Statistical learning theory

Joris LIMONIER

March 9, 2022

Contents

1 Inclass exercise January 12, 2022

1.1 Exercise 1

Show that

$$\mathbb{E}\left[\hat{\mathcal{R}}_S(h)\right] = \mathcal{R}_{D,f}(h) \tag{1}$$

$$\begin{aligned}
\mathbb{E}\left[\hat{\mathcal{R}}_S(h)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{h(x_i)\neq y_i}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbf{1}_{h(x_i)\neq y_i}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}\left(h(x_i)\neq y_i\right) \\
&= \frac{1}{n}n\mathbb{P}\left(h(x_i)\neq y_i\right) \\
&= \mathbb{P}\left(h(x_i)\neq y_i\right) \\
&= \mathbb{P}\left(h(x_i)\neq f(x)\right) \\
&= \mathcal{R}_{D,f}(h)
\end{aligned}$$

1.2 Exercise 2

We must prove that the variance of $\hat{\mathcal{R}}_S(h) \to 0$

$$\begin{aligned}
Var\left[\hat{\mathcal{R}}_S(h)\right] &= Var\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{h(x_i)\neq y_i}\right] \\
&= Var\frac{1}{n^2}\left[\sum_{i=1}^{n}\mathbf{1}_{h(x_i)\neq y_i}\right]
\end{aligned}$$

Let the Z_i be defined as follows:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{h(x_i)\neq f(x_i)} =: \frac{1}{n}\sum_{i=1}^{n}Z_i$$

(not finished, see lecture 1 slides)

2 Inclass exercise January 21, 2022

2.1 Exercise 1

Set $g(x) = \mathbb{P}(Y = 1 \mid X = x)$. We define the Bayes optimal predictor as:

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & g(x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

Question 1. Let $h : \mathcal{X} \to \{0, 1\}$ be a classifier. Show that

$$\begin{aligned}
&\mathbb{P}(h(X) \neq Y \mid X = x) \\
&= g(x) \cdot \mathbb{P}(h(X) = 0 \mid X = x)) + (1 - g(x)) \cdot \mathbb{P}(h(X) = 1 \mid X = x)
\end{aligned}$$

$$\begin{aligned}
&g(x) \cdot \mathbb{P}(h(X) = 0 \mid X = x)) + (1 - g(x)) \cdot \mathbb{P}(h(X) = 1 \mid X = x) \\
&= \mathbb{P}(Y = 1 \mid X = x) \cdot \mathbb{P}(h(X) = 0 \mid X = x)) \\
&+ (1 - \mathbb{P}(Y = 1 \mid X = x)) \cdot \mathbb{P}(h(X) = 1 \mid X = x) \\
&= \mathbb{P}(Y = 1 \cap h(X) = 0 \mid X = x)) \\
&+ \mathbb{P}(h(X) = 1 \mid X = x) - \mathbb{P}(Y = 1 \cap h(X) = 1 \mid X = x) \\
&= \mathbb{P}(Y = 1 \cap h(X) = 0 \mid X = x)) + \mathbb{P}(Y = 0 \cap h(X) = 1 \mid X = x) \\
&= \mathbb{P}(h(X) \neq Y \mid X = x)
\end{aligned}$$

Question 2. Deduce that

$$\mathbb{P}(f_D(X) \neq Y \mid X = x) = \min(g(x), 1 - g(x))$$

$$\begin{aligned}
&\mathbb{P}(f_D(X) \neq Y \mid X = x) \\
&= \begin{cases} \mathbb{P}(1 \neq Y \mid X = x), & g(x) \geq 1/2 \\ \mathbb{P}(0 \neq Y \mid X = x), & g(x) < 1/2 \end{cases} \\
&= \begin{cases} 1 - g(x), & g(x) \geq 1 - g(x) \\ g(x), & g(x) < 1 - g(x) \end{cases} \\
&= \min(g(x), 1 - g(x))
\end{aligned}$$

