Machine Learning Techniques, Spring 2020, HW3

B07902064 資工二 蔡銘軒

June 27, 2020

Problem 1

$$\begin{aligned}
E(s_j^{(\ell)}|\mathbf{x}^{(\ell-1)}) =& E(\sum_{i=0}^{d^{(\ell-1)}} w_{ij}^{(\ell)} x_i^{(\ell-1)}|\mathbf{x}^{(\ell-1)}) \\
=& \sum_{i=0}^{d^{(\ell-1)}} E(w_{ij}^{(\ell)}|\mathbf{x}^{(\ell-1)})E(x_i^{(\ell-1)}|\mathbf{x}^{(\ell-1)}) \text{ Independence between } w_{ij}^{(\ell)}, x_i^{(\ell-1)} \\
=& \sum_{i=0}^{d^{(\ell-1)}} E(w_{ij}^{(\ell)})E(x_i^{(\ell-1)}|\mathbf{x}^{(\ell-1)}) \text{ Independence between } w_{ij}^{(\ell)}, \mathbf{x}^{(\ell-1)} \\
=& \sum_{i=0}^{d^{(\ell-1)}} 0 \cdot E(x_i^{(\ell-1)}|\mathbf{x}^{(\ell-1)}) \\
=& 0
\end{aligned}$$

Intuitively, with $\mathbf{x}^{(\ell-1)}$ given as the condition, the dependence of each $s_j^{(\ell)}$ is determined by $w_{ij}^{(\ell)}$. As $w_{ij}^{(\ell)}$ are independent to one another and to all $x_i^{(\ell-1)}$, we conclude that $s_j^{(\ell)}$ are independent to one another as well.

Problem 2

We know for two independent random variables X, Y, we have

$$\text{Var}(XY) = E(X^2Y^2) - (E(XY))^2 = \text{Var}(X)\text{Var}(Y) + \text{Var}(X)(E(Y))^2 + \text{Var}(Y)(E(X))^2$$

Then we have

$$\begin{aligned}
\text{Var}(s_j^{(\ell)}) =& \sum_{i=0}^{d^{(\ell-1)}} \text{Var}(w_{ij}^{(\ell)} x_i^{(\ell-1)}) \\
=& \sum_{i=0}^{d^{(\ell-1)}} \text{Var}(w_{ij}^{(\ell)})\text{Var}(x_i^{(\ell-1)}) + \text{Var}(w_{ij}^{(\ell)})(E(x_i^{(\ell-1)}))^2 + \text{Var}(x_i^{(\ell-1)})(E(w_{ij}^{(\ell)}))^2 \\
=& \sum_{i=0}^{d^{(\ell-1)}} \sigma_w^2\sigma_x^2 + \sigma_w^2\bar{x}^2 + \sigma_x^2 0 \\
=& d^{(\ell-1)} \cdot \sigma_w^2(\sigma_x^2 + \bar{x}^2)
\end{aligned}$$

Problem 3

Let $f_i^{(\ell-1)}(s)$ be the **probability density function (pdf)** of random variable $s_i^{(\ell-1)}$. We have

$$E\left[(s_i^{(\ell-1)})^2\right] = \int_{-\infty}^{\infty} s^2 \cdot f_i^{(\ell-1)}(s)ds$$

and

$$\begin{aligned} E\left[(x_i^{(\ell-1)})^2\right] &= \int_{-\infty}^{\infty} (\max(s,0))^2 f_i^{(\ell-1)}(s)ds \\ &= \int_{-\infty}^{0} 0^2 f_i^{(\ell-1)}(s)ds + \int_{0}^{\infty} s^2 f_i^{(\ell-1)}(s)ds \\ &= \int_{0}^{\infty} s^2 f_i^{(\ell-1)}(s)ds \end{aligned}$$

Since $s_j^{(\ell-1)}$ are zero-mean and symmetric, we have

$$\begin{aligned} & f_i^{(\ell-1)}(s) = f_i^{(\ell-1)}(-s) \\ \Longrightarrow & s^2 f_i^{(\ell-1)}(s) = s^2 f_i^{(\ell-1)}(-s) \\ \Longrightarrow & \int_{0}^{\infty} s^2 f_i^{(\ell-1)}(s)ds = \frac{1}{2}\int_{-\infty}^{\infty} s^2 f_i^{(\ell-1)}(s)ds \end{aligned}$$

This shows $E\left[(x_i^{(\ell-1)})^2\right] = \frac{1}{2}E\left[(s_i^{(\ell-1)})^2\right]$

Problem 4

We have $\mathrm{Var}(s_i^{(\ell-1)}) = E\left[(s_i^{(\ell-1)})^2\right] - \left(E\left[s_i^{(\ell-1)}\right]\right)^2$.

