Name:

USC ID:

Read the following rules carefully:

- Write your name and ID number in the solution you submit.
- Please sign and submit the code of honor in the exam with your solutions. The exam cannot be graded without a signed code of honor. You are supposed to do all of the problems on your own without receiving help from others.
- Cheating in the exam will not be tolerated. If you are caught cheating, you will be reported to the authorities. The recommendation of the instructor will be at least an F in the course in such cases.
- Do not post any questions on Piazza about the exam. The TAs have been instructed to refrain from answering questions about the midterm. In case of ambiguity or problems in the questions, just do your best.
- The use of generative AI is prohibited in answering the questions.
- Problems are not sorted in terms of difficulty. Please avoid guess work and long and irrelevant answers.
- Instructions on submitting the solutions to paper and pencil and coding problems will be provided shortly by the TAs. You can handwrite or typeset your solutions to paper and pencil problems.
- Show all your work and your final answer. Showing only the final answer of a question may not receive full credit and you must show your solution and reasoning behind the answer. Simplify your answer as much as you can.
- The exam has 8 questions, 16 pages, and a total of 100 points.
- The submission deadline for this midterm is 11:59 PM, Friday, October 24, 2025.
- As this is a take home exam that extends over several days, OSAS students DO NOT receive extra time, per OSAS guidelines.
- Any change in the midterm (paper and pencil or coding) after the deadline is considered late submission. One second late is late. The midterm is graded based on *when it was submitted, not when it was finished.* The midterm can be submitted up to three days late, with 10% penalty per late day. Homework late days cannot be used for the midterm.
- Submission after the grace period will receive a zero. One second late is late.

Grading Breakdown

Problem	Score	Earned
1	10	
2	10	
3	10	
4	10	
5	10	
6	10	
7	20	
8	20	
Total	100	

Honor Code

I pledge on my honor that I have not given or received any unauthorized assistance on this examination.

Name:

Signature:

1. You are working with a dataset where you train five polynomial regression models (Model A to E) of increasing complexity (degrees 1, 3, 5, 7, and 9 respectively) on the same training data of size $n = 80$. The performance of these models has been evaluated using 5-fold cross-validation. The table below shows the average training and validation Mean Squared Errors (MSE) and Mean Absolute Errors (MAE) for each model:

Model	Degree	Training MSE	Validation MSE
A	1	21.4	24.7
B	3	12.8	14.6
C	5	6.9	8.2
D	7	3.0	9.9
E	9	1.2	23.3

Instructions:

(a) Define and explain the mathematical concepts of bias and variance. How do they relate to model complexity in supervised learning?

(b) Using the above table, calculate the bias^2, variance, and Expected Prediction Error (EPE) for each model. Assume:

- Training MSE $\approx$ Variance
- Validation MSE $\approx$ Total EPE
- Irreducible Error = 1.0

(c) For $x = 2, 4, 6, 8$, use the following table of true function values $f(x)$ and model predictions $\hat{f}(x)$ from Model C. Compute the absolute prediction error and squared error for each x. Then compute the average squared error (MSE):

x	2	4	6	8
$f(x)$ (True)	5.0	9.5	13.0	16.5
$\hat{f}(x)$ (Predicted)	4.8	10.0	12.2	17.0
$\lvert f(x) - \hat{f}(x) \rvert$				
$(f(x) - \hat{f}(x))^2$				

(d) Interpret the trends observed in the table. Which model provides the best tradeoff between bias and variance? Justify your choice using comparisons across at least three models, and classify which models suffer from underfitting or overfitting.

2. Consider a logistic regression problem in which there are no features, which means that:

$$\Pr(Y = 1) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

Assume that we have m data points with label $Y = 1$ and n data points with label $Y = 0$ (remember that features are irrelevant).

(a) Write down the likelihood function $l(\beta_0)$.

(b) Find the Maximum Likelihood estimate $\hat{\beta}_0$ for this data set. [**Hint**: maximize $\log_e l(\beta_0)$].

(c) Determine conditions under which this simple classifier classifies data points into $Y = 1$ or $Y = 0$.

3. For the following data set for classification:

Index	X	Y
1	-1	1
2	0	0
3	3	0
4	1	1
5	-2	0

Assume that we want to construct a regularized logistic regression model for this dataset.

(a) Write down the $\mathscr{L}_1$-regularized loss function $J(\beta_0, \beta_1)$ for this dataset with regularization parameter $\lambda = 2$.

