Introducing to Machine Learning (CS 419M)
Computer Science and Engineering
Indian Institute of Technology Bombay

Midterm Exam
2022-02-24 Thursday

Instructions

1. It is a CLOSED BOOK examination.
2. This paper has three questions. Each question carries 10 marks. Therefore, the maximum marks is 30.
3. Write your answers on a paper, scan and submit them at the end of the exam.
4. Write your name, roll number and the subject number (CS 419M) on the top of each of your answer script.
5. There are multiple parts (sub-questions) in each question. Some sub-questions are objective and some are subjective.
6. There will be partial credits for subjective questions, if you have made substantial progress towards the answer. However there will be NO credit for rough work.
7. Please keep your answer sheets different from the rough work you have made. Do not attach the rough work with the answer sheet. You should ONLY upload the answer sheets.

1. Consider the one dimensional linear regression problem where we neglect the bias term. Suppose that the dataset $\{(x_i, y_i)\}_{i=1}^m$is generated by sampling m samples from the following distribution D

$$(x,y) = \begin{cases} (1,0) & \text{with probability } \mu \\ (\mu,-1) & \text{with probability } (1-\mu) \end{cases} \tag{1}$$

1.a Find the probability that the dataset contains atleast one sample $(\mu, -1)$.

1.a /2

1.b Find $\mathbb{E}[x]$, $\mathbb{E}[y]$, $\mathrm{Var}(x)$, $\mathrm{Var}(y)$ where $\mathbb{E}[\cdot]$ denotes expectation and $\mathrm{Var}(\cdot)$ denotes the variance.

1.b /4

1.c Consider the loss function

$$l(w) = \sum_{i=1}^m (y_i - wx_i)^2 \tag{2}$$

Find the expected value of this loss function with respect to the distribution D and denote it $L_D(w)$. Then find w^* that minimizes

$$L(w) = L_D(w) + \lambda|w|^2 \tag{3}$$

1.c /4

1.a This will be $1 - \mu^m$

1.b

$$\begin{aligned}
\mathbb{E}[x] &= \mu\,1 + (1-\mu)\,\mu = 2\mu - \mu^2 && (4)\\
\mathbb{E}[y] &= \mu\,0 + (1-\mu)\,(-1) = \mu - 1 && (5)\\
\mathrm{Var}(x) &= \mathbb{E}[x^2] - \big(\mathbb{E}[x]\big)^2 && (6)\\
&= \mu\,1 + (1-\mu)\,\mu^2 - \big(2\mu-\mu^2\big)^2 && (7)\\
&= \mu - 3\mu^2 + 3\mu^3 - \mu^4 && (8)\\
\mathrm{Var}(y) &= \mathbb{E}[y^2] - \big(\mathbb{E}[y]\big)^2 && (9)\\
&= \mu\,0 + (1-\mu)\,1 - (\mu-1)^2 && (10)\\
&= \mu - \mu^2 && (11)
\end{aligned}$$

1.c

$$\begin{aligned}
L_D(w) &= \mathbb{E}\left[\sum_{i=1}^m (y_i - wx_i)^2\right] && (12)\\
&= \sum_{i=1}^m \mathbb{E}[y_i^2] + w^2\mathbb{E}[x_i^2] - 2\,w\,\mathbb{E}[y_i\,x_i] && (13)\\
&= m\Big((1-\mu) + w^2(\mu+\mu^2-\mu^3) - 2w(\mu^2-\mu)\Big) && (14)
\end{aligned}$$

Hence

$$\begin{aligned}
L(w) &= L_D(w) + \lambda|w|^2 && (15)\\
&= m\Big((1-\mu) + w^2(\mu+\mu^2-\mu^3) - 2w(\mu^2-\mu)\Big) + \lambda w^2 && (16)\\
\implies \frac{dL}{dw} &= m\Big(2w(\mu+\mu^2-\mu^3) - 2(\mu^2-\mu)\Big) + 2\lambda\,w = 0 && (17)
\end{aligned}$$