Question 3. Show that

$$\mathbb{P}(h(X) \neq Y \mid X = x) \geq \mathbb{P}(f_D(x) \neq Y \mid X = x)$$

$$\begin{aligned}
\mathbb{P}(f_D(x) \neq Y \mid X = x) &= \min(g(x), 1 - g(x)) \\
&= \min(g(x), 1 - g(x)) \\
&\cdot (\mathbb{P}(h(X) = 0 \mid X = x) + \mathbb{P}(h(X) = 1 \mid X = x)) \\
&\leq g(x) \cdot (\mathbb{P}(h(X) = 0 \mid X = x) \\
&+ (1 - g(x)) \cdot \mathbb{P}(h(X) = 1 \mid X = x)) \\
&= \mathbb{P}(h(X) \neq Y \mid X = x)
\end{aligned}$$

Question 4. Prove that

$$\mathcal{R}_{\mathcal{D}}(f_{\mathcal{D}}) \leq \mathcal{R}_{\mathcal{D}}(h)$$

$$\begin{aligned}
&\mathbb{P}(f_D(x) \neq Y \mid X = x) \leq \mathbb{P}(h(X) \neq Y \mid X = x) \\
\implies &\mathbb{E}\left[\mathbb{P}(f_D(x) \neq Y \mid X = x)\right] \leq \mathbb{E}\left[\mathbb{P}(h(X) \neq Y \mid X = x)\right] \\
\implies &\mathcal{R}_{\mathcal{D}}(f_{\mathcal{D}}) \leq \mathcal{R}_{\mathcal{D}}(h)
\end{aligned}$$

3 Inclass exercise January 28, 2022

3.1 Exercise 1

Let Z be a random variable with a second moment such that $\mathbb{E}[Z] = \mu$ and $\mathrm{Var}(Z) = \sigma^2$.

3.1.1 Question 1

Let $g : t \mapsto \mathbb{E}[(Z-t)^2]$. Show that g is minimum at $t = \mu$.

$$\begin{aligned}
g(t) &= \mathbb{E}\left[(Z-t)^2\right] \\
&= \mathbb{E}\left[Z^2 + t^2 - 2tZ\right] \\
&= \mathbb{E}\left[Z^2\right] + \mathbb{E}\left[t^2\right] - \mathbb{E}\left[2tZ\right] \\
&= \mathbb{E}\left[Z^2\right] + t^2 - 2t\mathbb{E}\left[Z\right] \\
&= \sigma^2 - \mu^2 + t^2 - 2t\mu \\
&= \sigma^2 - \mu^2 + t^2 - 2t\mu
\end{aligned}$$

We differentiate with respect to t:

$$\begin{aligned}
&\frac{\partial}{\partial t} g(t) = 0 \\
\implies &\frac{\partial}{\partial t}\left[\sigma^2 - \mu^2 + t^2 - 2t\mu\right] = 0 \\
\implies &2t - 2\mu = 0 \\
\implies &t = \mu
\end{aligned}$$

3.1.2 Question 2

Assume $Z \in [a, b]$ almost surely. Use the previous question to show that

$$\mathrm{Var}(Z) \leq \frac{(b-a)^2}{4}$$

$$
\begin{aligned}
& g(\mu) \leq g(t) \\
\implies & \operatorname{Var}(Z) \leq \mathbb{E}\left[(Z-t)^2\right] \\
\implies & \operatorname{Var}(Z) \leq \frac{1}{4}\mathbb{E}\left[(2Z-a-b)^2\right] \\
\implies & \operatorname{Var}(Z) \leq \frac{1}{4}\mathbb{E}\left[((Z-a)+(Z-b))^2\right] \\
\implies & \operatorname{Var}(Z) \leq \frac{1}{4}\mathbb{E}\left[((Z-a)-(b-Z))^2\right] \\
\implies & \operatorname{Var}(Z) \leq \frac{1}{4}\mathbb{E}\left[(|Z-a|-|b-Z|)^2\right] \\
\implies & \operatorname{Var}(Z) \leq \frac{1}{4}\mathbb{E}\left[|(Z-a)-(Z-b)|^2\right] \\
\implies & \operatorname{Var}(Z) \leq \frac{1}{4}\mathbb{E}\left[|b-a|^2\right] \\
\implies & \operatorname{Var}(Z) \leq \frac{(b-a)^2}{4}
\end{aligned}
$$