(b) Compare the bias variance of the regularized model with the unregularized model ($\lambda = 0$).

4. Consider the following data set for classification:

Index	X	Y
1	1	1
2	-1	0
3	-2	0

(a) Show all possible bootstrap samples of the dataset that have the same size as this dataset. Note that permutations of the same data set are considered the same dataset, for example $\{1, 2, 3\}$ and $\{2, 3, 1\}$ are the same dataset.

(b) Construct a KNN classification model for all bootstrap samples in part 4a with $K = 2$ and predict the label for the test point $x^* = 0$ by majority votes of the predictions of those bootstrap models. Break ties in favor of class 1.

5. You are given the following dataset containing short text sequences and their associated labels:

Text	**Label**
I loved the movie, it was fantastic!	Positive
The **film** was boring and too long.	Negative
What a great performance by the actors.	Positive
The storyline was weak and predictable.	Negative

Task: Explain and demonstrate how this text data can be processed and classified into the correct sentiment class (Positive or Negative). Your answer should cover the following parts:

(a) **Preprocessing**

i. Tokenize and lowercase the sentence:
"I Loved the movie, it was fantastic!"
Show the processed output.

(b) **Feature Representation**

i. Explain the difference between:
- Bag-of-Words (BoW) representation
- TF–IDF (Term Frequency–Inverse Document Frequency) representation

ii. For the word *"movie"*, calculate its TF–IDF value. **Use the tokenized version of the first sentence** (*"I loved the movie, it was fantastic!"*) **to compute TF, and use the entire dataset above to compute IDF.** Show the formula and calculation steps.

(c) **Model Building** Use the Naïve Bayes classifier and binary TF (TF=1 if the term exists in the document and TF=0 if it does not) to classify the sentence "What a great movie." Use histograms for your density estimates.

6. A researcher studies the relationship between weekly exercise hours (X) and stress level score (Y) in graduate students. The sample size is $n = 26$ and the sample Pearson correlation is $r = -0.39$.

   (a) Test at significance level $\alpha = 0.05$ the null hypothesis $H_0 : \beta_1 = 0$ versus the two-sided alternative $H_1 : \beta_1 \neq 0$. Show the test statistic, decision rule, and conclusion in context.

   (b) Report and interpret the coefficient of determination R^2.

   (c) Briefly explain what the negative sign of r indicates in this scenario.

   Note: Everything needed to solve this question is contained in the exam; do not consult external tables.

7. You are given a binary dataset with two real-valued features (x, y) and a label in $\{0, 1\}$. Class 0 points are denoted by $\circ$ and class 1 points by $\triangle$.

Important rules for *all* parts.

- A point *may not* be its own neighbor.
- Break ties in neighbor votes in favor of class 0.
- Show work to justify your neighbor choices (distances and votes) whenever asked.

Dataset

idx	x	y	label
1	1.0	6.0	0
2	2.2	7.0	0
3	3.1	8.2	0
4	4.0	6.1	0
5	5.2	6.0	1
6	6.0	6.2	0
7	6.2	4.8	1
8	7.0	3.8	1
9	8.2	4.6	1
10	8.8	6.0	0
11	3.2	5.4	1
12	2.8	6.2	0
13	7.6	7.6	1
14	4.8	7.4	1

Distance metrics. For $p = (x_1, y_1)$ and $q = (x_2, y_2)$:

$$d_1(p, q) = |x_1 - x_2| + |y_1 - y_2| \quad \text{(Manhattan)},$$
$$d_2(p, q) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad \text{(Euclidean)},$$
$$d_\infty(p, q) = \max\{|x_1 - x_2|, |y_1 - y_2|\} \quad \text{(Chebyshev)}.$$

Tasks.

(a) **LOOCV of 1-NN across metrics.**
Consider 1-nearest-neighbor classification on (x, y). The distance function will be selected from the options provided. For each option, calculate the LOOCV error of 1-NN and indicate which distance gives superior performance. The options are: Manhattan d_1, Euclidean d_2, and Chebyshev d_∞.

(b) **Effect of k.**
For each metric d_1, d_2, and d_∞, compute the LOOCV error for $k \in \{1, 3, 5\}$. Report the error for each (metric, k) and list the misclassified indices. For the *best LOOCV setting*, show the full 5-NN calculation (neighbor IDs, labels, distances, vote) for *every* misclassified point.

