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

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

Find the expected value of this loss function with respected 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	

$1 - \mu^m$

$E[x] = \mu(2-\mu)$, $E[y] = \mu - 1$, $E[x^2] = \mu(1 + \mu(1-\mu))$

$E[y^2] = 1 - \mu$ $\mathrm{Var}(x) = \mu(1-\mu)^3$ $\mathrm{Var}(y) = \mu(1-\mu)$

$L_D(w) := E[x^2]\|w\|^2 + (1-\mu) + 2\mu(1-\mu)\sum w_i$

$$w^* = -\frac{\mu(1-\mu)}{\lambda + E[x^2]}\mathbf{1} = -\frac{\mathrm{Var}(y)}{\lambda + E[x^2]}\mathbf{1}$$

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:

$$J(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)^2 \tag{4}$$

$\partial J / \partial w_1$

$0 = \sum_{i=1}^{n}(y_i - w_0^* - w_1^* x_i)x_i$

2.a Which of the following are true:

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

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

$\partial J / \partial w_0$

$0 = \sum_{i=1}^{n}(y_i - w_0^* - w_1^* x_i)$

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

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

Are linear combinations of above 2 equations

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	

2b $w_1^* = \frac{C_{xy}}{C_{xx}}, \quad w_0^* = \bar{y} - \bar{x}\frac{C_{xy}}{C_{xx}}$

2c $C_{\tilde{x}y} = C_{xy} + C_{\varepsilon y} = C_{xy}$

$C_{\tilde{x}\tilde{x}} = \mathrm{Cov}(x+\varepsilon, x+\varepsilon) = C_{xx} + 2C_{x\varepsilon} + C_{\varepsilon\varepsilon}$

$= C_{xx} + \sigma^2$

$\bar{\tilde{x}} = \bar{x}$ ($\because$ n is large, $\mathrm{Var}(\sum \varepsilon_i) = \sigma^2/n \to 0$ $\therefore$ Noise is almost zero in the mean)

$\therefore \tilde{w}_1^* = \frac{C_{xy}}{C_{xx} + \sigma^2}; \quad \tilde{w}_0^* = \bar{y} - \bar{x}\tilde{w}_1^*$

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

$$F_S(\mathbf{w}) = \lambda|S|\|\mathbf{w}\|^2 + \sum_{i \in s}[1 - y_i \mathbf{w}^\top \mathbf{x}_i]_+ \tag{5}$$

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

$$\|\mathbf{w}^*\| = \mathcal{O}\left(\frac{1}{\sqrt{\lambda}}\right) \tag{6}$$

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

$$a \leq \frac{k}{b} \tag{7}$$

for some constant k.

3.b	/3	

3.c Now define

$$F_S(\mathbf{w}) = \lambda|S|\|\mathbf{w}\|^2 + \sum_{i \in s}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \tag{8}$$

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

$$\|\mathbf{w}^*\| = \mathcal{O}\left(\frac{1}{\lambda}\right) \tag{9}$$

3.c	/5	

3a $\dfrac{\ln(1+e^{-yw^Tx})}{\ln 2}$

3b $2\lambda|S|w_* - \sum_{i\in S} \mathbb{1}_{w_*^T v_i < 1} v_i = 0$; $v_i = y_i x_i$

$\Rightarrow 2\lambda|S|\,\|w_*\|^2 = \sum_{i\in S} \mathbb{1}_{w^T v_i < 1}\, w^T v_i \leq |S|$

$\therefore 0 \leq 2\lambda\,|S|\,\|w_*\|^2 \leq |S| \Rightarrow \|w_*\| \leq \frac{1}{\sqrt{2\lambda}}$

3c $w^* = \dfrac{y\bar{x}}{\lambda + \|x\|^2}$ $\therefore \|w^*\| = O\left(\frac{1}{\lambda}\right)$

Total: 30