CS 550 Homework 2 (25 marks)

Instructions:

a. Due date is Sep 3.
b. Please type your solutions cleanly. We won't grade hand-written answers or poorly types answers. You can use LaTeX, Word or markdown etc. to do it.
c. Only individual attempts and original answers will get you the credits. Copying will lead to 0 marks and penalties will be imposed.

1. **(10 marks)** Deriving Decision Boundary for Learning with Prototypes **LwP Classifier**
   These are some training examples of +ve class: (2,3,4); (1,1,2); (0,2,0)
   These are some training examples of -ve class: (4,3,2); (2,1,1); (3,5,3)

   (1 mark) Compute the average feature vectors of +ve and -ve class examples.
   Let us call them $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$

   (1 marks) Determine the vector $\boldsymbol{w}$ which joins μ_+ and μ_- : $\boldsymbol{w} = \boldsymbol{\mu}_+ - \boldsymbol{\mu}_-$

   (2 marks) Equation of the decision boundary can be written as $\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{0}$
   We know that the decision boundary passes through the mid-point of the prototypes.
   Mid-point is given by $\frac{\mathbf{1}}{\mathbf{2}}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)$
   Compute the value of **b**

   (2 marks) Compute the distances of the training examples from the decision boundary. Are there any errors in classification using the LwP classifier?

   (4 marks) Find the equation of a linear classifier which separates the data without any error? If no such classifier exists, give a justification.

2. **(10 marks)** (1 mark each) **Objective type questions. Answer with a brief justification.**

i. Explain the importance of scaling features for training Large Margin Classifiers?

ii. Explain the effect of changing the value of C from very small to very large in a Soft margin Classifier's objective function?

Minimize $||w||_2^2 + \Sigma_{i=1}^{N} C z_i$

Subject to $y_i(w^T x_i + b) \geq 1 - z_i \; \forall i \in \{1,2,3,\ldots,N\}$

Where z_i are the slacks (errors).

iii. Consider the 2-D array, arr. What is the output of np.sum(arr, axis=1)?

iv. Which of the following are methods for indexing into a DataFrame?

a. Use the loc and iloc functions
b. Directly index columns similar to a Python dictionary
c. Using slices to retrieve a set of rows
d. All of the above

v. What is an indicator feature?

a. A feature that indicates whether or not the data has been pre-processed
b. A feature that represents categorical data, using 1's and 0's to denote which categories are present
c. An indicator of whether or not a category is quantitative or categorical
d. None of the above

vi. What is the main purpose of standardizing data?

a. It lets us view data in terms of standard deviations from the average case (i.e. mean)
b. It lets us use much smaller data values, which makes computation much quicker
c. It makes it easier to calculate the mean of each column in the data
d. It removes outliers from the dataset
e. All of the above

vii. For the given confusion matrix, please calculate accuracy, precision, recall and F1 score.

	Actual	
Predicted	Yes	No
Yes	12	3
No	1	9

viii. What is the difference between bagging and boosting? Which technique would be suitable when the data has high variance?

ix. Justify the cost function $L(\boldsymbol{w})$ used by logistic regression for binary classification and compute its gradient $\frac{\partial L}{\partial w}$ with respect to parameters $\boldsymbol{w}$.

$p = \sigma(\boldsymbol{w}^T \boldsymbol{x}_i) = \frac{1}{1+e^{-w^T x_i}}$ and $L(\boldsymbol{w}) = \Sigma_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$

x. It takes 10 minutes to train a SVM algorithm on a dataset with 100K data points. How much time do you think it might take to train the same algorithm on 1 M data points?

3. A car insurance company is building a risk assessment prediction system based on the age of the driver and the type of car. Can you help them by building a decision tree using information gain as the split point evaluation measure?
   Here is the sample data:

Age of Driver	**Car Type**	**Risk**
25	Sports	L
20	Vintage	H
25	Sports	L
45	SUV	H
20	Sports	H
25	SUV	H

a. (1 mark) Entropy of the dataset is:

Hint: Entropy = $\mathrm{H}(\boldsymbol{S}) = -\sum_{c \in C} p_c \log_2 p_c$

b. (1 mark) Calculate the information gain if we split the dataset by the following rule: Age<=22.5

Hint: $IG = H(S) - \frac{|S_1|}{|S|} H(S_1) - \frac{|S_2|}{|S|} H(S_2)$

c. (1 mark) Calculate the information gain if we split the dataset by the following rule:

Car Type = Sports
Hint: If the Car Type is not Sports, it can be either SUV or Vintage

d. (2 mark) Which of the above two rules should we use for the root-node of the decision tree? Assuming that this is the best choice, can you complete the decision tree?