Hence

$$w^* = \frac{m(\mu^2-\mu)}{m(\mu+\mu^2-\mu^3)+\lambda} \tag{18}$$

This is valid only if it is a minima, i.e., the second derivative should be non-negative, i.e.,

$$\frac{d^2L}{dw^2} = m(\mu + \mu^2 - \mu^3) + \lambda > 0 \tag{19}$$

2. Consider a 1-dimensional linear regression problem. The dataset corresponding to this problem has n examples $D = \{(x_i, y_i)\}_{i=1}^n$, where $x_i, y_i \in \mathbb{R} \quad \forall i$. We don't have access to the inputs or outputs directly and we don't know the exact value of n. However, we have a few statistics computed from the data.

Let $\mathbf{w}^* = [w_0^*, w_1^*]^\top$ be the unique solution that minimizes $J(\mathbf{w})$ given by:

$$J(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)^2 \tag{20}$$

2.a Which of the following are true:

i. $\frac{1}{n}\sum\limits_{i=1}^{n}(y_i - w_0^* - w_1^* x_i)y_i = 0$

ii. $\frac{1}{n}\sum\limits_{i=1}^{n}(y_i - w_0^* - w_1^* x_i)(y_i - \bar{y}) = 0$

iii. $\frac{1}{n}\sum\limits_{i=1}^{n}(y_i - w_0^* - w_1^* x_i)(x_i - \bar{x}) = 0$

iv. $\frac{1}{n}\sum\limits_{i=1}^{n}(y_i - w_0^* - w_1^* x_i)(w_0^* + w_1^* x_i) = 0$

where $\bar{x}$ and $\bar{y}$ are the sample means based on the same dataset D. Please provide justification for your answer.

2.a /2

2.b Suppose we have the following statistics computed from the dataset D

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i \qquad C_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$C_{yy} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad C_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Express w_1^* and w_0^* using (some or all of) these statistics. Show the derivation for full credit.

2.b /4

2.c Now suppose that the dataset D has been corrupted with some Gaussian noise, i.e. the new dataset will be $\tilde{D} = \{(\tilde{x}_i, y_i)\}_{i=1}^n$ where $\tilde{x}_i = x_i + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Now derive the expressions for $\tilde{w_0}^*$ and $\tilde{w_1}^*$ obtained by minimizing $J(\mathbf{w})$ on this noisy dataset in terms of σ and the true statistics (*i.e.*, when n is large) given in part (b) above.

2.c /4

2.a Equating the derivative of $J(\mathbf{w})$ with respect to w_0 and w_1 to zero, we get

$$\frac{\partial J}{\partial w_0} = \frac{1}{n}\sum_{i=1}^{n} 2\left(y_i - w_0^* - w_1^* x_i\right) = 0 \tag{21}$$

$$\frac{\partial J}{\partial w_1} = \frac{1}{n}\sum_{i=1}^{n} 2\left(y_i - w_0^* - w_1^* x_i\right) x_i = 0 \tag{22}$$

Based on these two equations, it is easy to see that only *iii* and *iv* are true.

2.b On expanding the equations 21 and 22 we get

$$w_0^* + w_1^* \bar{x} = \bar{y} \tag{23}$$

$$\bar{x}\, w_0^* + (C_{xx} + \bar{x}^2) w_1^* = C_{xy} + \bar{x}\bar{y} \tag{24}$$

On solving, we get

$$w_0^* = \bar{y} - \frac{C_{xy}}{C_{xx}}\bar{x} \tag{25}$$

$$w_1^* = \frac{C_{xy}}{C_{xx}} \tag{26}$$

2.c The new statistics of the data in terms of old statistics are as follows:

$$\bar{\tilde{x}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i = \bar{x} \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \tilde{C}_{xx} = \frac{1}{n}\sum_{i=1}^{n} (\tilde{x}_i - \bar{x})^2 = C_{xx} + \sigma^2$$

$$C_{yy} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \tilde{C}_{xy} = \frac{1}{n}\sum_{i=1}^{n} (\tilde{x}_i - \bar{x})(y_i - \bar{y}) = C_{xy} + \frac{1}{n}\sum_{i=1}^{n} \epsilon_i\, y_i = C_{xy}$$

