CSE 446 Midterm Review

Winter 2026 practice exam

Maximum Likelihood Estimation - Q2 (Sidharth)

2. 1 points | Select All That Apply

Let $X_1, \ldots, X_n$ be i.i.d. samples from a continuous distribution with density $p(x \mid \theta)$, where θ is an unknown parameter. Define the likelihood function $\ell(\theta) = \prod_{i=1}^{n} p(X_i \mid \theta)$ and the log-likelihood function $\log \ell(\theta) = \sum_{i=1}^{n} \log p(X_i \mid \theta)$.

Which statements about maximum likelihood estimation (MLE) are true?

(a) Any θ that maximizes $\ell(\theta)$ also maximizes $\log \ell(\theta)$.

(b) The value of the maximized $\ell(\theta)$ is equivalent to the value of the maximized $\log \ell(\theta)$.

(c) $\hat{\theta}_{MLE}$ is always unbiased given any sample size n.

(d) For a Bernoulli model $X_i \sim$ Bernoulli(p) (X_i is 1 with probability p and 0 with probability $1 - p$), the MLE is $\hat{p}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

(e) If the log-likelihood function is concave, then any θ that satisfies $\nabla_\theta \, log \, \ell(\theta) = 0$ is guaranteed to be a global minimum.

Q2 Solution

Solution: Correct choices: A, D.

(A) True: $\log(\cdot)$ is strictly increasing, so maximizing the product likelihood is equivalent to maximizing the sum log-likelihood. (B) False: In general $\ell(\theta) \neq \log \ell(\theta)$ (C) False: MLEs can be biased in finite samples (e.g., variance in a Normal model with unknown mean). (D) True: maximizing $p^{\sum X_i}(1-p)^{n-\sum X_i}$ gives $\hat{p} = \bar{X}$. (E) False: concavity implies critical points are global maxima (or part of a flat maximum set), not minima.

Linear Regression Q7 (Chuning)

7. [2 points]

Consider a linear regression problem with an intercept term. You are given the following small dataset with labels y and data matrix X:

$$y = \begin{pmatrix} 0 \\ 0 \\ 2 \\ 2 \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$$

(note that the data matrix includes a prepended column of ones).

(a) Compute the least-squares solution $\widehat{w} = \begin{pmatrix} \widehat{b} \\ \widehat{m} \end{pmatrix}$ where $\widehat{b}$ is the intercept and $\widehat{m}$ is the slope.

(b) Compute training mean squared error (MSE) over the training labels y and predicted labels $\hat{y} = X\widehat{w}$.

(c) Explain why the training MSE is not zero. (Hint: Look at how a few specific rows in X map to y).

Q7 part A - Solution

Solution:

Part A:

First compute

$$X^\top X = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \quad \Rightarrow \quad (X^\top X)^{-1} = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Next compute

$$X^\top y = \begin{pmatrix} 4 \\ 2 \end{pmatrix}.$$

Therefore,

$$\widehat{w} = (X^\top X)^{-1} X^\top y = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

Q7 - Parts B & C Solution

so $b = 1$ and $m = 1$.

Part B:

$$\hat{y} = Xw = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 2 \end{pmatrix}$$

$$r = y - \hat{y} = \begin{pmatrix} 0 \\ -1 \\ 1 \\ 0 \end{pmatrix}.$$

The sum of squared residuals is $0^2 + (-1)^2 + 1^2 + 0^2 = 2$, so the training MSE is $2/4 = 1/2$.

Part C:

Several potential correct answers to why MSE is not zero.

- Simple intuitive answer - if you observe the 2nd and 3rd row in X, the same vector maps to different y. It is impossible to fit such data points exactly and hence, MSE ¿ 0.
- There are two data samples with the same x and different y. Thus the irreducible error is non-zero, and the MSE (which includes the irreducible error) must also be non-zero.

Basis Functions - Q9 (Chuning)

9. 1 points | Select All That Apply

You are given a dataset of input-output pairs $(x_1, y_1), \ldots, (x_n, y_n)$ where each $x_i, y_i \in \mathbb{R}$. You wish to model the relationship between x and y using Linear Regression (Ordinary Least Squares). You are allowed to transform the inputs x using any fixed basis function

$$h(x) = \left[h_1(x), \ldots, h_k(x)\right]^\top \in \mathbb{R}^k$$

before learning the weights. Which of the following regression models **can** be learned using the standard closed-form solution $\mathbf{w} = (X^T X)^{-1} X^T \mathbf{y}$, where X is defined as

$$X = \begin{bmatrix} h(x_1)^\top \\ \vdots \\ h(x_n)^\top \end{bmatrix} \in \mathbb{R}^{n \times k},$$

for some choice of h?

