{i=1}^n X_i) - n\mu}{\sigma\sqrt{n}} \geq t\right] = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\,\mathrm{d}x\,.$$

The distribution of a properly normalized sum $\sum_{i=1}^n X_i$ converges to the normal distribution $N(0,1)$ with mean 0 and variance 1.

For any $t \geq 0$,

$$\frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\mathrm{d}x \leq \frac{1}{2} e^{-t^2/2}\,.$$

An application of the central limit theorem: Binomial tails

- Let $X_1, X_2, \ldots$ be i.i.d. Bernoullli(p) for some $p \in (0,1)$.
- Let $Z_n = \frac{1}{n}\sum_{i=1}^n X_i$.
- $\mu = \mathbf{E}[X_i] = p$ and $\sigma^2 = \mathbf{Var}[X_i] = p(1-p)$
- Central limit theorem implies that for any $t > 0$,

$$\Pr\left[|Z_n - p| \geq t\right] = \Pr\left[Z_n - p \geq t\right] + \Pr\left[(-Z_n) - (-p) \geq t\right]$$

$$= \Pr\left[(\textstyle\sum_{i=1}^n X_i) - np \geq nt\right] + \Pr\left[(\textstyle\sum_{i=1}^n -X_i) - n(-p) \geq nt\right]$$

$$= \Pr\left[\frac{(\sum_{i=1}^n X_i) - np}{\sqrt{np(1-p)}} \geq t\sqrt{\frac{n}{p(1-p)}}\right]$$

$$+ \Pr\left[\frac{(\sum_{i=1}^n -X_i) - n(-p)}{\sqrt{np(1-p)}} \geq t\sqrt{\frac{n}{p(1-p)}}\right]$$

$$\approx 2\frac{1}{\sqrt{2\pi}}\int_{t\sqrt{\frac{n}{p(1-p)}}}^{\infty} e^{-x^2/2}\,\mathrm{d}x \leq e^{-\frac{nt^2}{2p(1-p)}}$$

Binomial tails

- The bound from Chebyshev's inequality is

$$\Pr[|Z_n - p| \geq t] \leq \frac{p(1-p)}{nt^2} .$$

- The bound from Central limit theorem is

$$\Pr[|Z_n - p| \geq t] \lesssim e^{-\frac{nt^2}{2p(1-p)}} .$$

- Right-hand sides are functions of $\frac{nt^2}{p(1-p)}$.
- The second inequality is much stronger.

Hoeffding's inequality

Theorem (Hoeffding's inequality)

Let $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n$ be real numbers. Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $X_i \in [a_i, b_i]$ for $i = 1, 2, \ldots, n$ with probability one. Then, for any $\epsilon > 0$,

$$\Pr\left[\sum_{i=1}^{n} X_i \geq \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] + \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right), \quad (1)$$

$$\Pr\left[\sum_{i=1}^{n} X_i \leq \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] - \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right), \quad (2)$$

$$\Pr\left[\left|\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \quad (3)$$

Hoeffding's inequality

Proof of Hoeffding's inequality (part 1)

- Inequality (2) by applying (1) to $-X_1, -X_2, \ldots, -X_n$.
- Inequality (3) follows (1), (2) and union bound.
- It suffices to prove (1).
- Let $Y_i = X_i - \mathbf{E}[X_i]$.
- Let $Z = \sum_{i=1}^{n} Y_i$.
- We will use so called *Chernoff's bounding technique*. It involves bounding the *moment-generating functions* of Z,

$$M(t) = \mathbf{E}\left[e^{tZ}\right] = \mathbf{E}\left[\sum_{k=0}^{\infty} \frac{t^k Z^k}{k!}\right] = \sum_{k=0}^{\infty} \frac{t^k \, \mathbf{E}\left[Z^k\right]}{k!} \, .$$

Hoeffding's inequality

Proof of Hoeffding's inequality (part 2)