From **Problem 3**, $E\left[(s_i^{(\ell-1)})^2\right] = 2E\left[(x_i^{(\ell-1)})^2\right] = 2(\sigma_x^2 + \bar{x}^2)$, and $\left(E\left[s_i^{(\ell-1)}\right]\right)^2 = 0$ as given by the problem description. Using the result in **Problem 2**, we have

$$\begin{aligned} \mathrm{Var}(s_j^{(\ell)}) &= d^{(\ell-1)}\sigma_w^2(\sigma_x^2 + \bar{x}^2) \\ &= \frac{d^{(\ell-1)}}{2}\sigma_w^2 \cdot 2(\sigma_x^2 + \bar{x}^2) \\ &= \frac{d^{(\ell-1)}}{2}\sigma_w^2 \mathrm{Var}(s_i^{(\ell-1)}) \end{aligned}$$

Problem 5

Since we are using leaky ReLU, we assume that $0 < a < 1$. Using the notations and assumptions in **Problem 1** to **Problem 3**, we have

$$\begin{aligned} E\left[(x_i^{(\ell-1)})^2\right] &= \int_{-\infty}^{\infty} (\max(s, a\cdot s))^2 f_i^{(\ell-1)}(s)ds \\ &= \int_{-\infty}^{0} (a\cdot s)^2 f_i^{(\ell-1)}(s)ds + \int_{0}^{\infty} s^2 f_i^{(\ell-1)}(s)ds \\ &= (a^2+1)\int_{0}^{\infty} s^2 f_i^{(\ell-1)}(s)ds \\ &= \frac{(a^2+1)}{2}E\left[(s_i^{(\ell-1)})^2\right] \end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(s_j^{(\ell)}) &= d^{(\ell-1)}\sigma_w^2(\sigma_x^2 + \bar{x}^2)\\ &= d^{(\ell-1)}\sigma_w^2 \cdot \frac{a^2+1}{2} \cdot \frac{2}{a^2+1}(\sigma_x^2 + \bar{x}^2)\\ &= d^{(\ell-1)}\sigma_w^2 \cdot \frac{a^2+1}{2} \cdot \mathrm{Var}(s_i^{(\ell-1)})\end{aligned}$$

To make $\mathrm{Var}(s_j^{(\ell)}) = \mathrm{Var}(s_i^{(\ell-1)})$, we need $d^{(\ell-1)}\sigma_w^2 \cdot \frac{a^2+1}{2} = 1$. That is, $\sigma_w^2 = \frac{2}{d^{(\ell-1)}(a^2+1)}$. And to derive the result above, we require $w_{ij}^{(\ell)}$ to be zero-mean. We can use a normal distribution $N(0, \frac{2}{d^{(\ell-1)}(a^2+1)})$ as the initialization scheme.

Problem 6

We expand the equation

$$\begin{aligned}\mathbf{v}_T =& \beta\mathbf{v}_{T-1} + (1-\beta)\mathbf{\Delta}_T\\ =& \beta(\beta\mathbf{v}_{T-2} + (1-\beta)\mathbf{\Delta}_{T-1}) + (1-\beta)\mathbf{\Delta}_T = \beta^2\mathbf{v}_{T-2} + \beta(1-\beta)\mathbf{\Delta}_{T-1} + (1-\beta)\mathbf{\Delta}_T\\ =& \beta^2(\beta\mathbf{v}_{T-3} + (1-\beta)\mathbf{\Delta}_{T-2}) + \beta(1-\beta)\mathbf{\Delta}_{T-1} + (1-\beta)\mathbf{\Delta}_T\\ =& \dots\end{aligned}$$

It is easy to see the pattern that $\alpha_t = \beta^{T-t}(1-\beta)$

Problem 7

$$\begin{aligned}&\beta^{T-1}(1-\beta) \le \frac{1}{2}\\ \Longrightarrow& \beta^{T-1} \le \frac{1}{2(1-\beta)}\\ \Longrightarrow& T-1 \ge log_\beta(\frac{1}{2(1-\beta)})\\ \Longrightarrow& T \ge log_\beta(\frac{\beta}{2(1-\beta)})\end{aligned}$$

This shows the smallest T is $\max(1, \lceil log_\beta(\frac{\beta}{2(1-\beta)})\rceil)$