(a) $y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$

(b) $y = w_0 + e^{w_1 x}$

(c) $y = w_0 + w_1 \sin(x) + w_2 \cos(x)$

(d) $y = w_0 + w_1 e^x + w_2 \log(x)$

Q9 - Solution

Solution: All TRUE except for Option b. The weight w_1 is in the exponent—the model is not linear with respect to the parameters. The rest of the models are linear in parameters, making them solvable using the closed-form solution.

Cross-Validation: Q16 (Claudia)

16. [2 points]

You are training a model to predict whether a user will click an ad. Each data point corresponds to a single ad impression, and each user appears in many impressions. You have $n = 1{,}000{,}000$ total impressions but only 50,000 distinct users.

Your goal is to generalize to new users that do not appear in the training data.

(a) Explain why randomly splitting individual impressions into train/validation/test sets can lead to overly optimistic performance.

(b) Design a correct evaluation procedure. Your answer should specify: (i) how to split the data, (ii) how to perform cross-validation for hyperparameter tuning, and (iii) how to obtain the final test performance.

Q16 - Solution (Claudia)

Solution: **(a)** If we split individual impressions at random, the same user will likely appear in both training and validation/test sets. The model can then exploit user-specific patterns (e.g., embeddings or historical click behavior) learned from training impressions to perform well on validation/test impressions from the same users. This creates data leakage and leads to overly optimistic performance estimates, especially when the goal is to generalize to new users.

(b) Use a group-based split where the unit of splitting is the user:

- Split the set of users into disjoint training, validation, and test user groups (e.g., 80/10/10). All impressions from a given user must belong to exactly one split.
- For hyperparameter tuning, perform k-fold cross-validation over users within the training set: partition training users into k folds, train on $k-1$ folds of users and validate on the held-out fold of users. (All impressions of a user remain together.)
- After selecting hyperparameters, retrain the model on all training users (optionally training + validation users), and evaluate once on the held-out test users to report final performance.

Bias-Variance: Q13 Leo

13. [1 points] You are modeling delivery time in minutes for an online grocery service using squared loss. Features include distance, time of day, driver experience, and number of items. You observe the following behavior as model complexity increases:

- Training error decreases monotonically.
- Test error decreases at first, then increases.
- Even the best-performing model has high test error in regions with heavy traffic.

(a) Explain why the model has high test error in regions with heavy traffic (third bullet), even with a highly flexible model and an extremely large dataset. Your answer should reference the conditional distribution of Y given X.

(b) Suppose you could collect infinite training data, but must keep the same feature set. Which components of the prediction error would go to zero, and which would remain non-zero? Explain using the bias–variance decomposition.

Q13 - Solution

Solution: **(a)** The test error has a lower bound given by the irreducible error

$$\mathbb{E}\big[(Y - \mathbb{E}[Y \mid X])^2 \mid X\big],$$

which arises from randomness in the conditional distribution of Y given X (e.g., unpredictable traffic incidents). This error cannot be reduced by collecting more data or using a more complex model as long as the feature set is fixed.

(b) With infinite data, the variance goes to zero because the learned predictor converges to its expected value. Bias remains if the model class cannot represent the true conditional mean, and irreducible error remains non-zero.

Regularization: Q11 (Sidharth)

11. [1 points]

You are fitting a linear model with many similar (correlated and similar magnitude) features and relatively few samples. Hyperparameters are chosen via cross-validation. You train both ridge and lasso regression with a large regularization parameter λ.

(a) How will ridge and lasso distribute weight across these features differently?

(b) Which solution $\hat{w}_{\text{ridge}}$ or $\hat{w}_{\text{LASSO}}$ is more stable (i.e. lower variance)? In other words, if we were to sample a fresh dataset of the same size, which solution would change less? Explain.

Q11 - Solution

Solution: **(a)** Ridge shrinks all coefficients and distributes weight across correlated features, while lasso tends to select a single feature and set the other correlates to zero.

