Problem 0 [1 bonus mark]. Write your name and student ID (number) on every submitted answer sheet.

Problem 1 [10 marks]. In a typical case, more training samples will result in training mean squared error (MSE) loss going __up__ and validation error going _down_. Fill in the blanks with "up" and/or "down".

Problem 2 [30 marks]. discrete

The variance of a random variable X is defined as

$$\mathrm{Var}_{X\sim P(X)}[X] = \mathbb{E}_{X\sim P(X)}\left[\left(X - \mathbb{E}_{X\sim P(X)}[X]\right)^2\right].$$

a) Prove $\mathrm{Var}_{X\sim P(X)}[X] = \mathbb{E}_{X\sim P(X)}[X^2] - \left(\mathbb{E}_{X\sim P(X)}[X]\right)^2$.

b) Use the definition of expectation to prove its linearity:

$$\mathbb{E}_{X\sim P(X)}[af(X) + bg(X)] = a\mathbb{E}_{X\sim P(X)}[f(X)] + b\mathbb{E}_{X\sim P(X)}[g(X)]$$

c) Does linearity hold for variance? Prove or disprove the following statement:

$$\mathrm{Var}_{X\sim P(X)}[af(X) + bg(X)] = a\,\mathrm{Var}_{X\sim P(X)}[f(X)] + b\,\mathrm{Var}_{X\sim P(X)}[g(X)]$$

a) $\underset{X\sim p(X)}{\mathrm{Var}}[X] = \underset{X\sim p(X)}{\mathbb{E}}\left[X^2 + \left(\underset{X\sim p(X)}{\mathbb{E}}[X]\right)^2 - 2X\underset{X\sim p(X)}{\mathbb{E}}[X]\right]$

$= \underset{X\sim p(X)}{\mathbb{E}}[X^2] + \left(\underset{X\sim p(X)}{\mathbb{E}}[X]\right)^2 - 2\left(\underset{X\sim p(X)}{\mathbb{E}}[X]\right)^2$

$= \underset{X\sim p(X)}{\mathbb{E}}[X^2] - 2\left(\underset{X\sim p(X)}{\mathbb{E}}[X]\right)^2$

Dropping $X\sim p(X)$ is okay.

b) $\underset{X\sim p(X)}{\mathbb{E}}[a f(X) + b g(X)] = \sum_X \left[p(X)\cdot(a f(X) + b g(X))\right]$

$= a\sum_X p(X) f(X) + b\sum_X p(X) g(X)$

$= a\underset{X\sim p(X)}{\mathbb{E}}[f(X)] + b\underset{X\sim p(X)}{\mathbb{E}}[g(X)]$

c). No. Counter-example:

Let $X \sim \mathcal{N}(0,1)$

$f(X) = X$ $\quad a = b = 1$

$g(X) = -X$

$\underset{X\sim p(X)}{\mathrm{Var}}[X - X] = \underset{X\sim p(X)}{\mathrm{Var}}[0] = 0$

But $a\underset{X\sim p(X)}{\mathrm{Var}}[f(X)] + b\underset{X\sim p(X)}{\mathrm{Var}}[g(X)]$

$= a\cdot\sigma^2 + b\cdot\sigma^2 = 2$

Problem 3 [60 marks]. Consider (potentially problematic) maximum likelihood estimation (MLE) for feature selection. We assume $t = \sum_{i=0}^{d} \alpha_i w_i x_i + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and x_i is an input feature. Both $\alpha_i \in \{0, 1\}$ and $w_i \in \mathbb{R}$ are model parameters. We denote column vectors by bold letters, for example, $\boldsymbol{\alpha} = (\alpha_0, \cdots, \alpha_d)^\top$.

a) $\alpha_i = 1$ means the ith feature is selected, and $\alpha_i = 0$ means the ith feature is discarded. Fill in the blanks with "selected" and/or "discarded".

b) Consider a training set $\mathcal{D} = \{(\boldsymbol{x}^{(m)}, t^{(m)})\}_{m=1}^{M}$. Present the MLE principle and show its equivalence to mean squared error, denoted by $J(\boldsymbol{\alpha}, \boldsymbol{w})$. Provide derivation steps.

b) Maximum Likelihood estimation:

"," not ideal but earns marks ✓

okay { argmax ⋯ = argmax ⋯ ; max ⋯ = max ⋯

$$\underset{\alpha, w}{\text{maximize}} \sum_{m=1}^{M} \log p(t^{(m)} | x^{(m)}; w)$$

$$\Longleftrightarrow \underset{\alpha, w}{\text{maximize}} \sum_{m=1}^{M} \log c_1 \exp\left\{-c_2\left(t^{(m)} - \sum_{i=0}^{d} \alpha_i w_i x_i\right)^2\right\}$$

$$\Longleftrightarrow \underset{\alpha, w}{\text{minimize}} \frac{1}{2M} \sum_{m=1}^{M} \left(t^{(m)} - \sum_{i=0}^{d} \alpha_i w_i x_i\right)^2$$

↑ constant may vary

↖ $\sum_i$ ok, $\sum_{i=1}^{d}$ not okay

c) Is the loss function convex or not? Briefly explain why or why not. No proof is needed.

No. dom J is not convex as α_i is discrete

d) Closed-form solutions may still be possible for certain non-convex functions. Prove that $\min_{\alpha, w: \alpha_i = 1} J(\alpha, \boldsymbol{w}) \le \min_{\alpha, w: \alpha_i = 0} J(\boldsymbol{\alpha}, \boldsymbol{w})$, where ~~1 is an all-one vector.~~ $J(\boldsymbol{\alpha}, \boldsymbol{w})$ is defined above.

d) For any optimum α^*, w^* ($\alpha_i = 0$) that minimizes $J(\alpha, w)$ we may have optimium $\tilde{\alpha}^*, \tilde{w}^*$ such that

$$\tilde{\alpha}_i^* = 1, \quad \tilde{w}_i^* = 0, \quad \text{and} \quad \tilde{\alpha}_j^* = \alpha_j^*, \quad \tilde{w}_j^* = w_j^* \text{ for } j \ne i$$

that achieves the same value of $J(\alpha, w)$

Therefore $\min_{\alpha, w: \alpha_i = 1} J(\alpha, w) \le \min_{\alpha, w: \alpha_i = 0} J(\alpha, w)$

e) Give a closed-form solution to the problem $\text{minimize}_{\boldsymbol{\alpha}, \boldsymbol{w}} J(\boldsymbol{\alpha}, \boldsymbol{w})$. Provide derivation steps, including the derivation for the optimal $\boldsymbol{w}$.

e) a closed-form solution may have $\alpha^* = 1$

Then the problem redues to classic linear regression

$$J(w) = \frac{1}{2M} \| Xw - t \|^2$$

$$= \frac{1}{2M} (w^T X^T X w - 2 t^T X w + t^T t)$$

$$\frac{\partial J}{\partial w} = \frac{1}{2M} (2 X^T X w - 2 X^T t) \overset{\text{set}}{=} 0$$

$$w^* = (X^T X)^{-1} X^T t$$

f) Explain why such MLE cannot help feature selection.

because MLE will select all features

Hints:

- The optimization variables include α_i and w_i for $i = 0, \cdots, d$.
- The Gaussian assumption suggests $p(\epsilon) = c_1 \exp\{-c_2 \epsilon^2\}$ for some positive constants c_1 and c_2.
- For explanation questions, a few words would suffice. Then can be fit between question lines.

END OF THE EXAM