MML - Review for the Final Exam

The final exam will be during finals week on Monday, May 4 at 11:30 AM. The objective is just going to be to have a few conceptual questions to tie some things together. Here are some examples.

0. Machine learning algorithms are built on *data*, often stored in a data table or data frame. One such table is shown in Table 1.

Table 1: A data table

id	age	income	is_student	signup_date	score	group	target	category
1001	22	31500.50	TRUE	2024-01-15	78.2	control	0.134	A
1002	35	58210.00	FALSE	2023-11-03	88.9	treatment	0.842	B
1003	29	44120.75	TRUE	2024-03-22	69.5	control	0.256	A
1004	41	72000.00	FALSE	2022-07-30	91.3	treatment	0.913	C
1005	26	38990.20	TRUE	2024-05-10	74.8	control	0.478	B

   a. What generic terms might we use for the rows and columns of this type of data?
   b. Which variables are categorical?
   c. Which variables are numeric?
   d. How do you suppose the `id` column should be treated?

1. This problem continues with the data from problem 0.

   a. Suppose we'd like to create a parametric algorithm to predict the target from the other variables. What's the most basic type of algorithm to do so?
   b. Suppose we'd like to create a non-parametric algorithm to predict the target from the other variables. What's the most basic type of algorithm to do so?
   c. Suppose we'd like to create a neural network to predict the target from the other variables. What type of activation function should we use for the output?
   d. Suppose we'd like to create a parametric algorithm to predict the category from the other variables. What's the most basic type of algorithm to do so?
   e. Suppose we'd like to create a non-parametric algorithm to predict the category from the other variables. What's the most basic type of algorithm to do so?
   f. Suppose we'd like to create a neural network to predict the category from the other variables. What type of activation function should we use for the output?

2. Suppose we're given the data N points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N).$$

Let's consider the regression line that's the best least squares approximation to those points.

a. What is the form of your approximation?
(This should be a formula involving two parameters.)
b. Write down the associated least squared error E as a function of your two parameters.
c. Express your previous response using matrix multiplication, a difference, and a 2-norm.
d. The values of the parameters that solve the least squared error problem can be expressed as a projection of a point in N-dimensional space onto a two-dimensional subspace. Use your answer to part (c) to identify the point and the subspace.

3. What is regularization? Be sure to address the following in your response:

- What problem regularization attempts to mitigate,
- What technique is typically used in linear regression and why we might expect it to work, and
- What techniques are typically used for a neural network and why we might expect them to work.

4. What is validation? Be sure to address

- how cross validation is set up for linear or logistic regression and
- why we typically use just a single validation set for neural networks.

5. If I had to name the top three applications of linear algebra to machine learning, I might just say

- i. Norms,
- ii. Orthogonal projection, and
- iii. Diagonalization of a matrix A to SDS^{-1} using eigenvalues and eigenvectors.

Here's one question related to each of those:

- Which norm is used to measure error in a least square approximation and how?
- How is orthogonal projection related to the normal equations and the solution of least squares problems?
- How is diagonalization used in Principal Component Analysis?

6. Suppose we'd like to build a machine learning algorithm for each of the following situations:

    a. A medical diagnosis system to use patient similarity compared to past cases (based on features like age, symptoms, and blood pressure). The objective is to predict whether a patient has a certain condition.
    b. An online advertising platform to estimate the probability that a user will click on an ad, based on features like time of day, device type, and ad category.
    c. A model to predict the number of points that this team might defeat that team by based on prior performances.
    d. An algorithm for Spotify to predict song popularity. The measure of popularity might be number of streams per week and the predictors might be a number of disparate properties like tempo, genre, duration, release season, artist popularity, etc.

    Identify which of the following techniques might work best for each of these:

    - Linear regression,
    - Logistic regression,
    - KNN Classification, or
    - KNN Regression.

    Of course, we'd like to know why.

7. I've used a ratings algorithm to make the following assessments of the strengths of four teams (listed in alphabetical order) over the course of a season:

ID	Team	Strength
1	NCSt	4
2	OSU	10
3	UM	1
4	UNC	6

    The teams are about to play a tournament so, for each pair of teams, I'd like to compute the probability (Prob) that the first team will defeat the second, with the teams listed in the same order as above. I'd also like to make a *binary* prediction (Pred) indicating who the algorithm predicts to be the winner.

    a. Fill out the following table with the values of S1-S2 (the first team's strength minus the second). Also, fill out *reasonable* values for Prob and Pred.
    Note: Don't overthink this problem! You are not required to do any significant computation. This question is simply meant to see if we understand what kinds of numbers represent probabilities, how those probabilities relate to one another and team strength, and how that probability leads to a binary prediction.

Team 1	Team 2	S1-S2	Prob	Pred
NCSt	OSU			
NCSt	UM			
NCSt	UNC			
OSU	UNC			
OSU	UM			
UNC	UM			

b. Does the numeric value Prob represent regression or classification? Why?
c. Does the numeric value Pred represent regression or classification? Why?

8. The images shown in Figure 1 display the results of two different classification algorithms applied to the same data set. The data set consists of 300 different two-dimensional data points each indicated by color. What kinds of algorithms might have generated those plots? There might be more than one reasonable response for each plot but you should be able to articulate some reason for your response.

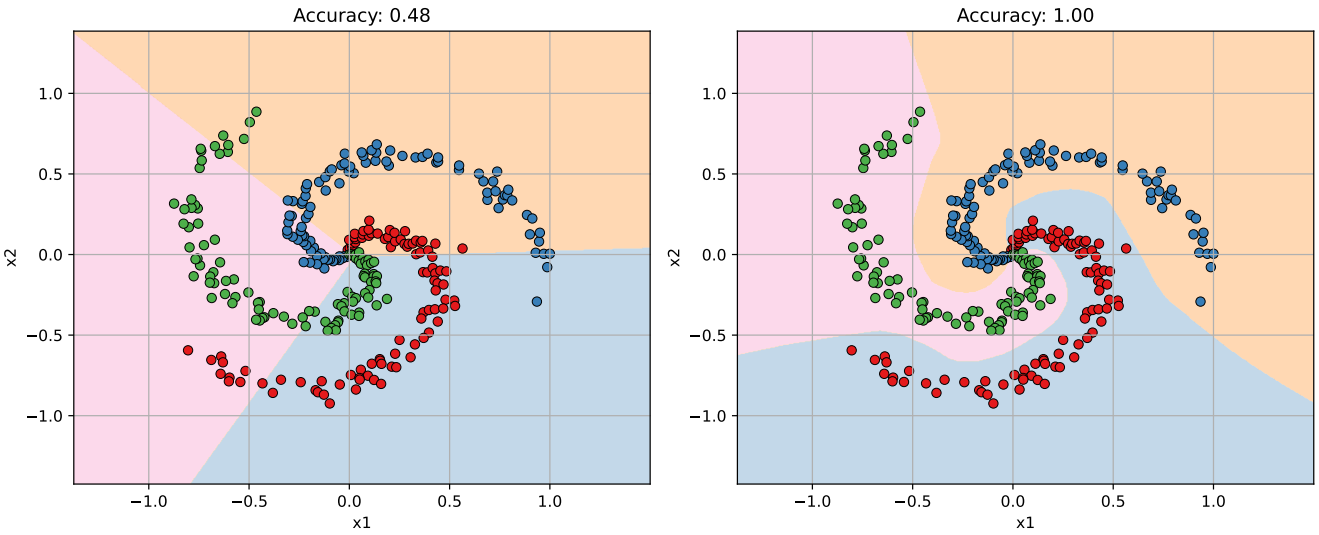

Figure 1: Classification plots