(b) The ridge regression solution will be more stable. With highly similar features, LASSO will tend to select either one or the other similar feature, so weights will vary significantly across new datasets.

Gradient Descent - Q26 (Claudia)

26. 1 points

Consider minimizing a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ using gradient descent. Starting from an initial point w_0, gradient descent updates are given by $w_{t+1} = w_t - \eta \nabla f(w_t)$, where $\eta > 0$ is the step size (learning rate).

(a) Under what conditions is gradient descent guaranteed to converge to a global minima?

(b) Give an example setting in which increasing the learning rate leads to gradient descent converging in fewer iterations?

(c) Give an example setting in which decreasing the learning rate leads to gradient descent converging in fewer iterations?

Q26 - Solution (Claudia)

Solution: (a) The learning rate must be sufficiently small (or follow a decaying schedule) AND the function f must be *strongly* convex. E.g., minimizing an affine function with nonzero slope will never converge. (b) When the function is very smooth. (c) When the function is not very smooth. Here, the update can overshoot the minimum, causing the iterates to oscillate or diverge instead of converging.

Convexity - Q25 Leo

25. 1 points

Is $f(x) = e^{-x^2}$ a convex function? Explain why or why not.

Q25 - Solution

Solution: $f''(x) = e^{-x^2}(4x^2 - 2)$. Since $e^{-x^2} > 0$ for all x, the sign of $f''(x)$ is determined by $4x^2 - 2$, and $4x^2 - 2 \geq 0 \iff |x| \geq \frac{1}{\sqrt{2}}$. Thus $f''(x) < 0$ for $|x| < \frac{1}{\sqrt{2}}$ and $f''(x) > 0$ for $|x| > \frac{1}{\sqrt{2}}$. Hence e^{-x^2} is not convex on $\mathbb{R}$; it is concave near 0 and convex for large $|x|$ (bell shape curve).

Logistic Regression: Q20 (Sidharth)

20. 1 points | Select All That Apply

Consider logistic regression for binary classification with labels $y \in \{0, 1\}$. The model is defined by

$$P(y = 1 \mid x) = \sigma(x^\top w), \quad \text{where } \sigma(z) = \frac{1}{1 + e^{-z}}.$$

Which of the following statements are true?

(a) The predicted probability $P(y = 1 \mid x)$ is always between 0 and 1.

(b) Increasing $x^\top w$ increases $P(y = 1 \mid x)$.

(c) Classifying an example as $y = 1$ when $P(y = 1 \mid x) \geq 0.5$ is equivalent to checking whether $x^\top w \geq 0$.

(d) The predicted probability $P(y = 1 \mid x)$ is a linear function of x.

Q20 - Solution

Solution: **(a)** TRUE. The sigmoid function maps any real value to the interval $(0, 1)$

(b) TRUE. The sigmoid function is monotonically increasing, so larger values of $x^\top w$ lead to higher predicted probabilities.

(c) TRUE. Because $\sigma(0) = 0.5$, thresholding the probability at 0.5 is equivalent to thresholding the linear score $x^\top w$ at zero.

(d) FALSE. Although $x^\top w$ is linear in x, applying the sigmoid makes the predicted probability a nonlinear function of x.

Classification Pitfalls - Q18 Leo

18. [1 points] [Select All That Apply]

You're given two datasets: $\mathcal{D}$ and $\mathcal{D}_{\text{test}}$, where $\mathcal{D}$ is 100 times larger than $\mathcal{D}_{\text{test}}$. You perform k-fold cross-validation with $\mathcal{D}$ and observe that the mean cross-validation accuracy is 92%, but the final test accuracy on $\mathcal{D}_{\text{test}}$ is 64%. Which of the following could be a reasonable explanation for the discrepancy between these two accuracies?

(a) The model is overfitting to the data distribution $\mathcal{D}$.

(b) The test set is too small.

(c) There is a difference in distribution between $\mathcal{D}$ and $\mathcal{D}_{\text{test}}$.

(d) Cross-validation always overestimates performance.

Q18 - Solution

Solution:

(a) Incorrect, the validation accuracy is 92%, which means we're not overfitting to $\mathcal{D}$.

(b) Correct, too small of a test set could mean the 62% could have wide confidence intervals.

(c) Correct, the model will perform poorly if the test data is completely different from its training data.

(d) Incorrect, cross-validation does not always overestimate performance.

Thanks for coming!
Good luck on the Midterm!!