Supervised Learning Coursework 2

Part 1

This all works if we treat $\bar{X}$ as a random variable

1.1 [2 marks] let $\bar{X} = \max_i X_i$. Show that for any $\lambda > 0$

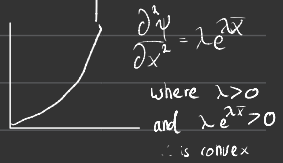
$$\mathbb{E}\bar{X} \leq \frac{1}{\lambda}\mathbb{E}e^{\lambda\bar{X}}$$

Note: We can treat $\bar{X}$ as a random variable because any function of a random variable is also a random variable.

Useful Inequalities

→ Jensen's Inequality: "For a random variable X and a convex function ψ, $\mathbb{E}(\psi(X)) \geq \psi(\mathbb{E}[X])$"

This directly applys to the inequality we're trying to prove, if we let $\psi(\bar{X}) = \frac{1}{\lambda}e^{\lambda\bar{X}}$. But is $\frac{1}{\lambda}e^{\lambda\bar{X}}$ a convex function



$$\therefore \quad \mathbb{E}\left[\frac{1}{\lambda}e^{\lambda\bar{X}}\right] \geq \frac{1}{\lambda}e^{\lambda\mathbb{E}[\bar{X}]}$$

$$\therefore \quad \frac{1}{\lambda}\mathbb{E}\left[e^{\lambda\bar{X}}\right] \geq \frac{1}{\lambda}e^{\lambda\mathbb{E}[\bar{X}]}$$

We can log both sides → Because $\forall x,y \in \mathbb{R}$ if $e^x \leq e^y$ then $x \leq y$ ∴ e^x is strictly increasing!

$$\log\left(\frac{1}{\lambda}\mathbb{E}\left[e^{\lambda\bar{X}}\right]\right) \geq \log\frac{1}{\lambda}e^{\lambda\mathbb{E}(\bar{X})}$$

$$\log\tfrac{1}{\lambda} + \log\left(\mathbb{E}\left[e^{\lambda\bar{X}}\right]\right) \geq \log\tfrac{1}{\lambda} + \lambda\mathbb{E}[\bar{X}]$$

$$\therefore \quad \lambda\mathbb{E}[\bar{X}] \leq \log\mathbb{E}\left[e^{\lambda\bar{X}}\right]$$

$$\mathbb{E}[\bar{X}] \leq \frac{1}{\lambda}\log\mathbb{E}\left[e^{\lambda\bar{X}}\right]$$

1.2 [5 marks] Show that

$$\frac{1}{\lambda}\log \mathbb{E}e^{\lambda \bar{X}} \leq \frac{1}{\lambda}\log m + \lambda \frac{(b-a)^2}{8}$$

(m: number of rvs)

→ Useful inequalities:

- Hoeffding's lemma: For any random variable X such that $X - \mathbb{E}X \in [a,b]$ with $a,b \in \mathbb{R}$ and for any $\lambda > 0$, we have:

$$\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \leq e^{\lambda^2(b-a)^2/8}$$

What does $X - \mathbb{E}X$ mean? → For a random sample, how far away from the mean was it? (or displacement as it could be -ve.)

- But our variables are centred so $\mathbb{E}X = 0$

∴

$$\mathbb{E}e^{\lambda X} \leq e^{\frac{\lambda^2(b-a)^2}{8}}$$

We are interested in $\bar{X}$ not X ∴ since e^x is strictly increasing we can write $e^{\lambda x}$ like so:

→ We know $\bar{X} = \max_i X_i$ ∴ $\lambda\bar{X} = \max_i \lambda X_i$ ∴ $e^{\lambda\bar{X}} = \max_i e^{\lambda X_i}$

But how do we find the expectation of a maximum?
We can bound it at least by summing over the expectations.....
inequality: $\mathbb{E}[\max_i X_i] \leq \sum_i \mathbb{E}X_i$

→ ∴ $\mathbb{E}\left[e^{\lambda\bar{X}}\right] \leq \sum_{i=1}^{m} \mathbb{E}\left[e^{\lambda X_i}\right]$

We can now bring the bound from Hoeffding's lemma back in!

$$\mathbb{E}[e^{\lambda \bar{X}}] \leq \sum_{i=1}^{m} e^{\lambda^2 (b-a)^2/8}$$

$$\mathbb{E}[e^{\lambda \bar{X}}] \leq m\, e^{\lambda^2 (b-a)^2/8}$$

We can now log both sides and divide by λ:

$$\frac{1}{\lambda} \log \mathbb{E}[e^{\lambda \bar{X}}] \leq \frac{1}{\lambda} \log m + \frac{\lambda (b-a)^2}{8}$$

1.3 [3 marks] Conclude that by choosing λ appropriately,

$$\mathbb{E}\left[\max_{i=1,\dots,m} X_i\right] \leq \frac{b-a}{2}\sqrt{2 \log m}$$

$= \bar{X}$?