3.1.3 Question 3

Let $Z_1, \ldots, Z_n \sim Z$ be i.i.d. Use Chebyshev inequality to obtain a concentration inequality for

$$
Z := \frac{1}{n}\sum_{i=1}^{n} Z_i
$$

Chebyshev inequality:

$$
\mathbb{P}\left(|Z-\mathbb{E}[Z]| \geq a\right) \leq \frac{\operatorname{Var} Z}{a^2} \tag{2}
$$

$$
\begin{aligned}
\operatorname{Var}(Z) &= \operatorname{Var}\left(\frac{1}{n^2}\sum_{i=1}^{n} Z_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}(Z_i) && \textit{(Z_i's independent)} \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n} \frac{(b-a)^2}{4} \\
&\leq \frac{(b-a)^2}{4n}
\end{aligned}
$$

Then we apply (2):

$$\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq \varepsilon\right) \leq \frac{\operatorname{Var} Z}{\varepsilon^2}$$
$$\implies \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| \geq \varepsilon\right) \leq \frac{(b-a)^2}{4n\varepsilon^2}$$

4 In-class exercise February 4, 2022

4.1 Question 1

We define our loss as:

$$\ell(y, y') = |y - y'|$$

Show:

$$\forall\ c \in \mathbb{R},\ \begin{cases} |c| = \min_{a \geq 0} a \\ s.t. \quad a \geq c \\ \qquad\quad a \geq -c \end{cases}$$

A function study of $x \mapsto |x|$ gives the result.

4.2 Question 2

ERM consists in finding the following quantity:

$$\min_{w \in \mathbb{R}} \mathcal{R}_S(w) = \min_{w \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |\langle w_i x_i \rangle - y_i|$$

5 In-class exercise February 22, 2022

5.1 Question 1

Show that ERM with the logistic loss is equivalent to minimizing

$$F(w) = \sum_{i=1}^{n} \log\left(1 + \exp\left(-\tilde{y}_i \langle w, x_i \rangle\right)\right)$$

with $\tilde{y}_i = \operatorname{sign}(y_i - 0.5)$. Deduce that $\hat{\mathcal{R}}$ is a convex function of w.
We have:

$$
\begin{aligned}
\ell(y, y_i) &= \begin{cases} -\log(1-\hat{y}) & y = 0 \\ -\log(\hat{y}) & y = 1 \end{cases} \\
&= \begin{cases} -\log\left(1 - \frac{1}{1+e^{-w^T x_i}}\right) & y = 0 \\ -\log\left(\frac{1}{1+e^{-w^T x_i}}\right) & y = 1 \end{cases} \\
&= \begin{cases} -\log\left(\frac{e^{-w^T x_i}}{1+e^{-w^T x_i}}\right) & y = 0 \\ -\log\left(\frac{1}{1+e^{-w^T x_i}}\right) & y = 1 \end{cases} \\
&= \begin{cases} \log\left(\frac{1+e^{-w^T x_i}}{e^{-w^T x_i}}\right) & y = 0 \\ \log\left(1+e^{-w^T x_i}\right) & y = 1 \end{cases} \\
&= \begin{cases} \log\left(1+\frac{1}{e^{-w^T x_i}}\right) & y = 0 \\ \log\left(1+e^{-w^T x_i}\right) & y = 1 \end{cases} \\
&= \begin{cases} \log\left(1+e^{w^T x_i}\right) & y = 0 \\ \log\left(1+e^{-w^T x_i}\right) & y = 1 \end{cases} \\
&= \log\left(1+e^{-\tilde{y}_i w^T x_i}\right) \\
&= \log\left(1+e^{(-1)^{y_i} w^T x_i}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial^2}{\partial w^2} \log\left(1+e^{(-1)^{y_i} w^T x_i}\right) \\
&= \frac{\partial}{\partial w} \frac{-y_i x_i e^{-\tilde{y}_i w^T x_i}}{1+e^{-\tilde{y}_i w^T x_i}} \\
&= \frac{(y_i x_i)^2 e^{-\tilde{y}_i w^T x_i}(1+e^{-\tilde{y}_i w^T x_i}) - (-y_i x_i e^{-\tilde{y}_i w^T x_i})^2}{(1+e^{-\tilde{y}_i w^T x_i})^2} \\
&= \frac{x_i^2 e^{-\tilde{y}_i w^T x_i}(1+e^{-\tilde{y}_i w^T x_i}) - x_i^2 e^{-2\tilde{y}_i w^T x_i}}{(1+e^{-\tilde{y}_i w^T x_i})^2} \\
&= \frac{x_i^2 e^{-\tilde{y}_i w^T x_i}}{(1+e^{-\tilde{y}_i w^T x_i})^2} \\
&\geq 0
\end{aligned}
$$

