Homework 10: finish by 6/22.

Reading: Notes: Chapters 9 - 10 **Videos:** 9.3 - 10.2

Problem 10.1 (**Video 9.1, 9.2, 9.3, 9.4**, **Lecture Problem**)

A biomedical engineer is testing whether their new device improves patient outcomes after knee surgery. They know that the typical patient population with this condition has a mean symptom severity rating of 7.2. As a preliminary study, they examine 100 patients who have received the device and find a reported average sample severity rating of 6.5 with a sample standard deviation of 1.5. Recall that the following useful values

$$\Phi(-1.64) = Q(1.64) = 0.05,\ \Phi(-1.96) = Q(1.96) = 0.025,\ \Phi(-2.57) = Q(2.57) = 0.005\ .$$

(a) Construct the 90% confidence interval for the average symptom severity rating for patients with the new device.

(b) You would like to construct a significance test to determine whether this new device improves patient outcomes. Explain your null hypothesis and select a test from amongst those covered in Video 9.3.

(c) Calculate the appropriate test statistic.

(d) Based on the observed test data, should the null hypothesis be rejected with significance level 0.05? Explain all steps in your calculation.

Problem 10.2 (**Video 9.1, 9.2, 9.3, 9.4**)

The weights in ounces of coffee in bags produced by Rhett's Roastery can be assumed to be i.i.d. from some single underlying Gaussian distribution. The null hypothesis H_0 is that the mean weight of a bag of coffee is 12 ounces.

You purchase 25 bags and weigh them to obtain values $X_1, X_2, \ldots, X_{25}$, from which you calculate sample mean $M_{25} = 11.92$ and sample variance $V_{25} = 0.04$.

Possibly useful:

$$2\Phi(-1.64) = 0.1, \qquad 2\Phi(-1.96) = 0.05, \qquad 2\Phi(-2.57) = 0.01,$$
$$2F_{T_{24}}(-1.71) = 0.1, \qquad 2F_{T_{24}}(-2.06) = 0.05, \qquad 2F_{T_{24}}(-2.80) = 0.01,$$

(a) What type of statistical test should you use to reject or not reject H_0?

(b) Write an expression for the test statistic to use for the test you selected in part (a).

(c) Do you reject or not reject H_0 at significance level 0.05? Do not merely answer yes or no. Instead, write a self-contained, clear sentence that states your null hypothesis, whether you reject it (or fail to reject it), and according to which significance level.

(d) Find a 90% confidence interval for the mean.

Problem 10.3 (**Video 10.1, 10.2**, **Lecture Problem**)

You are given the following 8 training data points and 8 testing data points:

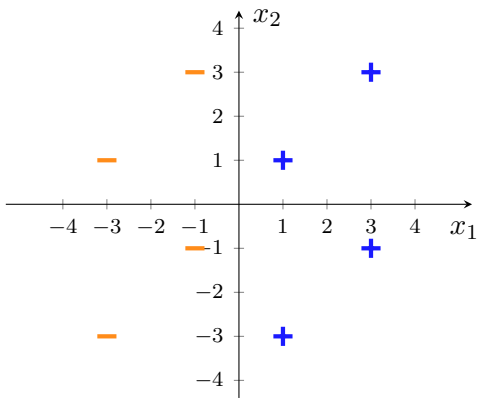
Training Data

x_1	x_2	label
1	1	+
3	3	+
1	-3	+
3	-1	+
-1	-1	-
-3	-3	-
-1	3	-
-3	1	-

Testing Data

x_1	x_2	label
2	3	+
0	-1	+
2	1	+
-0.5	-3	+
-2	1	-
-3	-2	-
0.5	3	-
0.5	1	-

The data is formatted as a table where the first column is the x_1 coordinate (i.e., feature), the second column is the x_2 coordinate (i.e., feature), and the third column is the label, which is either $+1$ or -1. For each part below, you will be asked to create a sketch of the decision region, place the testing points onto the sketch, and make the corresponding decisions. To help get you started, here is a sketch of the training data.



(a) Sketch the decision boundaries for the *nearest neighbor* classifier, add the testing data to your plot, and circle the testing points that will be misclassified. Determine the resulting estimate of the probability of error from the testing data.

(b) Determine the sample mean $\hat{\underline{\mu}}_+$ for the + training data and the sample mean $\hat{\underline{\mu}}_-$ for the − training data. On a single sketch, include the sample means for + and − data, the closest average decision boundary, the testing data, and circle the testing points that will be misclassified. Estimate the probability of error from the testing data.

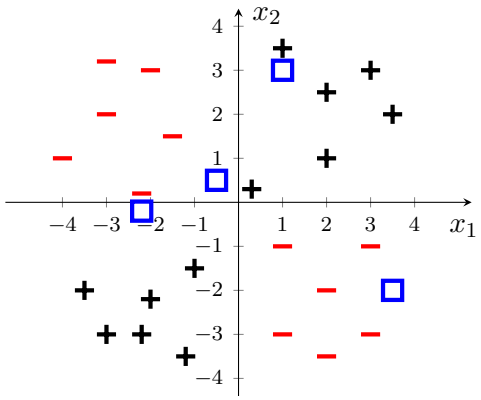
(c) Determine the slope a and offset b for the LDA decision boundary, expressed as a line of the form $x_2 = ax_1 + b$. For this training dataset, the pooled estimate of the covariance matrix is $\hat{\mathbf{\Sigma}} = \frac{4}{3}\begin{bmatrix} 1 & 1 \\ 1 & 5 \end{bmatrix}$ and its inverse is $\hat{\mathbf{\Sigma}}^{-1} = \frac{3}{16}\begin{bmatrix} 5 & -1 \\ -1 & 1 \end{bmatrix}$. On a single sketch, include the sample means for + and − data, the line for the LDA decision boundary, the testing data, and

circle the testing points that will be misclassified. Estimate the probability of error from the testing data.

(d) Argue why, for this particular training dataset, the QDA decision boundary will be exactly the same as the LDA decision boundary.

Problem 10.4 (**Video 10.1, 10.2**)

You are given the 24 training data points on the figure denoted by + and − symbols. You are also given four blue testing points, denoted by squares, for which you have no labels. The coordinates of the blue squares are: Square A = (-2.2, -0.2); Square B = (-0.5, 0.5); Square C = (1 3); Square D = (3.5 -2).



(a) Determine the labels for the four numbered squares that you would obtain using the *nearest neighbor classifier* studied in class.

(b) Explain why the *linear discriminant analysis (LDA)* classifier has a high training error rate.

(c) Is there a linear classifier that can attain a training error less than 1/8? Justify your answer.

(d) Consider a *quadratic* classifier of the form:

$$D_{\mathrm{q}}(x_1, x_2) = \begin{cases} +1, & a \cdot x_1^2 + b \cdot x_1 x_2 + c \cdot x_2^2 \geq 0 \\ -1, & a \cdot x_1^2 + b \cdot x_1 x_2 + c \cdot x_2^2 < 0 \end{cases}$$

where the values +1 and −1 correspond to the + and − symbols, respectively. Select values for the coefficients a, b and c that would achieve zero training error. Clearly justify your reasoning. (**Hint:** Don't look for formulas in your sheets. Instead, look at the quadrants that the training points of each class are in.)

(e) Determine the testing labels for the four numbered squares that you would obtain using the classifier that you specified in part (d).