Hence, the modified equations will be

$$w_0^* = \bar{y} - \frac{C_{xy}}{C_{xx} + \sigma^2}\bar{x} \tag{27}$$

$$w_1^* = \frac{C_{xy}}{C_{xx} + \sigma^2} \tag{28}$$

3. Consider the problem of learning a binary classifier on a dataset S using the 0-1 loss i.e.,

$$l^{0-1}(\mathbf{w}, (\mathbf{x}, y)) = \mathbb{I}_{[y \neq \text{sign}(\mathbf{w}^\top \mathbf{x})]} = \mathbb{I}_{[y\mathbf{w}^\top \mathbf{x} \leq 0]}$$

where $\mathbf{w}, \mathbf{x} \in \mathbb{R}^d$, $y \in \{-1, +1\}$ and $\mathbb{I}_{[\cdot]}$ denotes the indicator function.

3.a Design a convex surrogate loss function for this 0-1 loss. Note than a convex surrogate loss should (1) be convex and (2) should always upper bound the original loss function. You will get full marks only if you show that the convex surrogate you propose satisfies these two properties.

3.a /2

3.b Define

$$F_S(\mathbf{w}) = \lambda |S| \|\mathbf{w}\|^2 + \sum_{i \in s} [1 - y_i \mathbf{w}^\top \mathbf{x}_i]_+ \tag{29}$$

Show that, if $\mathbf{w}^*$ minimizes $F_S(\mathbf{w})$ then

$$\|\mathbf{w}^*\| = \mathcal{O}\left(\frac{1}{\sqrt{\lambda}}\right) \tag{30}$$

where we denote $a = \mathcal{O}(b)$ iff

$$a \leq \frac{k}{b} \tag{31}$$

for some constant k.

3.b /3

3.c Now define

$$F_S(\mathbf{w}) = \lambda |S| \|\mathbf{w}\|^2 + \sum_{i \in s} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \tag{32}$$

Show that, if $\mathbf{w}^*$ minimizes $F_S(\mathbf{w})$ and $||\mathbf{x}||_2 \leq x_{\max}$ then

$$\|\mathbf{w}^*\| = \mathcal{O}\left(\frac{1}{\lambda}\right) \tag{33}$$

3.c /5

3.a When y and $\mathbf{w}^\top \mathbf{x}$ are of opposite signs, only then value of the loss will be 1, otherwise it will be 0. So we want

$$y\mathbf{w}^\top \mathbf{x} \leq 0 \tag{34}$$

$$\iff 1 - y\mathbf{w}^\top\mathbf{x} \geq 0 \tag{35}$$

Based on the inequality 35, we can propose the following hinge loss:

$$l(\mathbf{w}, (\mathbf{x}, y)) = \max\{0, 1 - y\mathbf{w}^\top\mathbf{x}\} \tag{36}$$

It remains to show that the hinge loss is a valid convex surrogate for the 0-1 loss. We show this as follows:

- **Convexity:** We know that $1 - y\mathbf{w}^\top\mathbf{x}$ is convex as it is an affine transformation. Hinge loss is nothing but composition of a convex function $g(w) = \max 0, w$ and an affine transformation. Since composition of two convex functions is convex, the hinge loss is also convex.
- **Hinge loss upper bounds 0-1 loss:** Consider the case when $y\mathbf{w}^\top\mathbf{x} \leq 0$. Then $l^{0-1}(\mathbf{w}, (\mathbf{x}, y)) = 1$ and $l(\mathbf{w}, (\mathbf{x}, y)) = 1 - y\mathbf{w}^\top\mathbf{x}$. As $y\mathbf{w}^\top\mathbf{x} \leq 0$, we have $l(\mathbf{w}, (\mathbf{x}, y)) \geq l^{0-1}(\mathbf{w}, (\mathbf{x}, y))$. Similarly, consider the case when $y\mathbf{w}^\top\mathbf{x} \geq 1$. Then $l^{0-1}(\mathbf{w}, (\mathbf{x}, y)) = 0$ and $l(\mathbf{w}, (\mathbf{x}, y)) = 0$. If $0 \leq y\mathbf{w}^\top\mathbf{x} \leq 1$, then $l^{0-1}(\mathbf{w}, (\mathbf{x}, y)) = 0$ and $l(\mathbf{w}, (\mathbf{x}, y)) = 1 - y\mathbf{w}^\top\mathbf{x}$. Again in this case we have $l(\mathbf{w}, (\mathbf{x}, y)) \geq l^{0-1}(\mathbf{w}, (\mathbf{x}, y))$.

