IEMS 402 Statistical Learning - 2025 Winter
HW1

Tuorui Peng[1]

Exercise 1 Design of Loss Function

1.(a)

Consider decomposing $m(X)$ as follows:

$$m(X) := \mathbb{E}\left[Y|X=x\right] + \delta(X),$$

i.e. with $\delta(X)$ being the deviation of the true conditional expectation from the model. Then the expected ℓ_2 error can be written as:

$$\begin{aligned}
\mathbb{E}\left[(Y-m(X))^2\right] =& \mathbb{E}\left[(Y-\mathbb{E}\left[Y|X=x\right]-\delta(X))^2\right]\\
=& \mathbb{E}\left[(Y-\mathbb{E}\left[Y|X=x\right])^2\right] + \mathbb{E}\left[\delta^2(X)\right] - 2\mathbb{E}\left[(Y-\mathbb{E}\left[Y|X=x\right])\delta(X)\right]\\
\overset{(i)}{=}& \mathbb{E}\left[(Y-\mathbb{E}\left[Y|X=x\right])^2\right] + \mathbb{E}\left[\delta^2(X)\right]\\
\overset{(ii)}{\geq}& \mathbb{E}\left[(Y-\mathbb{E}\left[Y|X=x\right])^2\right].
\end{aligned}$$

Thus proved that the expected ℓ_2 error is minimized by the conditional expectation $\mathbb{E}\left[Y|X=x\right]$. Here in the proof, (i) is due to the fact that $\mathbb{E}_{Y|X=x}\left[Y-\mathbb{E}\left[Y|X=x\right]\big|X=x\right]=0$, and (ii) is due to the non-negativity of $\mathbb{E}\left[\delta^2(X)\right]$, and equality holds if and only if $\delta(X)=0$ almost surely.

1.(b)

The expected ℓ_1 error can be written as:

$$\mathbb{E}\left[|Y-m(X)|\right] = \mathbb{E}_X\left[\int_{Y|X=x} |Y-m(x)|\,\mathrm{d}F(y|X=x)\right]$$

taking variation with respect to $m(X)$, we have (here δ refers to the variation operator):

$$\begin{aligned}
\delta\mathbb{E}\left[|Y-m(X)|\right] =& \mathbb{E}_X\left[\int_{Y|X=x} \delta m(x)\cdot \operatorname{sgn}(Y-m(x))\,\mathrm{d}F(y|X=x)\right]\\
=& \mathbb{E}_X\left[\delta m(x)\int_{Y|X=x} \operatorname{sgn}(Y-m(x))\,\mathrm{d}F(y|X=x)\right]
\end{aligned}$$

To minimize the expected ℓ_1 error, we requre the variation taking value of zero, i.e.:

$$0 = \delta\mathbb{E}\left[|Y-m(X)|\right] = \mathbb{E}_X\left[\delta m(x)\int_{Y|X=x} \operatorname{sgn}(Y-m(x))\,\mathrm{d}F(y|X=x)\right],\quad \forall \delta m(\,\cdot\,)$$

which requires choosing $m(\,\cdot\,)$ s.t. $\int_{Y|X=x}\operatorname{sgn}(Y-m(x))\,\mathrm{d}F(y|X=x) = \mathbb{E}_{Y|X=x}\left[\operatorname{sgn}(Y-m(x))|X=x\right] = 0$ almost surely. This is equivalent to using the conditional median as the prediction function $m(x) = \operatorname{median}(Y|X=x)$.

[1] TuoruiPeng2028@u.northwestern.edu

1.(c)

Similarly we write the following differentiation w.r.t. β

$$\frac{\partial}{\partial\beta}\mathbb{E}\left[(Y-\beta' X)^2\right] = -\,\mathbb{E}\left[-2X(Y-\beta' X)\right]$$

and set it to zero, we have:

$$\begin{aligned} 0 =& -\,\mathbb{E}\left[-2X(Y-\beta' X)\right] \\ &\Rightarrow \beta_* = \mathbb{E}\left[XX'\right]^{-1}\mathbb{E}\left[XY\right] \end{aligned}$$

1.(d)