Problem 8

$$\begin{aligned}\alpha'_t &= \frac{\alpha_t}{\sum_{t=1}^T \alpha_t}\\ &= \frac{\beta^{T-t}(1-\beta)}{\sum_{t=1}^T \beta^{T-t}(1-\beta)}\\ &= \frac{\beta^{T-t}}{\frac{1-\beta^T}{1-\beta}}\\ &= \frac{\beta^{T-t}(1-\beta)}{1-\beta^T}\end{aligned}$$

Problem 9

$$
\begin{aligned}
&\frac{\beta^{T-1}(1-\beta)}{1-\beta^T} \le \frac{1}{2} \\
\Longrightarrow &\beta^{T-1} - \frac{1}{2}\beta^T \le \frac{1}{2} \\
\Longrightarrow &\frac{2-\beta}{2\beta}\beta^T \le \frac{1}{2} \\
\Longrightarrow &\beta^T \le \frac{\beta}{2-\beta} \\
\Longrightarrow &T \ge log_\beta \frac{\beta}{2-\beta}
\end{aligned}
$$

Since $log_\beta \frac{\beta}{2-\beta} \ge 1$ for all $\beta > 0$, so the the smallest T is $\lceil log_\beta \frac{\beta}{2-\beta} \rceil$

Problem 10

We assume the dimension of each matrix or vector as follows

- $\mathbf{w}$ and $\mathbf{p}$: $D \times 1$
- X: $N \times D$
- $\mathbf{y}$: $N \times 1$

For some $N, D \in \mathbb{N}$. Then we have

$$
\begin{aligned}
E_{\mathbf{p}}(\|\mathbf{y} - \mathrm{X}(\mathbf{w} \odot \mathbf{p})\|^2) =& E_{\mathbf{p}}((\mathbf{y} - \mathrm{X}(\mathbf{w} \odot \mathbf{p}))^T(\mathbf{y} - \mathrm{X}(\mathbf{w} \odot \mathbf{p}))) \\
=& E_{\mathbf{p}}((\mathbf{y}^T - (\mathbf{w} \odot \mathbf{p})^T \mathrm{X}^T)(\mathbf{y} - \mathrm{X}(\mathbf{w} \odot \mathbf{p})) \\
=& \mathbf{y}^T\mathbf{y} - \frac{1}{2}\mathbf{y}^T\mathrm{X}\mathbf{w} - \frac{1}{2}\mathbf{w}^T\mathrm{X}^T\mathbf{y} + E_{\mathbf{p}}((\mathbf{w} \odot \mathbf{p})^T \mathrm{X}^T \mathrm{X}(\mathbf{w} \odot \mathbf{p}))
\end{aligned}
$$

Note that $\mathbf{y}^T\mathrm{X}\mathbf{w}$ and $\mathbf{w}^T\mathrm{X}^T\mathbf{y}$ are the same scaler. Let $Z = \mathrm{X}^T\mathrm{X}$, we have

$$
\begin{aligned}
E_{\mathbf{p}}(\|\mathbf{y} - \mathrm{X}(\mathbf{w} \odot \mathbf{p})\|^2) =& \mathbf{y}^T\mathbf{y} - \mathbf{w}^T\mathrm{X}^T\mathbf{y} + E_{\mathbf{p}}(\sum_{i=1}^{D}\sum_{j=1}^{D} w_i p_i Z_{ij} w_j p_j) \\
=& \mathbf{y}^T\mathbf{y} - \mathbf{w}^T\mathrm{X}^T\mathbf{y} + \sum_{i=1}^{D} E_{\mathbf{p}}(p_i^2) w_i^2 Z_{ii} + \sum_{i,j,i \ne j}^{D} E_{\mathbf{p}}(p_i p_j) w_i w_j Z_{ij} \\
=& \mathbf{y}^T\mathbf{y} - \mathbf{w}^T\mathrm{X}^T\mathbf{y} + \frac{1}{2}\sum_{i=1}^{D} w_i^2 Z_{ii} + \frac{1}{4}\sum_{i,j,i \ne j}^{D} w_i w_j Z_{ij} \\
=& \mathbf{y}^T\mathbf{y} - \mathbf{w}^T\mathrm{X}^T\mathbf{y} + \frac{1}{4}\sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j Z_{ij} + \frac{1}{4}\sum_{i=1}^{D} w_i^2 Z_{ii} \\
=& \mathbf{y}^T\mathbf{y} - \mathbf{w}^T\mathrm{X}^T\mathbf{y} + \frac{1}{4}\mathbf{w}^T Z \mathbf{w} + \frac{1}{4}\sum_{i=1}^{D} w_i^2 Z_{ii}
\end{aligned}
$$

Let

$$
\tilde{Z} = \begin{bmatrix} Z_{11} & 0 & \dots & 0 \\ 0 & Z_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Z_{dd} \end{bmatrix}
$$

