CS7545, Spring 2024: Machine Learning Theory - Solutions #1

Jacob Abernethy, Tyler Labonte, Guanghui Wang, Yeojoon Youn Due: Friday, February 2 at 11:59 p.m.

1) **Norm.** We will prove a generic statement which implies (a)-(d).

Let $p > q \geq 1$, and r be a number such that $\frac{1}{p} + \frac{1}{r} = \frac{1}{q}$. Then, $q < p, r$ and $(\frac{p}{q}, \frac{r}{q})$ is a conjugate norm pair. Let $\mathbf{a} \in \mathbb{R}^N$ such that $a_i = |x_i|^q$, and let $\mathbf{y} = (1, \ldots, 1) \in \mathbb{R}^N$. Now we use Holder's inequality:

$$\mathbf{a}^\top \mathbf{y} = \sum_{i=1}^N |x_i|^q \leq \|\mathbf{a}\|_{\frac{p}{q}} \|\mathbf{y}\|_{\frac{r}{q}} = \left(\sum_{i=1}^N |x_i|^p\right)^{\frac{q}{p}} n^{\frac{q}{r}}.$$

By exponentiating each side with $1/q$, we get

$$\|\mathbf{x}\|_q \leq \|\mathbf{x}\|_p n^{\frac{1}{r}} = \|\mathbf{x}\|_p n^{\frac{1}{q} - \frac{1}{p}}$$

Also note that

$$\|\mathbf{x}\|_q^p = \left(\sum_{i=1}^N |x_i|^q\right)^{\frac{p}{q}} \geq \sum_{i=1}^N |x_i|^p = \|\mathbf{x}\|_p^p$$

which implies $\|\mathbf{x}\|_q \geq \|\mathbf{x}\|_p$. The inequality follows since $(\sum |x_i|)^\alpha \geq \sum |x_i|^\alpha$ whenever $\alpha \geq 1$.

For part e), from the above result, we get

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_p \leq N^{\frac{1}{p}} \|\mathbf{x}\|_\infty$$

Thus, when we apply $\lim_{p \to +\infty}$ to the above inequality, we finally obtain

$$\|\mathbf{x}\|_\infty \leq \lim_{p \to +\infty} \|\mathbf{x}\|_p \leq \|\mathbf{x}\|_\infty \Rightarrow \lim_{p \to +\infty} \|\mathbf{x}\|_p = \|\mathbf{x}\|_\infty$$

2) **Hölder.**

(a) Let $p > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$.Consider the following two vectors:

$$\mathbf{x}, \mathbf{y} \in \mathbb{R}^N : x_i = p_i^{\frac{1}{q} - 1}, y_i = p_i^{\frac{1}{p}},$$

then by Hölder's Inequality,

$$\|\mathbf{x}\|_q \|\mathbf{y}\|_p \geq \mathbf{x}^T \mathbf{y} = \sum_i p_i^{\frac{1}{p} + \frac{1}{q} - 1} = N.$$

where $\|\mathbf{x}\|_q = (\sum_i p_i^{1-q})^{\frac{1}{q}}$ and $\|\mathbf{y}\|_p = 1$. Therefore,

$$\sum_i \left(\sum_i \frac{1}{p_i^{q-1}}\right)^{\frac{1}{q}} \geq N \Rightarrow \sum_i \frac{1}{p_i^{q-1}} \geq N^q.$$

Remark. You can also use Jensen's inequality. Consider the function $f(p) = \frac{1}{p^q}$ and note that $f(\sum_{i=1}^N p_i p_i) \le \sum_{i=1}^N p_i f(p_i)$.

(b) By Jensen's Inequality,

$$\sum_i p_i^2 \ge \sum_i \frac{p_i}{N} = \frac{1}{N}.$$

Therefore, we have

$$\sum_i \left(\frac{1}{p_i} + p_i\right)^2 = \sum_i p_i^2 + \sum_i 2 + \sum_i \frac{1}{p_i^2} \ge \frac{1}{N} + 2N + N^3.$$

Remark. You can use 1(a) to show that $\sum_i p_i^2 = \|\mathbf{p}\|_2^2 \ge \frac{\|\mathbf{p}\|_1^2}{N}$.

