Midterm 1 Practice Problems

Topics 1 and 2: Linear Regression and Model Selection

Name: _____ **Total: 100 points**

Instructions. Use the supplied formula sheet. **No calculators.** Show substitutions; exact fractions and fully substituted, unevaluated expressions are acceptable. No decimal approximations are required. Justify any requested comparison or model choice. For Q1, select *every* true statement; select E only when A–D are all false. For Q2–Q3, choose one response and give two lines of reasoning. For Q4–Q5, explain the mechanism. For Q6–Q7, show your work. Assume iid samples come from the same population. Include an intercept only where specified; it is not counted among measured features.

Q1. Modeling Basics [15 points]

A campus courier service records 100 iid deliveries. Each example has two numerical features: route distance x_1 and number of stops x_2. The regression target y is delivery time. The data are split into 60 training, 20 validation, and 20 test examples. Test data remain untouched during all fitting and selection.

The team considers linear regression, polynomial regression, ridge, lasso, and kNN. Unless specified otherwise, all preprocessing is fitted on training data. Interpret each subquestion under its own stated conditions. Each subquestion is worth 5 points.

1.1. Feature Representation

☐ A. The augmented design matrix for the original two-feature linear model has shape 60×3.

☐ B. The predictor $w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_1x_2 + w_5x_2^2$ is linear in its fitted coefficients.

☐ C. Because the polynomial model in B has six coefficients and 60 training examples, its least-squares coefficient solution is guaranteed to be unique.

☐ D. If a three-category zone variable is added using all three one-hot indicators alongside an intercept, the resulting columns are linearly dependent.

☐ E. None of these.

1.2. Cross-Validation

As an alternative to the original training/validation split, the team combines those 80 development examples and performs five-fold cross-validation over eight configurations specified in advance. Each fold has the same size, and the original 20 test examples remain reserved.

☐ A. Comparing all configurations requires 40 model fits, before any final refit.

☐ B. It is valid to learn the feature means and standard deviations from all 80 development examples before cross-validation, because scaling does not use the targets.

☐ C. The smallest observed cross-validation error is guaranteed to equal the population error of the model refitted on all 80 examples.

☐ D. After selecting a configuration, the team may refit its scaler and model on all 80 development examples, then apply that fitted scaler to the test examples.

☐ E. None of these.

1.3. Regularization

Return to the original 60 training examples. The augmented model has three coefficients, and $\lambda > 0$. The objectives are

$$E_{\text{ridge}}(w) = \frac{\text{RSS}(w)}{60} + \lambda(w_1^2 + w_2^2), \qquad E_{\text{lasso}}(w) = \frac{\text{RSS}(w)}{60} + \lambda(|w_1| + |w_2|).$$

Here $I' = \text{diag}(0, 1, 1)$.

☐ A. For the stated ridge objective, the normal-equation matrix is $X^T X + \lambda I'$.

☐ B. Adding a penalty on w_0 is required to make the ridge solution unique when the two slope columns are dependent.

☐ C. Multiplying the entire ridge objective, including its penalty, by 3 changes its minimizer.

☐ D. The lasso objective has an ordinary partial derivative with respect to every coefficient at every coefficient vector, including vectors with $w_1 = 0$.

☐ E. None of these.

Must be true, or might not be true?

Use **only the stated assumptions**. Choose one response and give two lines of reasoning. For "Might not be true," a counterexample that disproves the guarantee is sufficient.

Q2. Polynomial Regression [5 points]

On the same training data, fit these two classes by minimizing unregularized MSE exactly:

$$\mathcal{H}_1 = \{w_0 + w_1 x_1\}, \qquad \mathcal{H}_2 = \{w_0 + w_1 x_1 + w_2 x_1^2\}.$$

Statement: The fitted model from $\mathcal{H}_2$ has training MSE no larger than the fitted model from $\mathcal{H}_1$, regardless of which relationship generated the data.

◯ Must be true ◯ Might not be true

Q3. Bias and Variance [5 points]

At a fixed input x_*, repeated independent training samples produce fitted predictors $g^{(D)}$. Suppose

$$\mathbb{E}_D[g^{(D)}(x_*)] = f(x_*).$$

A future label is $Y = f(x_*) + \epsilon$, where $\mathbb{E}[\epsilon] = 0$, $\text{Var}(\epsilon) = \sigma^2$, and future noise is independent of the training sample.

Statement: The expected generalization error at x_*, averaged over training datasets, must equal σ^2.

◯ Must be true ◯ Might not be true

Open-ended conceptual problems

Q4. Learning Rate and Feature Scale [15 points]

Fit $\widehat{y} = w_0 + w_1 x_1$ by batch gradient descent on $J(w) = \mathrm{RSS}(w)/(2N)$, with no regularization. Use relevant course terminology; avoid vague claims.

4.1. Learning Rate

Explain how the choice of a fixed learning rate $\alpha > 0$ affects convergence over the course of training. Discuss the tradeoffs involved in choosing it and how you would tell whether a choice is working well.