For any $t > 0$,

$$\Pr\left[\sum_{i=1}^{n} X_i \geq \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] + \epsilon\right] = \Pr\left[\sum_{i=1}^{n} Y_i \geq \epsilon\right] = \Pr\left[Z \geq \epsilon\right]$$

$$= \Pr\left[e^{tZ} \geq e^{t\epsilon}\right] \leq \frac{\mathbf{E}\left[e^{tZ}\right]}{e^{t\epsilon}} \quad \text{(Markov's inequality)}$$

$$= \frac{\mathbf{E}\left[e^{t\sum_{i=1}^{n} Y_i}\right]}{e^{t\epsilon}} = \frac{\mathbf{E}\left[\prod_{i=1}^{n} e^{tY_i}\right]}{e^{t\epsilon}}$$

$$= \frac{\prod_{i=1}^{n} \mathbf{E}\left[e^{tY_i}\right]}{e^{t\epsilon}} \quad \text{(independence)}$$

It remains to upper bound $\mathbf{E}\left[e^{tY_i}\right]$ i.e. the moment generating function of Y_i.

Hoeffding's inequality

Lemma (Hoeffding's lemma)

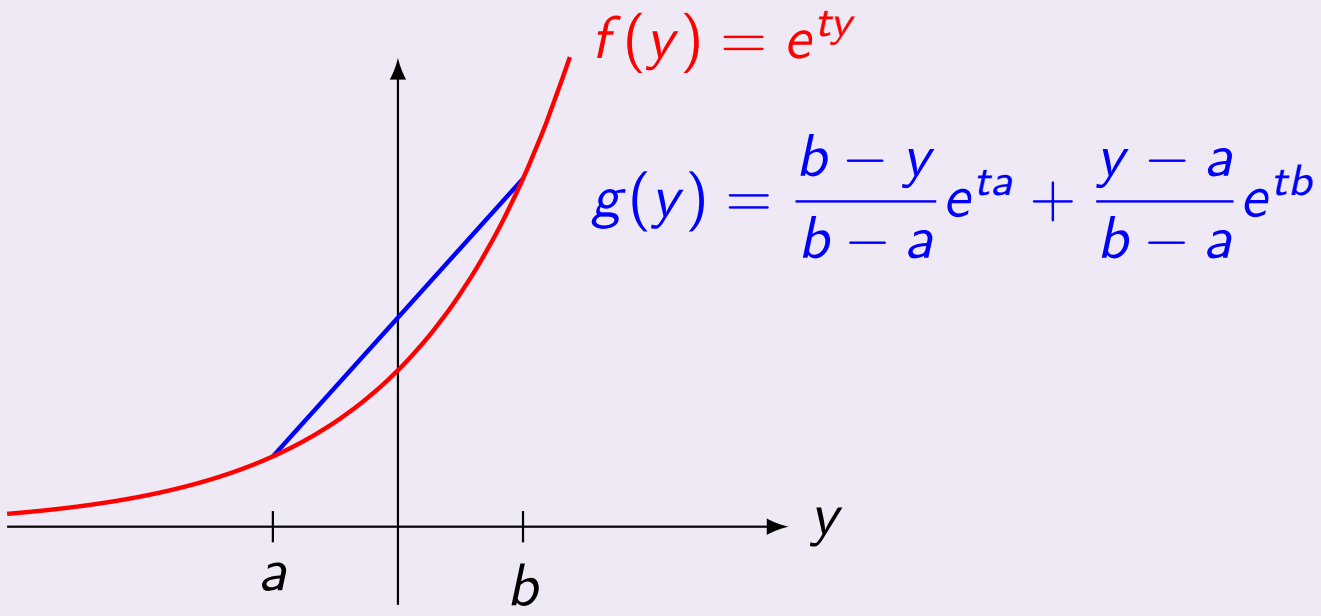
Let Y be a random variable such that $Y \in [a, b]$ with probability one and $\mathbf{E}[Y] = 0$. Then, for any real number t,

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

Hoeffding's inequality

Proof of Hoeffding's lemma (part 1)

The function $f(y) = e^{ty}$ is a convex.