We know $\mathbb{E}\bar{X} \leq \frac{1}{\lambda} \log \mathbb{E} e^{\lambda \bar{X}} \leq \frac{1}{\lambda} \log m + \frac{\lambda (b-a)^2}{8}$

What is the smallest this can be? To get it as close to $\frac{1}{\lambda} \log \mathbb{E} e^{\lambda \bar{X}}$ as possible?

- well we can only change λ
- lets minimise w.r.t λ:

$$\frac{\partial}{\partial \lambda}\left[\frac{1}{\lambda} \log m + \frac{\lambda (b-a)^2}{8}\right] = 0$$

$$\frac{\partial}{\partial \lambda}\left[\frac{1}{\lambda} \log m\right] + \frac{\partial}{\partial \lambda}\left[\frac{\lambda (b-a)^2}{8}\right] = 0$$

$$\frac{d}{d\lambda}\left[\lambda^{-1}\log m\right] + \frac{d}{d\lambda}\left[\frac{\lambda(b-a)^2}{8}\right] = 0$$

$$-\lambda^{-2}\log m + \frac{(b-a)^2}{8} = 0$$

$$(b-a)^2 = 8\lambda^{-2}\log m$$

$$\lambda_{min} = \pm\sqrt{\frac{8\log m}{(b-a)^2}} \rightsquigarrow \frac{1}{\lambda_{min}} = \frac{1}{b-a}\sqrt{\frac{1}{8\log m}} = \frac{1}{2(b-a)}\sqrt{\frac{1}{2\log m}}$$

but $\lambda > 0$ ∴ just the +ve

sub λ_{min} : $\quad \mathbb{E}\bar{X} \le \frac{1}{\lambda}\log \mathbb{E}e^{\lambda\bar{X}} \le \frac{1}{\lambda}\log m + \frac{\lambda(b-a)^2}{8}$

$$\lambda_{min}^{-1} = \frac{1}{\sqrt{\frac{8\log m}{(b-a)^2}}} \qquad \therefore \mathbb{E}\bar{X} \le \frac{\log m}{\sqrt{\frac{8\log m}{(b-a)^2}}} + \frac{\sqrt{\frac{8\log m}{(b-a)^2}}(b-a)^2}{8}$$

$$\therefore \mathbb{E}\bar{X} \le \frac{8\log m + \frac{8\log m\,(b-a)^2}{(b-a)^2}}{8\sqrt{\frac{8\log m}{(b-a)^2}}}$$

$$\frac{x}{\sqrt{2x}} \cdot \frac{\sqrt{2x}}{\sqrt{2x}} = \frac{x\sqrt{2x}}{2x} = \frac{\sqrt{2x}}{2}$$

$$\frac{\sqrt{2x}}{2} = \frac{x}{\sqrt{2x}}$$

$$\mathbb{E}\bar{X} \le \frac{\log m + \log m}{2\sqrt{\frac{2\log m}{(b-a)^2}}}$$

$$\mathbb{E}\bar{X} \le \frac{\log m \times (b-a)}{\sqrt{2\log m}}$$

$$\mathbb{E}\bar{X} \le \frac{\sqrt{2\log m}}{2}(b-a)$$

1.4 [3 marks]

Useful Notes on R.C: (finite hypothesis space)

- Overall goal of question to bound the R.C for $\mathcal{H}$
- S is a finite set of points $\in \mathbb{R}^n$, $|S| = m$

$$R(S) = \mathbb{E}_{\sigma}\left[\max_{x \in S} \frac{1}{n} \sum_{j=1}^{n} \sigma_j x_j\right]$$

(summed over the features)

(σ: Rademacher variable $\{-1, 1\}$ $\sigma_1 \ldots \sigma_n \sim U\{-1, 1\}$)

Show that:

$$R(S) \leq \max_{x \in S} \|\vec{x}\|_2 \frac{\sqrt{2\log(m)}}{n}$$

($\|\vec{x}\|_2$: Euclidean norm or just the magnitude of the vector $\vec{x}$)

We need to show that $\mathbb{E}_{\sigma} \max_{x \in S} \frac{1}{n} \sum_{j=1}^{n} \sigma_j x_j$ is no bigger than $\max_{x \in S} \|x\|_2 \frac{\sqrt{2\log(m)}}{n}$.