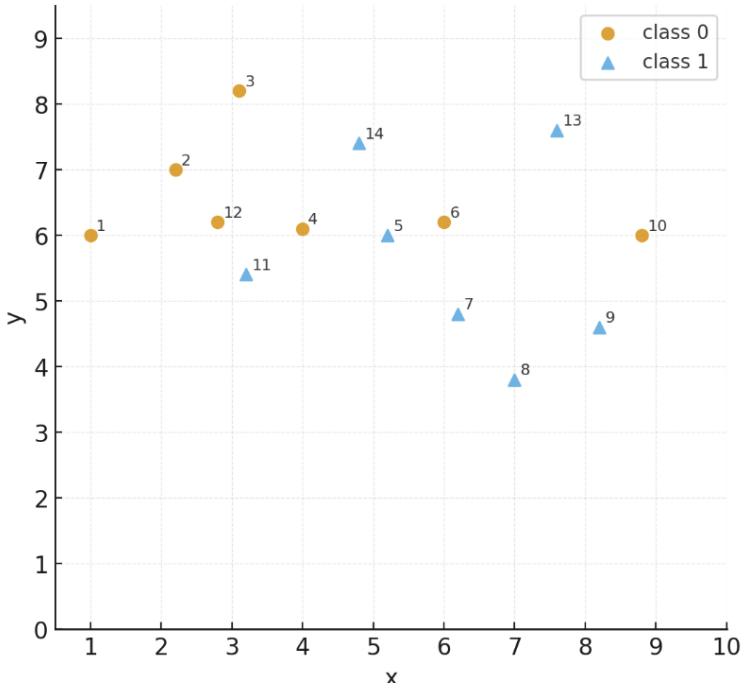

Figure 1: Dataset with indices and class markers. Use this for geometric intuition; grading is based on your calculations.

(c) **Fixed 5-fold CV (Euclidean).**
Using Euclidean d_2 and $k \in \{1, 3, 5\}$, evaluate the fixed folds (do not reshuffle):

$$\text{Fold } 1 = \{1, 7, 11\}, \quad \text{Fold } 2 = \{2, 8, 12\}, \quad \text{Fold } 3 = \{3, 9, 5\}, \quad \text{Fold } 4 = \{4, 6, 10\},$$

$$\text{Fold } 5 = \{13, 14\}.$$

For each k, report the error on each fold, the mean error across folds, and the misclassified index set per fold. Select the k recommended by this split and justify briefly.

(d) **1-NN decision boundary (Euclidean).**
Sketch the qualitative 1-NN (k=1) decision regions for d_2. Indicate at least two places where the boundary is clearly non-linear due to local class interleaving It is fine if you use software to plot it and you do not need to submit the code.

8. **Programming Question: Predicting Housing Prices with Linear Regression**

Dataset: `student-mat.csv`

Link: https://archive.ics.uci.edu/dataset/320/student+performance

In this problem, you will build a linear regression model to predict housing prices in the USA based on various features. You will use Python and `scikit-learn` for this task.

(a) **Data Exploration and Pre-processing**

i. Import the `student-mat.csv` dataset as a pandas DataFrame.

ii. Select the following features: `age, studytime, schoolsup, goout, Dalc, Walc, health, absences, G3` (target/dependent variable). Encode the binary variable (`schoolsup`) values as 0s (no) and 1s (yes). Combine `Dalc` and `Walc` into one variable `alc` by taking averge, then remove `Dalc` and `Walc`. Display the first five rows of the pre-processed dataset.

iii. Find the number of outliers for each independent variable using the IQR method.

iv. Standardize and run PCA on the dataset. Create a scatterplot of PC1 vs PC2, coloring the dots by their final grade `G3`. Inspect the component loadings and determine which features contribute the most to PC1 and PC2. *Keep and use standardized data for remaining problems.*

(b) **Linear Regression**

i. Split data into training set and testing set with an 80:20 ratio. Use random seed 552 for reproducibility.

ii. Build three models using the training set: A. Linear Regression Model, B. Linear Regression Model with Ridge Regularization, and C. Linear Regression Model with Lasso Regularization. Set α=0.1.

iii. Test all three models on the test set. Find out the best performing model with repsect to each of the metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2.

iv. How do you interpret R^2 values from the three models?

v. Print coefficients of independent variables from the three models in one table.

vi. What's the relationship between each independent variable and the dependent variable?

vii. How do the regularization methods differ? What can you conclude about the dataset and features given the results?

viii. If the regularization strength (α) is increased, what would happen to performance metrics?

ix. What are some feature engineering methods you would suggest to improve model performance?