Then we have

$$E_{\mathbf{p}}(\|\mathbf{y} - \mathrm{X}(\mathbf{w} \odot \mathbf{p})\|^2) = \mathbf{y}^T\mathbf{y} - \mathbf{w}^T\mathrm{X}^T\mathbf{y} + \frac{1}{4}\mathbf{w}^T Z\mathbf{w} + \frac{1}{4}\mathbf{w}^T\tilde{Z}\mathbf{w}$$

To find the optimal $\mathbf{w}$, consider $\frac{\partial E_{\mathbf{p}}(\|\mathbf{y}-\mathrm{X}(\mathbf{w}\odot\mathbf{p})\|^2)}{\partial \mathbf{w}} = 0$, we have

$$\begin{aligned} &-\mathbf{y}^T\mathrm{X} + \frac{1}{4}\mathbf{w}^T(Z + Z^T) + \frac{1}{4}\mathbf{w}^T(\tilde{Z} + \tilde{Z}^T) = 0 \\ &\Longrightarrow \mathbf{w} = 4(\mathbf{y}^T\mathrm{X}(Z + Z^T + \tilde{Z} + \tilde{Z}^T)^\dagger)^T \end{aligned}$$

To simplify the result, we note the following properties

- Both Z and $\tilde{Z}$ are symmetric
- Since $(Z + \tilde{Z})$ is symmetric, its pseudo inverse $(Z + \tilde{Z})^\dagger$ is symmetric as well

we simplify the result as

$$\begin{aligned} \mathbf{w} &= 4(\mathbf{y}^T\mathrm{X}(2 \cdot (Z + \tilde{Z}))^\dagger)^T \\ &= 2(\mathbf{y}^T\mathrm{X}(Z + \tilde{Z})^\dagger)^T \\ &= 2(Z + \tilde{Z})^\dagger\mathrm{X}^T\mathbf{y} \end{aligned}$$

Problem 11

Let M_i be the set of data g_i makes mistakes on.

- **Minimum:** Consider the case where M_1, M_2, M_3 are disjoint. That is, no more than one classifier makes a mistake on a piece of data. This is possible given the E_{out} of g_1, g_2, g_3. In this case, uniform blending yields $E_{out}(G) = 0$.
- **Maximum:** Consider the case where $M_1 \subset M_3$, $M_2 \subset M_3$ and $M_1 \cap M_2 = \emptyset$. In this case, when either g_1 or g_2 makes a mistake, g_3 is guaranteed to err as well. This yields $E_{out}(G) = 0.08 + 0.16 = 0.24$.

So we have $0 \leq E_{out}(G) \leq 0.24$.

Problem 12

A uniform blending classifier G makes a mistake only if no less than $\frac{K+1}{2}$ of the K classifiers $\{g_k\}_{k=1}^K$ make the same mistake, where K is an odd integer. For a blending classifier G with $E_{out}(G)$, the number of mistakes made by $\{g_k\}_{k=1}^K$ on a set of N data is at least $N \cdot E_{out}(G) \cdot \frac{K+1}{2}$, and it is bounded by the total number of mistakes made by $\{g_k\}_{k=1}^K$. So we have

$$\begin{aligned} N \cdot E_{out}(G) \cdot \frac{K+1}{2} &\leq \sum_{k=1}^{K} N \cdot e_k \\ \Longrightarrow E_{out}(G) &\leq \frac{2}{K+1}\sum_{k=1}^{K} e_k \end{aligned}$$

which proves the given statement.

Problem 13

The probability of an example being sampled in an operation is $\frac{1}{N}$. It follows that the probability of an example not being sampled at least once after pN operations is

$$(1 - \frac{1}{N})^{pN} = ((1 - \frac{1}{N})^N)^p$$

As N is **very large**, we have

$$\lim_{N\to\infty}((1 - \frac{1}{N})^N)^p = (e^{-1})^p = e^{-p}$$

So the number of examples not that are not sampled is approximately $N \cdot e^{-p}$. That is, approximately $N - (e^{-p} \cdot N)$ examples have been sampled at least once.

Problem 14

We can categorize $g_{s,i,\theta}(\mathbf{x})$ into two categories:

- Independent of $\mathbf{x}$: If θ falls within $(-\infty, L]$ or (R, ∞), then regardless of i and $\mathbf{x}$, the output of $g_{s,i,\theta}$ depends only on s. It can be either positive or negative for all $\mathbf{x} \in \mathcal{X}$. In this case, there are 2 different decision stumps, all positive and all negative.
- dependent of $\mathbf{x}$: In contrast to the first category, if θ falls within $(L, R]$, then both i and $\mathbf{x}$ will affect the output. In this case, there are $d = 4$ choices of i, and for each i, there are 5 intervals to choose from, i.e. $(0,1], (1,2], (2,3], (3,4], (4,5]$. For each θ, s can be either -1 or 1. This gives us $4 \cdot 5 \cdot 2 = 40$ different decision stumps.