5.2 Question 2

Compute the gradient of $\hat{\mathcal{R}}$ with respect to w.
Hint: show that $\phi'(z) = \phi(z)(1 - \phi(z))$

$$
\begin{aligned}
\frac{d}{dz}\phi(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\
&= -\frac{-e^{-z}}{(1+e^{-z})^2} \\
&= \frac{1-1+e^{-z}}{(1+e^{-z})^2} \\
&= \frac{1}{1+e^{-z}}\frac{-1+e^{-z}}{1+e^{-z}} \\
&= \frac{1}{1+e^{-z}}\frac{e^{-z}}{1+e^{-z}} \\
&= \frac{1}{1+e^{-z}}\frac{1+e^{-z}-1}{1+e^{-z}} \\
&= \phi(z)(1-\phi(z))
\end{aligned}
$$

$$
\begin{aligned}
\hat{\mathcal{R}}(w) &= \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) \\
&= \sum_{i=1}^{n} -(1-y)\log(1-\hat{y}_i) - y_i \log \hat{y}_i \\
&= \sum_{i=1}^{n} -(1-y)\log\left(1-\phi(w^T x_i)\right) - y_i \log \phi(w^T x_i)
\end{aligned}
$$

For some $1 \leq j \leq n$:

$$
\frac{\partial}{\partial w_j}\ell(y, \hat{y}) = \frac{\partial}{\partial w_j} - (1-y)\log\left(1-\phi(w^T x_i)\right) - y_i \log \phi(w^T x_i)
$$

Final result:

$$
\hat{\mathcal{R}} = \sum_{i=1}^{n} \left(\phi(w^T x_i) - y_i\right) x_{ij}
$$

6 In class exercises February 25, 2022

6.1 Question 1

Prove that the following function is a kernel:

$$k(x, y) = 2^{x+y}$$

Symmetry Trivial

Positive definiteness

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j 2^{x_i + x_j} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j 2^{x_i} 2^{x_j} \\
&= \sum_{i=1}^{n} c_i 2^{x_i} \sum_{j=1}^{n} c_j 2^{x_j} \\
&= \left[\sum_{i=1}^{n} c_i 2^{x_i}\right]^2 \\
\geq 0
\end{aligned}$$

6.2 Question 2

Prove that the following function is a kernel:

$$k(x, y) = (x^T y)^2$$

Symmetry Trivial

Positive definiteness

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j (x_i^T x_j)^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j (x_i^T x_j)(x_i^T x_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j (x_j^T x_i)(x_i^T x_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \operatorname{tr}(x_j^T x_i x_i^T x_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \operatorname{tr}(x_i x_i^T x_j x_j^T) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \langle x_i x_i^T, x_j x_j^T \rangle \\
&= \sum_{i=1}^{n} c_i x_i x_i^T \sum_{j=1}^{n} c_j x_j x_j^T \\
&= \left[\sum_{i=1}^{n} c_i x_i x_i^T\right]^2 \\
&\geq 0
\end{aligned}$$