Hoeffding's inequality

Proof of Hoeffding's lemma (part 2)

For any $y \in [a, b]$,

$$e^{ty} \leq \frac{b-y}{b-a} e^{ta} + \frac{y-a}{b-a} e^{tb} .$$

Thus, with probability one,

$$e^{tY} \leq \frac{b-Y}{b-a} e^{ta} + \frac{Y-a}{b-a} e^{tb} .$$

Taking expectation and since $\mathbf{E}[Y] = 0$,

$$\mathbf{E}\left[e^{tY}\right] \leq \frac{b}{b-a} e^{ta} - \frac{a}{b-a} e^{tb} .$$

Hoeffding's inequality

Proof of Hoeffding's lemma (part 3)

It remains to prove that

$$\frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} \leq e^{-\frac{t^2(b-a)^2}{8}} .$$

Let $p = \frac{-a}{b-a}$. (Note that $p \in [0,1]$, since $a \leq 0 \leq b$.)

$$\begin{aligned}
\frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} &= (1-p)e^{ta} + pe^{tb} \\
&= e^{ta}(1-p+pe^{t(b-a)}) \\
&= e^{-pt(b-a)}(1-p+pe^{t(b-a)}) \\
&= e^{-pu}(1-p+pe^{u})
\end{aligned}$$

where $u = t(b-a)$.

Hoeffding's inequality

Proof of Hoeffding's lemma (part 4)

We have

$$e^{-pu}(1-p+pe^u) = e^{\phi(u)}$$

where

$$\phi(u) = -pu + \log(1-p+pe^u)\,.$$

It remains to prove that

$$e^{\phi(u)} \leq e^{t^2(b-a)^2/8}\,.$$

Equivalently, we need to prove that

$$\phi(u) \leq u^2/8\,.$$

Hoeffding's inequality

Proof of Hoeffding's lemma (part 5)

We compute second order Taylor approximation of $\phi(u)$ at $u = 0$:

$$\phi(u) = -pu + \log(1 - p + pe^u) \qquad \phi(0) = 0$$

$$\phi'(u) = -p + \frac{pe^u}{1 - p + pe^u} \qquad \phi'(0) = 0$$

$$\phi''(u) = \frac{pe^u(1 - p + pe^u) - (pe^u)^2}{(1 - p + pe^u)^2}$$

$$= \frac{p(1-p)e^u}{(1 - p + pe^u)^2} = \frac{AB}{(A+B)^2} \leq 1/4$$

since $\sqrt{AB} \leq \frac{A+B}{2}$.

Hoeffding's inequality

Proof of Hoeffding's lemma (part 6)

By Taylor's theorem there exists v between 0 and u such that

$$\phi(u) = \phi(0) + \phi'(0)u + \phi''(v)\frac{u^2}{2}$$

Therefore,

$$\phi(u) \leq u^2/8\,.$$

■

Hoeffding's inequality

Proof of Hoeffding's inequality (part 3)

For any $t > 0$,

$$\Pr\left[\sum_{i=1}^{n} X_i \geq \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] + \epsilon\right] \leq \frac{\prod_{i=1}^{n} \mathbf{E}\left[e^{tY_i}\right]}{e^{t\epsilon}} \leq \frac{\prod_{i=1}^{n} e^{\frac{t^2(b_i-a_i)^2}{8}}}{e^{t\epsilon}}$$

$$= \exp\left(\frac{t^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - t\epsilon\right)$$

Choose t that minimizes the last expression. (Note that the exponent is a quadratic function of t.) Set

$$t = \frac{4\epsilon}{\sum_{i=1}^{n}(b_i - a_i)^2}$$

Hoeffding's inequality

Proof of Hoeffding's inequality (part 3)

For $t = \frac{4\epsilon}{\sum_{i=1}^n (b_i - a_i)^2}$,

$$\Pr\left[\sum_{i=1}^n X_i \geq \mathbf{E}\left[\sum_{i=1}^n X_i\right] + \epsilon\right]$$

$$\leq \exp\left(\frac{t^2}{8}\sum_{i=1}^n (b_i - a_i)^2 - t\epsilon\right)$$

$$= \exp\left(\frac{16\epsilon^2}{8\left(\sum_{i=1}^n (b_i - a_i)^2\right)^2}\sum_{i=1}^n (b_i - a_i)^2 - \frac{4\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

$$= \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

■

No free lunch theorem

No free lunch theorem

Theorem (No free lunch theorem)