3) **Convexity.**

(a) For the convexity of the given function, we need to show $f(\frac{p+q}{2}) \le \frac{f(p)+f(q)}{2}$ for $\forall p, q \in \Delta_N$. By the definition,

$$f(\frac{p+q}{2}) = \sum_{i=1}^N \frac{p_i + q_i}{2} \log(\frac{p_i + q_i}{2})$$

Here, let $g(x) = x \log x$ for a scalar x $(0 < x < 1)$. Since $g''(x) = \frac{1}{x} > 0$, we know that $g(x)$ is convex. Thus, we get

$$\begin{aligned} f(\frac{p+q}{2}) &= \sum_{i=1}^N \frac{p_i + q_i}{2} \log(\frac{p_i + q_i}{2}) \\ &\le \sum_{i=1}^N \frac{p_i \log p_i + q_i \log q_i}{2} = \frac{f(p) + f(q)}{2} \end{aligned}$$

(b) Since the function g is convex, we know

$$\begin{aligned} \nabla g(x)^T (y - x) \le g(y) - g(x) \\ \nabla g(y)^T (x - y) \le g(x) - g(y) \end{aligned}$$

Thus, we obtain

$$\begin{aligned} (\nabla g(x) - \nabla g(y))^T (x - y) &= \nabla g(x)^T (x - y) - \nabla g(y)^T (x - y) \\ &\ge g(x) - g(y) - (g(x) - g(y)) = 0 \end{aligned}$$

4) **Fenchel.**

(a) The conjugate of f_α is defined as

$$f_\alpha^*(\theta) = \sup_{\mathbf{x}} \mathbf{x}^T \theta - f_\alpha(\mathbf{x}) = \alpha \left(\sup_{\mathbf{x}} \mathbf{x}^T \frac{\theta}{\alpha} - f(\mathbf{x})\right) = \alpha g \left(\frac{1}{\alpha}\theta\right).$$

(b) The conjugate of f is defined as
$$f^*(\theta) = \sup_x x\theta - \sqrt{1+x^2}.$$
Let $h(x,\theta) = x\theta - \sqrt{1+x^2}$. As h is strictly concave in x, $\frac{\partial h(x,\theta)}{\partial x}$ has at most one zero for a fixed θ. We have
$$\frac{\partial h(x,\theta)}{\partial x} = \theta - \frac{x}{\sqrt{1+x^2}}.$$
As $|\frac{x}{\sqrt{1+x^2}}| < 1$ for all $x \in \mathbb{R}$, consider the three cases:

- $|\theta| > 1$, then $h(x,\theta)$ is monotonic in x since $|\frac{\partial h(x,\theta)}{\partial x}| > |\theta| - 1 > 0$. Therefore $f^*(\theta)$ is not defined.
- $|\theta| < 1$, then the supremum is achieved where the gradient is zero, i.e., $x = \frac{\theta}{\sqrt{1-\theta^2}}$. Therefore we have $f^*(\theta) = -\sqrt{1-\theta^2}$.
- $|\theta| = 1$. For $\theta = 1$ the gradient approaches 0 as x goes to infinity, and hence
$$f^*(\theta) = \lim_{x\to\infty} x - \sqrt{1+x^2} = 0.$$
Similarly, we have $f^*(-1) = 0$.

To summerize, we have $f^*(\theta) = -\sqrt{1-\theta^2}, \theta \in [0,1]$.

5) **Hoeffding.** Often Hoeffding's Inequality is stated in a different way. Make sure to use Hoeffding to prove this version.

Let $X_1, \ldots, X_m$ be m independent random variables sampled from the same distribution D, where D has support on $[-1,1]$, and the mean of D is μ. Then for some $\alpha, \beta, \gamma > 0$ we have the following statement: with probability at least $1-\delta$
$$\left|\frac{1}{m}\sum_{i=1}^m X_i - \mu\right| \le \alpha\sqrt{\frac{\log(\beta/\delta)}{\gamma m}}.$$
When you solve this problem, make sure to get the best values of α, β, γ!