We consider the following function:

$$s_\alpha(y,\hat{y}) := \begin{cases} \alpha, & \text{if } y-\hat{y}>0 \\ \alpha-1, & \text{if } y-\hat{y}<0 \end{cases}$$

and notice that $s_\alpha(y,\hat{y}) = \dfrac{\partial}{\partial y}\alpha\cdot\mathrm{sgn}(y-\hat{y})$[2]. Then we can write the deviation of expected ℓ_α error as:

$$\delta\mathbb{E}\left[\rho_\alpha(y,m(x))\right] = \mathbb{E}_X\left[\delta m(x)\int_{Y|X=x} s_\alpha(y,m(x))\,\mathrm{d}F(y|X=x)\right]$$

minimizing the ρ_α loss function requires the variation to be zero for any δm, i.e.:

$$0 = \mathbb{E}_{Y|X=x}\left[s_\alpha(y,m(x))|X=x\right] \Rightarrow m(x) = q_\alpha(x)$$

where $q_\alpha(x)$ is the α-quantile of Y given $X=x$.

Exercise 2 Central Limit Theorem

2.(a)

First by SLLN we definitely have $\bar{X}\xrightarrow{\text{a.s.}}\mathbb{E}\left[X\right]=\mu$.

Then we re-write the expression of s_n^2 as:

$$\begin{aligned} s_n^2 =& \frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2 \\ =& \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 \end{aligned}$$

[2]Omitting the discontinuity at $y-\hat{y}=0$, which won't be a big problem if we have continuous and strictly increasing loss function.

Then we notice that:

$$\text{by LLN: } \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathbb{E}\left[X^2\right]$$

$$\text{by LLN and continuous mapping theorem: } \bar{X}^2 \xrightarrow{p} (\mathbb{E}[X])^2$$

thus we have by slutsky's theorem:

$$s_n^2 \xrightarrow{\mathrm{d}} \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \sigma^2$$

2.(b)

Note that $s_n^2 \xrightarrow{\mathrm{d}} \sigma^2$ which is a constant, thus we also have $s_n \xrightarrow{\mathrm{p}} \sigma$. Then by Slutsky's theorem and CLT, we have the following:

$$\begin{aligned}
&\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathrm{d}} N(0,1)\\
&\frac{s_n}{\sigma} \xrightarrow{\mathrm{p}} 1\\
\Rightarrow &\frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \xrightarrow{\mathrm{d}} N(0,1)
\end{aligned}$$

Exercise 3 Curse of Dimensionality: Asymptotic scaling of nearest neighbor distances

3.(a)

$$\begin{aligned}
\mathbb{P}\left(\left\|x_{i(X_0)} - x_0\right\| > \delta\right) =&\mathbb{P}\left(\bigcap_{i=1}^{n}\{\|x_i - x_0\|_2 > \delta\}\right)\\
=&\int \mathrm{d}P_{x_0}\int \mathrm{d}P_{x_1^n}\prod_{i=1}^{n}\mathbf{1}_{\|x_i - x_0\|_2 > \delta}\\
=&\int \left(1 - P\left(B_d(x,\delta)\right)\right)^n \mathrm{d}P(x)
\end{aligned}$$

3.(b)

We can construct the partition as follows: at each dimention, construct cutting points $\{-kr, (-k+1)r, \ldots, (k-1)r, kr\}$ where k chosen s.t. $k = \lceil\frac{R}{\delta/d}\rceil \leq 2Rd/\delta$ and $r = \delta/d$. And each $U_\cdot$ is constructed by the combination of the cutting points. Then we have number of partition

$$N(\delta) =(2k+1)^d \leq \frac{8(Rd)^d}{\delta^d} = \frac{c}{\delta^d}$$

in this way, each "block" of the partition is at most a hypercube with side length r, and diameter diam $= r\sqrt{d} < \delta$.