Let $\mathcal{X}$ be an infinite domain. Let $\mathcal{Y} = \{0, 1\}$. Let $A : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$ be any learning algorithm. Let m be a positive integer. There exists a distribution D over $\mathcal{X} \times \mathcal{Y}$ and a classifier[1] $h^* : \mathcal{X} \to \mathcal{Y}$ such that

$$L_D(h^*) = 0$$

and, for a labeled sample $S \sim D^m$,

$$L_D(A(S)) > 1/8$$

with probability at least $1/7$.

[1] h^* is the Bayes optimal classifier for D.

No free lunch theorem

Proof of No free lunch theorem (part 1)

- Let $C \subseteq \mathcal{X}$ be of size $2m$.
- Let U be uniform distribution over C.
- There are 2^{2m} different functions in $\mathcal{Y}^C$.
- For any $h \in \mathcal{Y}^C$ and $x \in \mathcal{X} \setminus C$, we define $h(x) = 0$.
- For any $h \in \mathcal{Y}^C$, define distribution D_h over $\mathcal{X} \times \mathcal{Y}$ as follows

$$D_h(\{(x,y)\}) = \begin{cases} U(\{x\}) & \text{if } h(x) = y, \\ 0 & \text{if } h(x) \neq y. \end{cases}$$

- $L_{D_h}(h) = 0$ by construction of D_h.
- $D = D_h$ and $h^* = h$ for some $h \in \mathcal{Y}^C$; to be specified later.

No free lunch theorem

Proof of No free lunch theorem (part 2)

We lower bound

$$\max_{h \in C} \mathbf{E}_{S \sim D_h^m}[L_{D_h}(A(S))] \,.$$

Let $T \sim U^m$ be an unlabeled sample $T = (x_1, x_2, \ldots, x_m)$. For any classifier h, let

$$T_h = ((x_1, h(x_1)), (x_2, h(x_2)), \ldots, (x_m, h(x_m)))$$

be a labeled sample.

$$\begin{aligned}\max_{h \in C} \mathbf{E}_{S \sim D_h^m}[L_{D_h}(A(S))] &= \max_{h \in C} \mathbf{E}_{T \sim U^m}[L_{D_h}(A(T_h))] \\ &\geq \frac{1}{2^{2m}} \sum_{h \in \mathcal{Y}^C} \mathbf{E}_{T \sim U^m}[L_{D_h}(A(T_h))]\end{aligned}$$

No free lunch theorem

Proof of No free lunch theorem (part 3)

$$
\begin{aligned}
&\max_{h \in C} \mathbf{E}_{S \sim D_h^m}[L_{D_h}(A(S))] \\
&= \max_{h \in C} \mathbf{E}_{T \sim U^m}[L_{D_h}(A(T_h))] \\
&\geq \frac{1}{2^{2m}} \sum_{h \in \mathcal{Y}^C} \mathbf{E}_{T \sim U^m}[L_{D_h}(A(T_h))] \\
&= \mathbf{E}_{T \sim U^m}\left[\frac{1}{2^{2m}} \sum_{h \in \mathcal{Y}^C} L_{D_h}(A(T_h))\right] \\
&= \mathbf{E}_{T \sim U^m}\left[\frac{1}{2^{2m}} \sum_{h \in \mathcal{Y}^C} \frac{1}{2m} \sum_{x \in C} \mathbf{1}\left[A(T_h)(x) \neq h(x)\right]\right]
\end{aligned}
$$

No free lunch theorem

Proof of No free lunch theorem (part 4)

$$= \mathbf{E}_{T\sim U^m}\left[\frac{1}{2^{2m}}\sum_{h\in\mathcal{Y}^C}\frac{1}{2m}\sum_{x\in C}\mathbf{1}\left[A(T_h)(x)\neq h(x)\right]\right]$$

$$= \mathbf{E}_{T\sim U^m}\left[\frac{1}{2m}\sum_{x\in C}\frac{1}{2^{2m}}\sum_{h\in\mathcal{Y}^C}\mathbf{1}\left[A(T_h)(x)\neq h(x)\right]\right]$$

$$= \mathbf{E}_{T\sim U^m}\left[\frac{1}{2m}\sum_{x\in C\setminus V_T}\frac{1}{2^{2m}}\sum_{h\in\mathcal{Y}^C}\mathbf{1}\left[A(T_h)(x)\neq h(x)\right]\right]$$

where $V_T = \{x_1, x_2, \ldots, x_m\}$ is the set of (distinct) examples in $T = (x_1, x_2, \ldots, x_m)$.