Expected Output: Your submission should include:

- Jupyter Notebook `.ipynb` with all the steps clearly commented.
- The output of each step as specified above (e.g., head of DataFrame, info, describe, missing value counts, evaluation metrics for all models, coefficients, and intercept).
- Visualizations for outlier detection and residual analysis.
- A brief discussion answering the interpretation questions.

Scratch paper

Name:

USC ID:

Scratch paper

Name:

USC ID:

F - Distribution ($\alpha = 0.01$ in the Right Tail)

Denominator Degrees of Freedom df_2 \ Numerator Degrees of Freedom df_1	1	2	3	4	5	6	7	8	9
1	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158
6	13.745	10.925	9.7795	9.1483	8.7459	8.4661	8.2600	8.1017	7.9761
7	12.246	9.5466	8.4513	7.8466	7.4604	7.1914	6.9928	6.8400	6.7188
8	11.259	8.6491	7.5910	7.0061	6.6318	6.3707	6.1776	6.0289	5.9106
9	10.561	8.0215	6.9919	6.4221	6.0569	5.8018	5.6129	5.4671	5.3511
10	10.044	7.5594	6.5523	5.9943	5.6363	5.3858	5.2001	5.0567	4.9424
11	9.6460	7.2057	6.2167	5.6683	5.3160	5.0692	4.8861	4.7445	4.6315
12	9.3302	6.9266	5.9525	5.4120	5.0643	4.8206	4.6395	4.4994	4.3875
13	9.0738	6.7010	5.7394	5.2053	4.8616	4.6204	4.4410	4.3021	4.1911
14	8.8616	6.5149	5.5639	5.0354	4.6950	4.4558	4.2779	4.1399	4.0297
15	8.6831	6.3589	5.4170	4.8932	4.5556	4.3183	4.1415	4.0045	3.8948
16	8.5310	6.2262	5.2922	4.7726	4.4374	4.2016	4.0259	3.8896	3.7804
17	8.3997	6.1121	5.1850	4.6690	4.3359	4.1015	3.9267	3.7910	3.6822
18	8.2854	6.0129	5.0919	4.5790	4.2479	4.0146	3.8406	3.7054	3.5971
19	8.1849	5.9259	5.0103	4.5003	4.1708	3.9386	3.7653	3.6305	3.5225
20	8.0960	5.8489	4.9382	4.4307	4.1027	3.8714	3.6987	3.5644	3.4567
21	8.0166	5.7804	4.8740	4.3688	4.0421	3.8117	3.6396	3.5056	3.3981
22	7.9454	5.7190	4.8166	4.3134	3.9880	3.7583	3.5867	3.4530	3.3458
23	7.8811	5.6637	4.7649	4.2636	3.9392	3.7102	3.5390	3.4057	3.2986
24	7.8229	5.6136	4.7181	4.2184	3.8951	3.6667	3.4959	3.3629	3.2560
25	7.7698	5.5680	4.6755	4.1774	3.8550	3.6272	3.4568	3.3239	3.2172
26	7.7213	5.5263	4.6366	4.1400	3.8183	3.5911	3.4210	3.2884	3.1818
27	7.6767	5.4881	4.6009	4.1056	3.7848	3.5580	3.3882	3.2558	3.1494
28	7.6356	5.4529	4.5681	4.0740	3.7539	3.5276	3.3581	3.2259	3.1195
29	7.5977	5.4204	4.5378	4.0449	3.7254	3.4995	3.3303	3.1982	3.0920
30	7.5625	5.3903	4.5097	4.0179	3.6990	3.4735	3.3045	3.1726	3.0665
40	7.3141	5.1785	4.3126	3.8283	3.5138	3.2910	3.1238	2.9930	2.8876
60	7.0771	4.9774	4.1259	3.6490	3.3389	3.1187	2.9530	2.8233	2.7185
120	6.8509	4.7865	3.9491	3.4795	3.1735	2.9559	2.7918	2.6629	2.5586
∞	6.6349	4.6052	3.7816	3.3192	3.0173	2.8020	2.6393	2.5113	2.4073