6.3 Question 3

Prove that the following function is a kernel:

$$k(x, y) = \cos(x - y)$$

Symmetry Trivial

Positive definiteness

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \cos(x_i - x_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \left[\cos(x_i)\cos(x_j) + \sin(x_i)\sin(x_j)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \cos(x_i)\cos(x_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \sin(x_i)\sin(x_j) \\
&= \left[\sum_{i=1}^{n} c_i \cos(x_i)\right]^2 + \left[\sum_{i=1}^{n} c_i \sin(x_i)\right]^2 \\
&\geq 0
\end{aligned}$$

6.4 Exercise

Show that the Gaussian kernel, which is given by:

$$k(x, y) := \exp\left(-\frac{\|x-y\|^2}{2\nu^2}\right)$$

is actually a kernel.
We have that:

$$\exp\left(-\frac{\|x-y\|^2}{2\nu^2}\right) = \lim_{n\to\infty}\sum_{p=1}^{n}\frac{1}{p!}\left(\frac{-\|x-y\|^2}{2\nu^2}\right)^p$$

$$\begin{aligned}
&\|x-y\|^2 = (x-y)^T(x-y) \\
\implies &\|x-y\|^2 = x^T x - x^T y - y^T x + y^T y \\
\implies &\|x-y\|^2 = \|x\|^2 - 2x^T y + \|y\|^2 \\
\implies &-\|x-y\|^2 = \underbrace{2x^T y}_{\text{kernel}} - \|x\|^2 - \|y\|^2
\end{aligned}$$

Let us show that $\exp\left(-\|x\|^2 - \|y\|^2\right)$ is a kernel:

$$\begin{aligned}
\exp\left(-\|x\|^2 - \|y\|^2\right) &= \exp\left(-\|x\|^2\right)\exp\left(-\|y\|^2\right) \\
&= \left\langle \exp\left(-\|x\|^2\right), \exp\left(-\|y\|^2\right)\right\rangle
\end{aligned}$$

7 In-class exercises March 9, 2022

7.1 Exercise

Let $S = \{x_1, \ldots, x_n\}$ be a finite set of points in $\mathcal{X}$. Compute the distance to the barycenter of S in the RKHS.
The distance to the barycenter is given by:

$$
\begin{aligned}
d^2 &:= \left\| \Phi(x_{i_0}) - \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \right\|^2 \\
&= \left\langle \Phi(x_{i_0}) - \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i), \Phi(x_{i_0}) - \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \right\rangle \\
&= \langle \Phi(x_{i_0}), \Phi(x_{i_0}) \rangle + \left\langle \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i), \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \right\rangle - 2 \left\langle \Phi(x_{i_0}), \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \right\rangle \\
&= \langle k(x_{i_0}), k(x_{i_0}) \rangle + \frac{1}{n^2} \left\langle \sum_{i=1}^{n} \Phi(x_i), \sum_{i=1}^{n} \Phi(x_i) \right\rangle - 2 \left\langle \Phi(x_{i_0}), \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \right\rangle
\end{aligned}
$$

7.2 Exercise

Show that for any matrices A, B, $A\left(BA + \gamma I\right)^{-1} = \left(AB + \gamma I\right)^{-1} A$.

$$
\begin{aligned}
& A\left(BA + \gamma I\right)^{-1} = \left(AB + \gamma I\right)^{-1} A \\
\implies & \left(AB + \gamma I\right) A = A\left(BA + \gamma I\right) \\
\implies & ABA + \gamma A = ABA + \gamma A
\end{aligned}
$$

The linear kernel is given by $k(x, y) = x^T y$, so K