In total, we have $2 + 40 = 42$ different decision stumps.

Problem 15

Consider $g_{s,i,\theta}(\mathbf{x}) \cdot g_{s,i,\theta}(\mathbf{x}')$ for some $i \in \{1, 2, \dots, d\}$, $s \in \{+1, -1\}$, $\theta \in \mathbb{R}$, we have

$$\begin{aligned} g_{s,i,\theta}(\mathbf{x}) \cdot g_{s,i,\theta}(\mathbf{x}') =& s \cdot \operatorname{sign}(x_i - \theta) \cdot s \cdot \operatorname{sign}(x_i' - \theta) \\ =& \operatorname{sign}(x_i - \theta) \cdot \operatorname{sign}(x_i' - \theta) \end{aligned}$$

we can see that

$$g_{s,i,\theta}(\mathbf{x}) \cdot g_{s,i,\theta}(\mathbf{x}') = \begin{cases} -1 & \text{if } \min(x_i, x_i') < \theta \le \max(x_i, x_i') \\ 1 & \text{else} \end{cases}$$

Since both $\mathbf{x}$ and $\mathbf{x}'$ are integer vectors, for each $i \in \{1, 2, \dots, d\}$, there are $2 \cdot |x_i - x_i'|$ (coefficient 2 is for two choices of s) different decision stumps that yield -1.
Now we consider $|\mathcal{G}|$. Following the logic in the previous problem, we know for a given tuple (d, L, R), there are $2 + 2d \cdot (R - L)$ different desicion stumps. That is, $|\mathcal{G}| = 2 + 2d \cdot (R - L)$, and the number of $+1$ in $K_{ds}(\mathbf{x}, \mathbf{x}')$ is

$$2 + 2d \cdot (R - L) - \sum_{i=1}^{d} 2 \cdot |x_i - x_i'|$$

Then we have

$$\begin{aligned} K_{ds}(\mathbf{x}, \mathbf{x}') &= 2 + 2d \cdot (R - L) - \sum_{i=1}^{d} 2 \cdot |x_i - x_i'| - \sum_{i=1}^{d} 2 \cdot |x_i - x_i'| \\ &= 2(1 + d \cdot (R - L) - \sum_{i=1}^{d} 2 \cdot |x_i - x_i'|) \end{aligned}$$

Problem 16

Following the similar idea in the previous problem, we consider the *i-th* dimension for some $i \in \{1, 2, \dots, d\}$. But since $\mathbf{x}$ and $\mathbf{x}'$ are real vectors, we should use integration in this case. The sum of all decision stumps regarding the *i-th* dimension is given by

$$\begin{aligned} & 2 \cdot \int_L^R s \cdot \operatorname{sign}(x_i - \theta) \cdot s \cdot \operatorname{sign}(x_i' - \theta) d\theta \\ =& 2 \cdot \int_L^R \operatorname{sign}(x_i - \theta) \cdot \operatorname{sign}(x_i' - \theta) d\theta \\ =& 2 \cdot (\int_L^R 1 d\theta - 2 \cdot \int_{\min(x_i, x_i')}^{\max(x_i, x_i')} 1 d\theta) \\ =& 2 \cdot (R - L - 2 \cdot |x_i - x_i'|) \end{aligned}$$

The coefficient 2 reflects the two choices of s. Now consider every dimension of $\mathbf{x}$ and $\mathbf{x}'$, we have

$$
\begin{aligned}
K_{ds}(\mathbf{x}, \mathbf{x}') =& \sum_{i=1}^{d} 2 \cdot (R - L - 2 \cdot |x_i - x_i'|) \\
=& 2 \cdot (d \cdot (R - L) - 2 \sum_{i=1}^{d} |x_i - x_i'|)
\end{aligned}
$$

Problem 17

My favorite lecture is **Blending / Bagging / Adaptive Boosting**. First of all, the lecturer explained the techniques very clearly and thorougly. I enjoyed the lecture very much. Secondly, I was amazed by how these techniques could combine several weak models to form a strong model. It seemed like very useful techniques that I could use often when training machine learning models.

Problem 18

The lecture I like the least is the first lecture of **Convolutional Neural Network**. Although the topic itself sounded interesting, the learning experience was not so good due to technical issues and unfamiliar teaching style.