3.b As $\mathbf{w}^*$ minimizes $F_S(\mathbf{w})$, we have

$$F_S(\mathbf{w}^*) \leq F_S(0) \tag{37}$$

$$\lambda|S|\|\mathbf{w}^*\|^2 + \sum_{i\in s}(y_i - \mathbf{w}^{*\top}\mathbf{x}_i)^2 \leq \sum_{i\in s} 1 = |S| \tag{38}$$

Now, the left hand side of inequality 38 has a lower bound given by

$$\lambda|S|\|\mathbf{w}^*\|^2 \leq \lambda|S|\|\mathbf{w}^*\|^2 + \sum_{i\in s}(y_i - \mathbf{w}^{*\top}\mathbf{x}_i)^2 \tag{39}$$

By combining the inequalities 39 and 38, we have

$$\lambda|S|\|\mathbf{w}^*\|^2 \leq |S| \tag{40}$$

$$\implies \|\mathbf{w}^*\|^2 \leq \frac{1}{\lambda} \tag{41}$$

$$\implies \|\mathbf{w}^*\| \leq \frac{1}{\sqrt{\lambda}} \tag{42}$$

$$\implies \|\mathbf{w}^*\| = \mathcal{O}\left(\frac{1}{\sqrt{\lambda}}\right) \tag{43}$$

3.c We have

$$F_S(\mathbf{w}) = \lambda|S|\|\mathbf{w}\|^2 + \sum_{i\in S}(y_i - \mathbf{w}^\top\mathbf{x}_i)^2 \tag{44}$$

$$= \lambda|S|\|\mathbf{w}\|^2 + \sum_{i\in S} y_i^2 - 2y_i\mathbf{w}^\top\mathbf{x}_i + (w^\top\mathbf{x}_i)^2 \tag{45}$$

Ignoring the terms independent of w above, define

$$G_S(\mathbf{w}) = \lambda|S|\|\mathbf{w}\|^2 + \sum_{i\in S}(w^\top\mathbf{x}_i)^2 - 2y_i\mathbf{w}^\top\mathbf{x}_i \tag{46}$$

Now, it is easy to see that w^* also minimizes $G_S(\mathbf{w})$. Hence, we will have

$$G_S(\mathbf{w}^*) \leq G_S(0) \tag{47}$$

$$\implies \lambda|S|\|\mathbf{w}^*\|^2 + \sum_{i\in S}(w^{*\top}\mathbf{x}_i)^2 - 2y_i\mathbf{w}^{*\top}\mathbf{x}_i \leq 0 \tag{48}$$

The left side of the inequality 48 can be lower bounded using the Schwarz inequality, giving

$$\lambda|S|\|\mathbf{w}^*\|^2 - 2\,|S|\|\mathbf{w}^*\|\,x_{max} \leq \lambda|S|\|\mathbf{w}^*\|^2 + \sum_{i\in S}(w^{*\top}\mathbf{x}_i)^2 - 2y_i\mathbf{w}^{*\top}\mathbf{x}_i \tag{49}$$

By combining the inequalities 48 and 49, we have

$$\lambda |S| \|\mathbf{w}^*\|^2 - 2\,|S| \|\mathbf{w}^*\|\, x_{max} \leq 0 \tag{50}$$

$$\implies \|\mathbf{w}^*\| \leq \frac{2\, x_{max}}{\lambda} \tag{51}$$

$$\implies \|\mathbf{w}^*\| = \mathcal{O}\left(\frac{1}{\lambda}\right) \tag{52}$$

Total: 30