4.2. Feature Scaling

Now stretch the feature: replace x_1 by cx_1, where $c > 1$, and replace the current slope w_1 by w_1/c. Leave the intercept unchanged, so the current predictions match. Does a learning rate that gave reliable convergence before necessarily do so now? Explain why or why not, and justify any change to the learning rate or preprocessing.

Q5. Learning Curves [15 points]

Two regression procedures use the same available features. A is linear and B uses degree-six polynomial features. Both minimize their unregularized training MSE successfully. The table gives mean errors across repeated representative training samples; the irreducible MSE is approximately 1.

Training size	A training	A validation	B training	B validation
30	4.8	5.3	0.1	8.0
120	5.0	5.2	0.6	3.1
480	5.1	5.2	0.9	1.4

A colleague prefers A at $N = 480$ because it has the smaller generalization gap. **Explain what the curves suggest, assess that recommendation, and propose the next experiment for each procedure.** Discuss whether more data or changing flexibility/regularization is likely to help, and how you would compare the changes while preserving an honest final evaluation.

Q6. Ridge Regression [25 points]

Consider the training data

$$(x_1, y) = (-1, 0),\ (0, 1),\ (1, 3), \qquad N = 3,$$

and the model $\widehat{y} = w_0 + w_1 x_1$. This problem uses the following objective:

$$\boxed{J_*(w) = \frac{3}{2N} \sum_{i=1}^{N} \left(w_0 + w_1 x_1^{(i)} - y^{(i)} \right)^2 + \frac{1}{2} w_1^2.}$$

The intercept is not penalized. Start at $w^{[0]} = (1, -1)^T$ and use learning rate $\alpha = 1/6$. All coordinates must be updated simultaneously.

6.1. Gradient [7 points]

Find both partial derivatives of J_* as sums over examples. Then express the complete gradient in matrix form using the augmented design matrix X and target vector y.

6.2. Gradient Descent Update [6 points]

Write X and y. Find the predictions, signed errors $Xw^{[0]} - y$, and gradient at $w^{[0]}$. Show the simultaneous update with the old weights, learning rate, and gradient substituted. Report $w^{[1]}$ using exact fractions or equivalent expressions.

Q6. Ridge Regression (continued)

Use the data and objective from the preceding page. Exact expressions are sufficient.

6.3. Loss and R^2 [7 points]

Write fully substituted expressions for $J_*(w^{[0]})$, $J_*(w^{[1]})$, the unpenalized training RSS at $w^{[1]}$, TSS, and training R^2 at $w^{[1]}$. Sums, squares, and fractions may remain unevaluated. Determine whether R^2 is negative, zero, or positive; justify your answer and explain why it is possible despite the intercept.

6.4. Ridge Solution [5 points]

Find $\lambda_{\text{adjusted}}$ and a positive constant c such that

$$J_*(w) = c\left[\frac{1}{N}\,\mathrm{RSS}(w) + \lambda_{\text{adjusted}} w_1^2\right].$$

Use the corresponding normal equations to find the minimizer w_*. The matrix for this dataset is diagonal, so no general matrix inversion is needed.

Q7. kNN and Generalization [20 points]

This question uses a separate one-dimensional training set:

Input x	-2	0	3	5	8
Regression target y	1	2	4	7	8
Class label	A	A	A	B	B

Distance is $|x - x'|$. Regression uses an unweighted average; classification uses majority vote. If two training points are equally distant, order the smaller input first. Use exactly k neighbors. Candidate values are $k \in \{1, 3, 5\}$.

7.1. Nearest Neighbors [5 points]

For $x = 6$, list the training inputs in increasing order of distance and give their distances. Then fill in the regression and classification predictions for each k.

k	Regression prediction	Predicted class
1		
3		
5		

7.2. Validation and k [7 points]

For held-out regression validation pairs $(x, y) = (1, 3)$ and $(6, 6)$, write predictions and validation-MSE expressions. Arithmetic may remain unevaluated. Select k and justify your choice.

k	Prediction at $x = 1$	Prediction at $x = 6$	Validation MSE
1			
3			
5			

Q7. kNN and Generalization (continued)

7.3. Training Error [2 points]

The $k = 1$ regressor has training MSE zero when each example is allowed to be its own neighbor. Does that give a reason to replace the value selected in 7.2 with $k = 1$? Explain briefly.

7.4. Hoeffding Bound [6 points]

The team also uses the selected value of k for its classification rule. All choices are now frozen. On **200 new iid test examples**, this fixed classifier makes **24 mistakes**. Its loss is 0 for a correct prediction and 1 for an incorrect prediction.

Using the supplied Hoeffding inequality:

(a) Identify ν, K, a, and b in the formula sheet's notation. Here K means the number of evaluation examples.

(b) For tolerance $\varepsilon = 1/10$, write exact expressions for the bound on failure probability δ and the guaranteed confidence level $1 - \delta$. Leave exponentials unevaluated.

(c) State the resulting interval for the classifier's true error E_{out}. Exact fractional expressions for its endpoints are sufficient.

(d) Explain why the same single-classifier calculation would not apply directly if the team had first compared $k = 1, 3, 5$ on these 200 test examples and reported the best one's error.