- $\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$ $\therefore \|x\|_2^2 = \sum_{j=1}^{n} x_j^2$ but σ_j makes things weird $\because$ $p(\sigma_j = 1) = \frac{1}{2}$, $p(\sigma_j = -1) = \frac{1}{2}$

Useful Inequalities and lemmas

~> Massarts Lemma

If we have a finite set of points $S \in \mathbb{R}^n$ and we let $r = \max_{x \in S} \|x\|_2$

$$\mathbb{E}_{\sigma}\left[\max_{x \in S} \sum_{i=1}^{n} x_i \sigma_i\right] \leq r\sqrt{2\log(|S|)}$$

($|S| = m$ in our case)

We want to upper bound this:

$\frac{1}{n} \mathbb{E}_{\sigma}\left[\max_{x \in S} \frac{1}{n} \sum_{j=1}^{n} \sigma_j x_j\right]$ with this $\max_{x \in S} \|\vec{x}\|_2 \frac{\sqrt{2\log(m)}}{n}$

- There is a difference of $\frac{1}{n}$ in the LHS of Massarts lemma.

note: $\frac{1}{a}\mathbb{E}X = \mathbb{E}\frac{X}{a}$ if a is a constant. $\therefore$ we can take that $\frac{1}{n}$ factor out.

Coming back to Massarts lemma

$$\mathbb{E}_\sigma\left[\max_{x \in S} \sum_{i=1}^{n} x_i \sigma_i\right] \leq r\sqrt{2\log(|S|)}$$

- In our problem we know

- $r = \max_{x \in S} \|\vec{x}\|_2$

- $|S| = m$

$\Rightarrow$ So $\quad \mathbb{E}_\sigma\left[\max_{x \in S} \sum_{i=1}^{n} x_i \sigma_i\right] \leq \max_{x \in S} \|\vec{x}\|_2 \sqrt{2\log m}$

[dividing by n] $\rightarrow \quad \frac{1}{n}\mathbb{E}_\sigma\left[\max_{x \in S} \sum_{i=1}^{n} x_i \sigma_i\right] \leq \frac{1}{n} \max_{x \in S} \|\vec{x}\|_2 \sqrt{2\log m}$

- The n can go inside the maximum because its a constant

$$\therefore \quad \mathbb{E}_\sigma\left[\max_{x \in S} \frac{1}{n} \sum_{i=1}^{n} \sigma_i x_i\right] \leq \max_{x \in S} \frac{1}{n} \|x\|_2 \sqrt{2\log m}$$

Which is the expression we needed to show.

1.5 [7 marks]

Q: Let $\mathcal{H}$ be a set of hypotheses $f: X \to \mathbb{R}$. Assume $\mathcal{H}$ to have finite cardinality $|\mathcal{H}| < -\infty$. Let $S = (x_i)_{i=1}^n$ be a set of points in X, an input set. Use the reasoning above (in previous parts) to prove an upper bound for empirical Rademacher complexity $R_S(\mathcal{H})$, where the cardinality of $\mathcal{H}$ appears logarithmically.

Definition of the empirical Radamacher complexity:

$$R_S(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right]$$

→ The big change is that we are now working with a set of functions rather than of points...

What do we know?

1. $R(S) = \mathbb{E}_\sigma \left[\max_{x \in S} \frac{1}{n} \sum_{j=1}^{n} \sigma_j x_j \right]$, where σ_j are the Rademacher variables sampled with equal prob from $\{-1, 1\}$

(1.4)

2. $R(S) \leq \max_{x \in S} \|x\|_2 \frac{1}{n} \sqrt{2 \log m}$

(1.3)

3. $\mathbb{E}\,\bar{X} \leq \frac{b-a}{2} \sqrt{2 \log m}$

$\max_{i=1,\dots,m} \vec{x}_i$

What do we have to do again?

- We want to upper bound the empirical Rademacher complexity

Could we just treat $h(x_i)$ as a real valued vector of points (just like part 1.4) $\rightarrow h(x_i) \in \mathbb{R}^n$?

In which case, we could use 1.4's equation

$$R_S(\mathcal{H}) = \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(x_i)\right] \leq \max_{h\in\mathcal{H}} \|h(x)\|_2 \frac{\sqrt{n \log(|\mathcal{H}|)}}{n}$$

rather than $\|\cdot\|_2$ (on $\max$); rather than $|S|$ (on $\log$); rather than just x_i (on $h(x_i)$)

- $|\mathcal{H}|$ has to be finite

for us to be able to represent $h(x_i)$ as a set of points and therefore use eq 1.4

∴ our final bound:

$$R_S(\mathcal{H}) \leq \max_{h\in\mathcal{H}} \|h(x)\|_2 \frac{\sqrt{n \log(|\mathcal{H}|)}}{n}$$