No free lunch theorem

Proof of No free lunch theorem (part 5)

We lower bound

$$\frac{1}{2m} \sum_{x \in C \setminus V_T} \frac{1}{2^{2m}} \sum_{h \in \mathcal{Y}^C} \mathbf{1}\left[A(T_h)(x) \neq h(x)\right] .$$

For any $g \in \mathcal{Y}^{V_T}$, let

$$H_g = \left\{ h \in \mathcal{Y}^C \ : \ h(x) = g(x) \text{ for all } x \in V_T \right\} .$$

Note that $|H_g| = 2^{2m - |V_T|}$ and $\bigcup_{g \in \mathcal{Y}^{V_T}} H_g = \mathcal{Y}^C$ is a partition of $\mathcal{Y}^C$ into $2^{|V_T|}$ sets.

No free lunch theorem

Proof of No free lunch theorem (part 6)

$$
\begin{aligned}
&\frac{1}{2m}\sum_{x\in C\setminus V_T}\frac{1}{2^{2m}}\sum_{h\in\mathcal{Y}^C}\mathbf{1}\left[A(T_h)(x)\neq h(x)\right]\\
&=\frac{1}{2m}\sum_{x\in C\setminus V_T}\frac{1}{2^{|V_T|}}\sum_{g\in\mathcal{Y}^{V_T}}\frac{1}{2^{2m-|V_T|}}\sum_{h\in H_g}\mathbf{1}\left[A(T_g)(x)\neq h(x)\right]\\
&=\frac{1}{2m}\sum_{x\in C\setminus V_T}\frac{1}{2^{|V_T|}}\sum_{g\in\mathcal{Y}^{V_T}}\frac{1}{2}\\
&=\frac{1}{2m}\sum_{x\in C\setminus V_T}\frac{1}{2}\\
&=\frac{2m-|V_T|}{4m}\geq\frac{2m-m}{4m}=\frac{1}{4}\,.
\end{aligned}
$$

No free lunch theorem

Proof of No free lunch theorem (part 7)

Thus,

$$\max_{h \in C} \mathbf{E}_{S \sim D_h^m}[L_{D_h}(A(S))] \geq 1/4 \,.$$

Therefore, there exists $h^* \in \mathcal{Y}^C$ such that

$$\mathbf{E}_{S \sim D_{h^*}^m}[L_{D_{h^*}}(A(S))] \geq 1/4 \,.$$

We define $D = D_{h^*}$. Thus, there exists D such that

$$\mathbf{E}_{S \sim D^m}[L_D(A(S))] \geq 1/4 \qquad \text{and} \qquad L_D(h^*) = 0 \,.$$

No free lunch theorem

Proof of No free lunch theorem (part 8)

Let $Z = L_D(A(S))$. We know that $Z \in [0,1]$ and

$$\mathbf{E}[Z] \geq 1/4\,.$$

Let $W = 1 - Z$. Clearly $W \in [0,1]$ and $\mathbf{E}[W] \leq 3/4$. By Markov's inequality,

$$\Pr[W \geq 7/8] \leq \frac{\mathbf{E}[W]}{7/8} \leq \frac{3/4}{7/8} = 6/7\,.$$

Therefore,

$$\begin{aligned}\Pr[L_D(A(S)) > 1/8] &= \Pr[Z > 1/8] = \Pr[W < 7/8] \\ &= 1 - \Pr[W \geq 7/8] \geq 1 - 6/7 = 1/7\,. \quad \blacksquare\end{aligned}$$