3.(c)

Since the partition $U_1^{N(\delta)}$ has diameter at most δ, we consider THE block $U_\imath$ that contains x for each given x. Then we have:

$$U_\imath \subseteq B_d(x,\delta) \Rightarrow P\left(U_\imath\right) \leq P\left(B_d(x,\delta)\right)$$

thus we have:

$$\begin{aligned}
\mathbb{P}\left(\left\|x_{i(X_0)} - x_0\right\| > \delta\right) &= \int \left(1 - P\left(B_d(x,\delta)\right)\right)^n \mathrm{d}P(x) \\
&\leq \sum_{i=1}^{N(\delta)} \int_{U_i} \left(1 - P\left(P\left(U_i\right)\right)\right)^n \mathrm{d}P(x) \\
&= \sum_{i=1}^{N(\delta)} \left(1 - P\left(P\left(U_i\right)\right)\right)^n P\left(U_i\right) \\
&\overset{(i)}{\leq} \frac{c}{en\delta^d}.
\end{aligned}$$

Thus finished the proof. Here in the proof, (i) is due to the fact that $x \mapsto x(1-x)^n$ reaches maximum at $x = 1/(n+1)$, with maximum value

$$\frac{1}{n+1}(1 - \frac{1}{n+1})^n = \frac{1}{n}\left(1 - \frac{1}{n+1}\right)^{n+1} \leq \frac{1}{en}.$$

3.(d)

With the probabilistic bound, we note that to maintain the bound at $O(1)$, we should choose $\delta \asymp n^{-1/d}$ (so that $c/en\delta^d = O(1)$). Which indicates that

$$\mathbb{P}\left(\left\|x_{i(X_0)} - x_0\right\| \lesssim n^{-1/d}\right) \geq 1 - C$$

i.e. with certain minimal probability, the nearest neighbor distance is at most $\lesssim n^{-1/d}$.

Exercise 4

4.(a)

Note that $f_\theta(x) = 0$ is a hyper plane in $\mathbb{R}^d$, the distance from $x^{(i)}$ to which is

$$\begin{aligned}
\text{distance} =& \frac{\left|\theta' x^{(i)} + \theta_0\right|}{\|\theta\|} \\
=& \begin{cases} \dfrac{\left|\theta' x^{(i)} + \theta_0\right|}{\|\theta\|}, & \text{if } \theta' x^{(i)} + \theta_0 > 0 \\ -\dfrac{\left|\theta' x^{(i)} + \theta_0\right|}{\|\theta\|}, & \text{if } \theta' x^{(i)} + \theta_0 < 0 \end{cases}
\end{aligned}$$

further for hard margin SVM, $\theta' x^{(i)} + \theta_0$ has the same sign as $y^{(i)}$, thus we have:

$$\text{distance} = \frac{y^{(i)}(\theta' x^{(i)} + \theta_0)}{\|\theta\|} = \gamma^(i)$$

4.(b)

Optimization problem for hard margin SVM can be written as:

$$\begin{aligned}\arg\min_{\theta,\theta_0} & \frac{1}{2}\left\|\theta\right\|^2 \\ & w.r.t.y^{(i)}(\theta' x^{(i)} + \theta_0) \geq 1\end{aligned}$$

and note that the decision boundary is determined by the (θ, θ_0), which has an extra degree of freedom w.r.t. scale transformation. We cancel this degree of freedom by setting a constraint $\|\theta\| = \dfrac{1}{M}$.

The Lagrangian can be written as:

$$\mathcal{L}(\theta, \theta_0, \alpha_1^n) = \frac{1}{2}\left\|\theta\right\|^2 - \sum_{i=1}^{n} \alpha_i \left(y^{(i)}(\theta' x^{(i)} + \theta_0) - 1\right)$$

and the optimization problem can be solved by minimizing the Lagrangian w.r.t. θ, θ_0

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \theta} &= \theta - \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} = 0 \\ \frac{\partial \mathcal{L}}{\partial \theta_0} &= -\sum_{i=1}^{n} \alpha_i y^{(i)} = 0\end{aligned}$$

to get the dual problem:

$$\begin{aligned}\theta_D(\alpha) &= -\frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)\prime} x^{(j)} + \sum_{i=1}^{n} \alpha_i \\ s.t.\alpha_i &\geq 0, \sum_{i=1}^{n} \alpha_i y^{(i)} = 0\end{aligned}$$

The dual problem is usually easier to solve because we can see that the dual problem has less (non-trivial) constraints.