Since $\mathbb{E}[X_i] = \mu, X_i \in [-1,1]$, we know that $\mathbb{E}[X_i - \mu] = 0, X_i - \mu \in [-1-\mu, 1-\mu]$. Thus, by the Hoeffding's Inequality, we get
$$\begin{aligned}\Pr(\frac{1}{m}\sum_{i=1}^m X_i - \mu > \frac{t}{m}) &= \Pr(\sum_{i=1}^m (X_i - \mu) > t)\\ &\le \exp(-\frac{2t^2}{\sum_{i=1}^m (a_i - b_i)^2}) = \exp(-\frac{t^2}{2m})\end{aligned}$$
Similarly, we obtain
$$\Pr(\frac{1}{m}\sum_{i=1}^m X_i - \mu < -\frac{t}{m}) \le \exp(-\frac{t^2}{2m}) \tag{1}$$
Therefore, when $\exp(-\frac{t^2}{2m}) = \frac{\delta}{2}$, we get $\left|\frac{1}{m}\sum_{i=1}^m X_i - \mu\right| \le \frac{t}{m}$ with probability at least $1-\delta$. From $\exp(-\frac{t^2}{2m}) = \frac{\delta}{2}$, we represent t with δ.
$$t = \sqrt{2m\log\frac{2}{\delta}}$$

Therefore, we finially get

$$\frac{t}{m} = \sqrt{\frac{2}{m}\log\frac{2}{\delta}} = \alpha\sqrt{\frac{\log(\beta/\delta)}{\gamma m}}$$
$$\to \alpha = 2, \beta = 2, \gamma = 2$$

6) **Bayes classifier.**

(a) Recall that $\eta(x) = \Pr[Y = 1|X = x]$. Show that

$$\eta(x) = \frac{1}{1 + \exp(\frac{-x\mu}{\sigma^2})}.$$

Hint. Use Bayes' rule.

Using Bayes' rule we have

$$\eta(x) = \frac{\Pr[X = x|Y = 1]\Pr[Y = 1]}{\Pr[X = x]}. \tag{2}$$

Denote the pdf of the first Gaussian by f_1 and the second Gaussian by f_2. Then, $\Pr[Y = 1] = 1/2$, $\Pr[X = x|Y = 1] = f_1(x)$, and $\Pr[X = x] = f_1(x)/2 + f_2(x)/2$. Hence, $\eta(x) = f_1(x)/(f_1(x) + f_2(x))$. Substituting in for the Gaussian pdf,

$$\eta(x) = \frac{\exp\left(-\frac{1}{2}\left(\frac{x-\frac{\mu}{2}}{\sigma}\right)^2\right)}{\exp\left(-\frac{1}{2}\left(\frac{x-\frac{\mu}{2}}{\sigma}\right)^2\right) + \exp\left(-\frac{1}{2}\left(\frac{x+\frac{\mu}{2}}{\sigma}\right)^2\right)} \tag{3}$$

$$= \frac{1}{1 + \frac{\exp\left(-\frac{1}{2}\left(\frac{x+\frac{\mu}{2}}{\sigma}\right)^2\right)}{\exp\left(-\frac{1}{2}\left(\frac{x-\frac{\mu}{2}}{\sigma}\right)^2\right)}} \tag{4}$$

$$= \frac{1}{1 + \exp\left(\frac{-x\mu}{\sigma^2}\right)}. \tag{5}$$

(b) Compute an analytical expression for h_{Bayes} (*i.e.*, substitute $\eta(x)$ and simplify the resulting expression).

We have

$$h_{\text{Bayes}}(x) = \begin{cases} 1 & \frac{1}{1+\exp(\frac{-x\mu}{\sigma^2})} \geq 1/2, \\ -1 & \text{else.} \end{cases} \tag{6}$$

Simplifying the inequality, we find that it reduces to $x \geq 0$. So, $h_{\text{Bayes}}(x) = \text{sgn}(x)$.

(c) Compute the Bayes error $L(h_{\text{Bayes}})$. You can leave your answer in terms of the Gaussian cdf Φ.

We have

$$\begin{aligned} L(h_{\text{Bayes}}) &= \Pr[h_{\text{Bayes}}(x) \neq Y] \\ &= \Pr[\text{sgn}(x) \neq Y] \\ &= \Pr[x > 0 \text{ and } Y = -1] + \Pr[x < 0 \text{ and } Y = 1] \\ &= 2\Pr[x < 0 \mid Y = 1]\Pr[Y = 1] \\ &= \Pr[x < 0 \mid Y = 1] \\ &= \Phi\left(-\frac{\mu}{2\sigma}\right). \end{aligned}$$

(d) On which point(s) $x \in \mathbb{R}$ is h_{Bayes} most likely to make a mistake? Why?

h_{Bayes} is most likely to make a mistake on the point $x = 0$. This is because $\eta(0) = 1/2$, which means the value of Y is essentially a coin flip. Hence, h_{Bayes} will make a mistake with probability $1/2$ (which is the maximum probability for a mistake in binary classification).