Problem 1

Open true_labels_pred_probs.csv (same file you used in HW4).

Calculate the points on the ROC curve. (2 points)

Plot the ROC curve. (1 points)

Print what is the best critical probability by finding the point on the ROC curve that's closest to the perfect classifier (0,1) point. (5 points)

In [ ]:

Problem 2:

In class we have seen both l1 (Lasso) and l2 (Ridge) regression. These are two basic ways to perform regularization. In the following problem, we will explore the **elastic net**, a third regularization technique that combines both l1 and l2 penalties. We will use this in a classification context.

The basic idea of the elastic net is that the cost function in regression becomes

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n}[(\theta_0 + \sum_{j=1}^{m}\theta_j x_{ij} - y_i)^2] + \frac{\alpha\rho}{m}\sum_{j=0}^{m}|\theta_j| + \frac{\alpha(1-\rho)}{m}\sum_{j=0}^{m}\theta_j^2,$$

and the cost function in classification is

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{n}[y_i \ln(\frac{1}{1+e^{-\theta_0+\sum_{j=1}^{m}\theta_j x_{ij}}}) + (1-y_i)\ln(1-\frac{1}{1+e^{-\theta_0+\sum_{j=1}^{m}\theta_j x_{ij}}}))] + \frac{\alpha\rho}{m}\sum_{j=0}^{m}|\theta_j| + \frac{\alpha(1-\rho)}{m}\sum_{j=0}^{m}\theta_j^2,$$

where α is the regularization parameter and ρ is the l1 ratio (how much weight we assign to the l1 term over the l2 term in the cost function). Basically, an elastic net uses the weighted sum of the l1 and l2 regularization terms. The weight of the l1 term is ρ and the weight of the l2 term is $(1-\rho)$ where ρ is between 0 and 1.

You can read more about the elastic net here (https://scikit-learn.org/stable/modules/linear_model.html#elastic-net).

Problem 2a

Load the training and test sets from train.csv and test.csv. (1 point)

Run a logistic regression model without regularization on the data and print the accuracy score of the test set. Use the 'saga' solver. (2 points)

In [ ]:

Problem 2b

Perform l1 regularization on the data.

The value of the alpha should contain 21 uniformly spaced values in log from 1e-2 to 1e2. (1 point)

Again, use the 'saga' solver and if you see a converge warning, fix it without ignoring the warning. (2 points)

Plot the train and test accuracy scores. (2 points)

Print the best test accuracy score and the corresponding alpha value. (2 points)

In [ ]:

Problem 2c

Perform l2 regularization on the data. The alpha values and all the other instructions are the same as in 2b. (1 point)

In [ ]:

Problem 2d

Let's train an elastic net now. The elastic net has two parameters: alpha and rho (the l1 ratio). The l1 ratio should be linearly spaced between 0 and 1 with 26 values in between. (1 point) Use the 'saga' solver. The reason we use the saga solver everywhere in Problem 2 is that it is the only solver in LogisticRegression that works with an elastic net.

Calculate the train and test accuracy scores for all combinations of alpha and rho. (4 points)

Print the best test score and the corresponding alpha and rho values. (3 points)

Prepare heatmaps to show the train and test scores. Make sure that the data range covered by the two heatmaps are the same so you can easily compare the two heatmaps and you can identify the high bias and high variance regions. Label the plot and add a colorbar. Make the x and y ticks look pretty. (3 points)

Which of the four approaches gave you the best test score?

In [ ]: