NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Week 14: Lecture 2
Classification

Center for Data Science

NYU

How can we use logistic regression for classification?

DS-UA 112
Introduction to Data Science

Week 14: Lecture 2

Classification

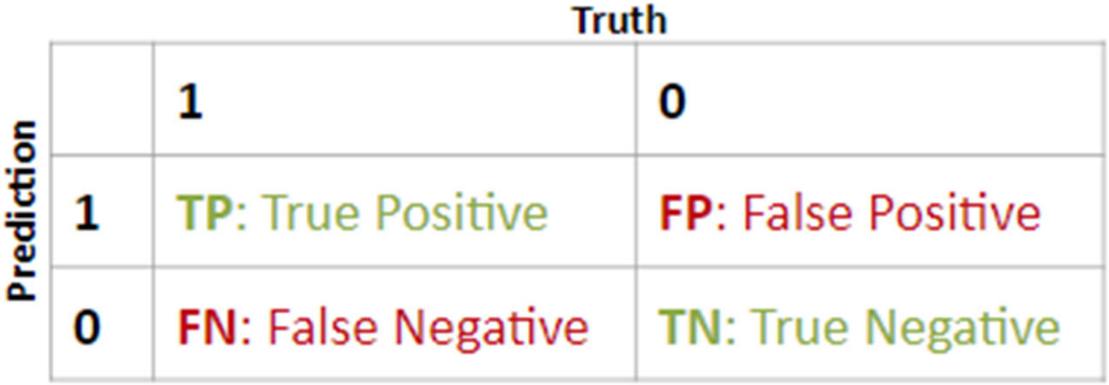
Adapted from Nolan, Speed, Gonzalez, Lau

Announcements

- Please check Week 14 agenda on NYU Classes
  - Lab 13
    - Due on Friday May 1 at 11:59PM EST
  - Project 2
    - Due on Tuesday May 12 at 11:59PM EST

Review

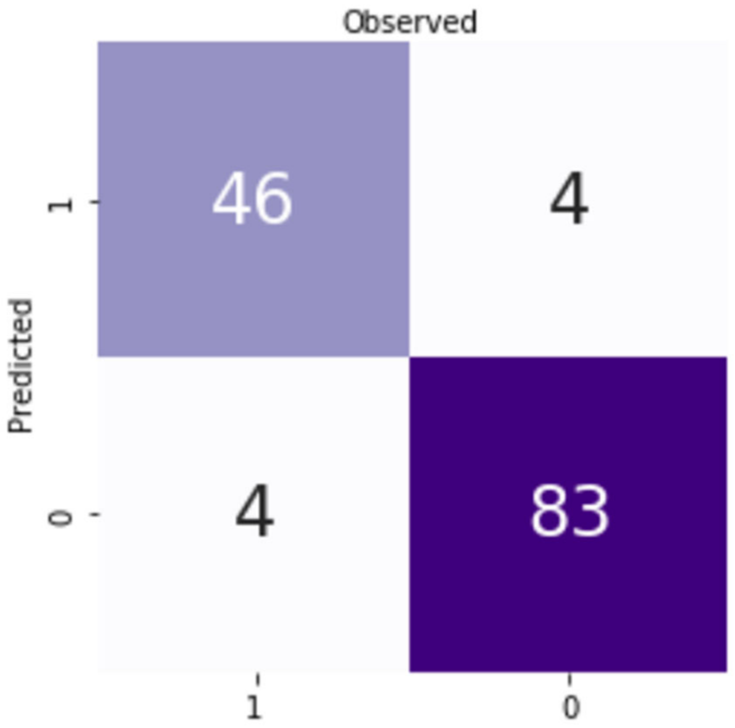
- The observation take the value 1 or 0. The predictions take the value 1 or 0. So we have four possibilities
  - True Positive
  - False Positive
  - False Negative
  - True Negative



		Truth	
Prediction		1	0
	1	**TP**: True Positive	**FP**: False Positive
	0	**FN**: False Negative	**TN**: True Negative

Review

- The observation take the value 1 or 0. The predictions take the value 1 or 0. So we have four possibilities
  - True Positive
  - False Positive
  - False Negative
  - True Negative
- We can visualize the number of each possibility for a dataset with a confusion matrix



	1	0
1	46	4
0	4	83

Review

- We can determine metrics from different combinations of these four possibilities.

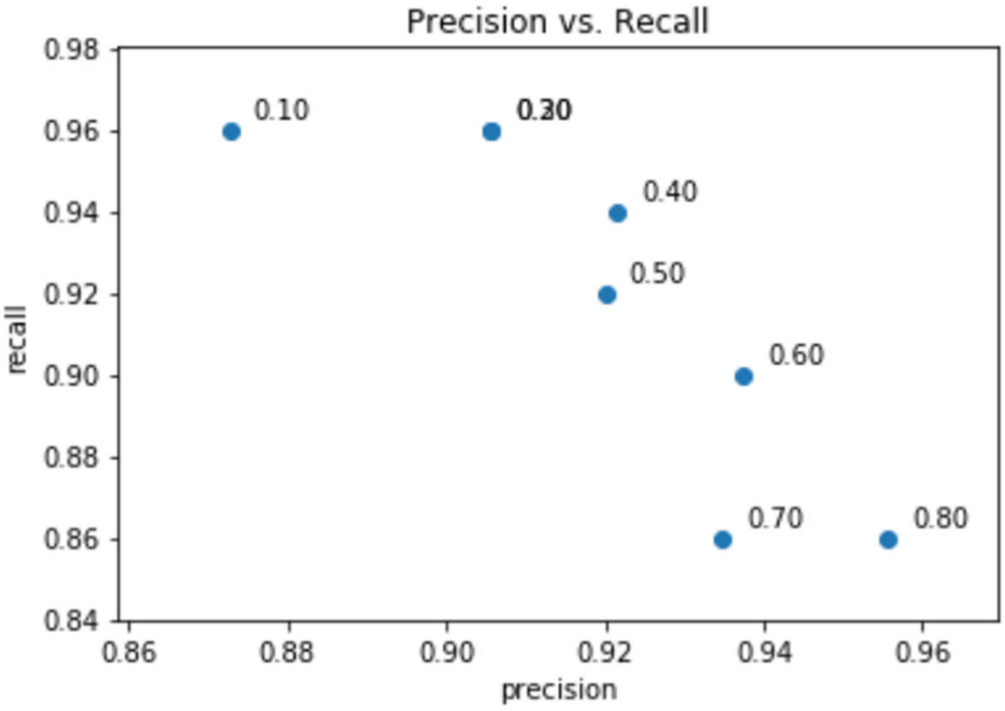
$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{n}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

Review

- Accuracy might not capture the differences between observations and prediction with an imbalance between categories
  - Precision penalizes false positives
  - Recall penalizes false negative
- We can visualize the trade-off between recall and precision through a precision-recall curve



Agenda

- Gradient Descent for Logistic Regression
- True Positive Rate and False Negative Rate
- Multiple Categories

ROC Curve

- The phrase true positive rate means recall

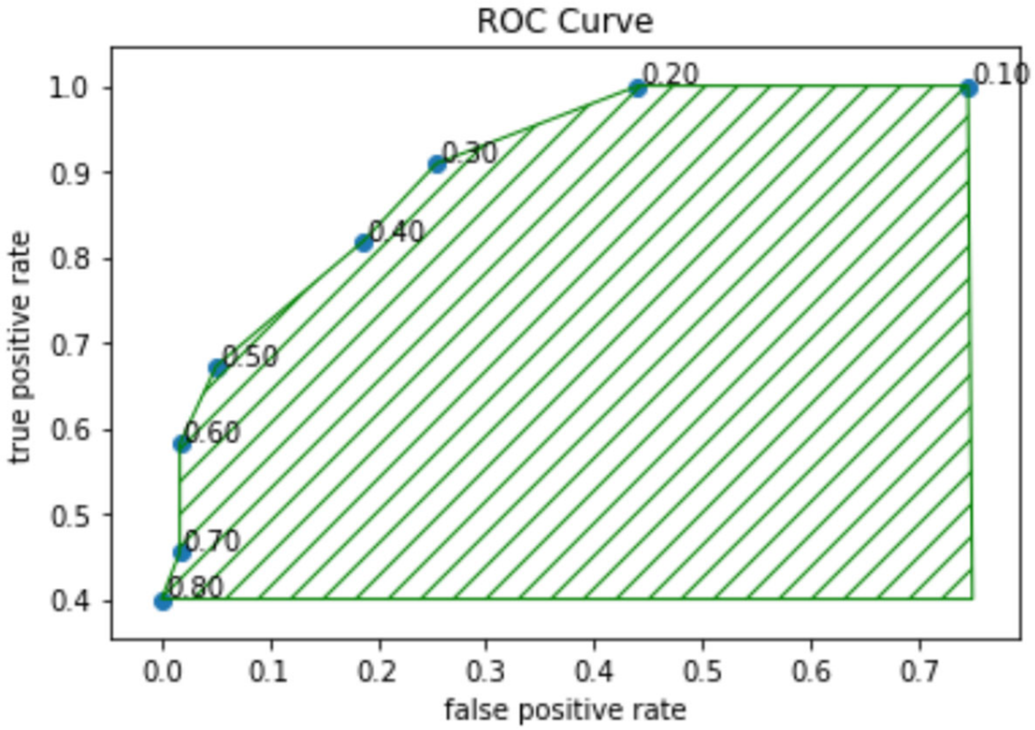
$$\text{True Positive Rate} = \frac{\#\text{True Positive}}{\#\text{True Positive} + \#\text{False Negative}}$$

- The false positive rate complements the true positive rate.

$$\text{False Positive Rate} = \frac{\#\text{False Positive}}{\#\text{True Negative} + \#\text{False Positive}}$$

ROC Curve

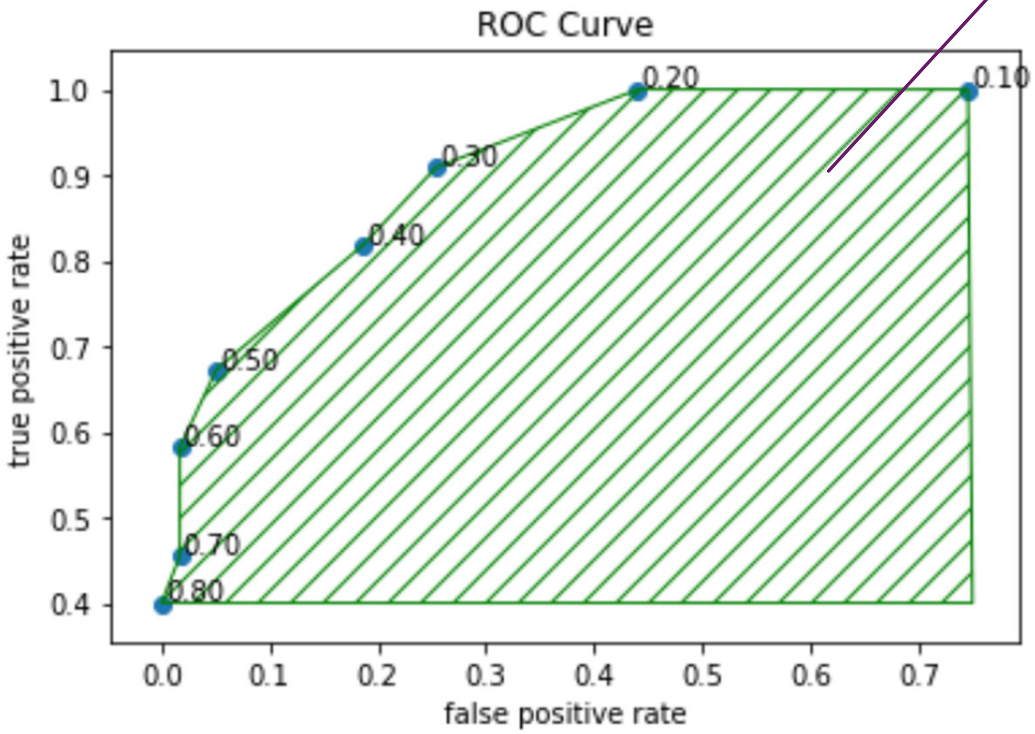
- A ROC curve plots the true positive rate and the false positive rate
- The acronym ROC stands for Receiver Operating Characteristic.
- We can summarize the ROC curve with the area under the curve. We abbreviate the area under the curve as AUC.



ROC Curve

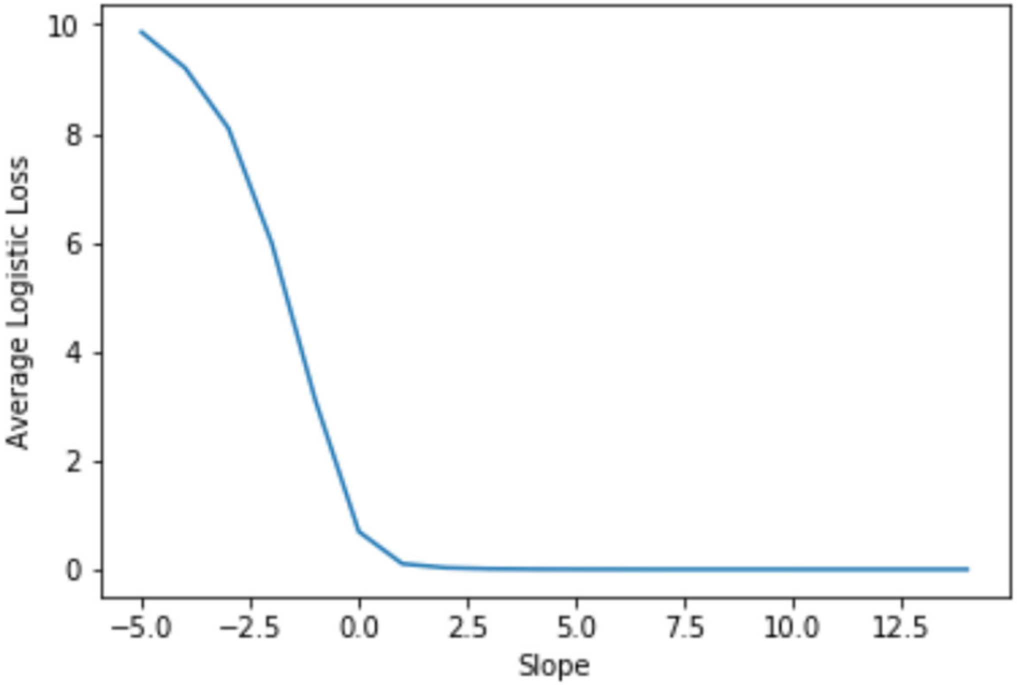
If AUC is close to 1, then we have high true positive rate and low false positive rate

- A ROC curve plots the true positive rate and the false negative rate
- The acronym ROC stands for Receiver Operating Characteristic.
- We can summarize the ROC curve with the area under the curve. We abbreviate the area under the curve as AUC.



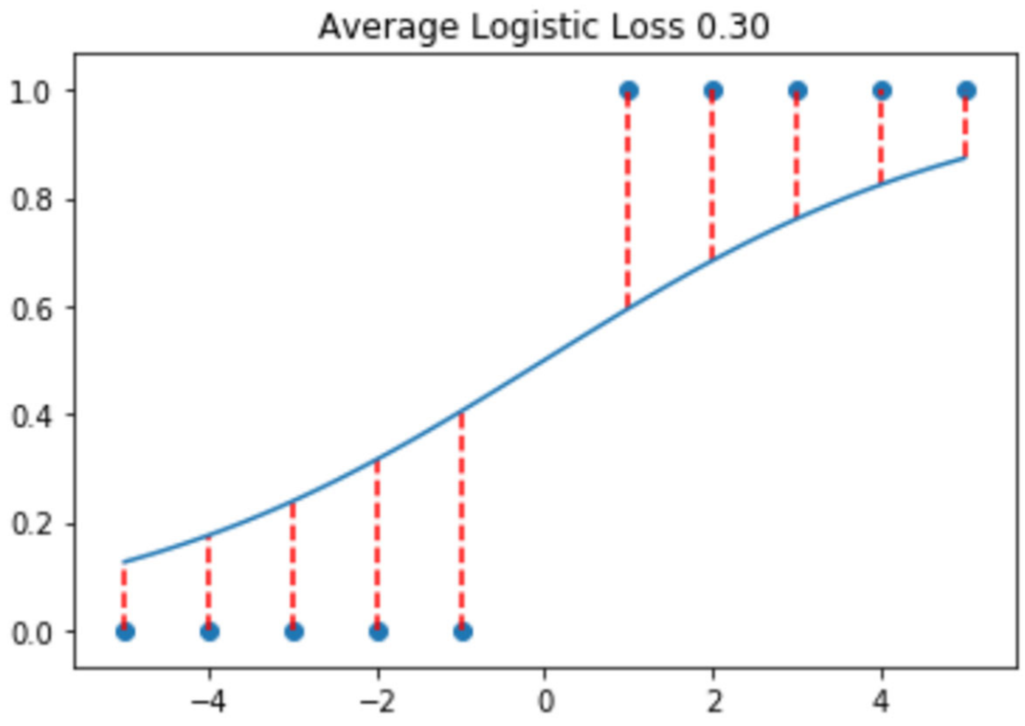
Gradient Descent

- The sigmoid function never attains the value 0 or 1. So the average logistic loss might not attain its minimum value.
- If we can completely separate the two categories into regions divided by a decision boundary, then we need to add regularization for convergence of gradient descent



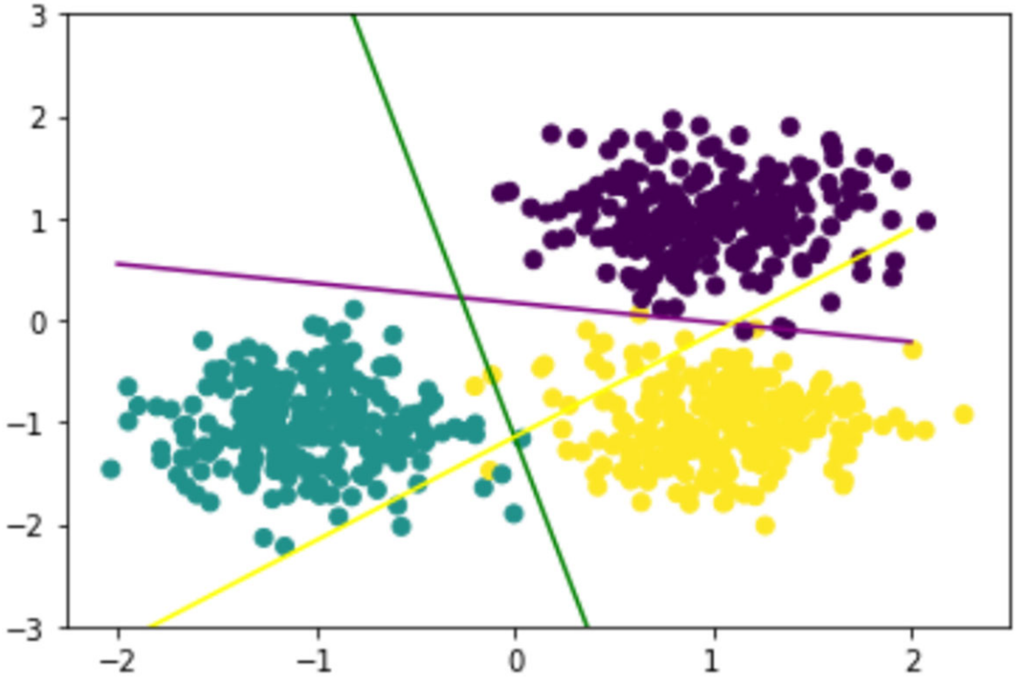
Gradient Descent

- The sigmoid function never attains the value 0 or 1. So the average logistic loss might not attain its minimum value.
- If we can completely separate the two categories into regions divided by a decision boundary, then we need to add regularization for convergence of gradient descent



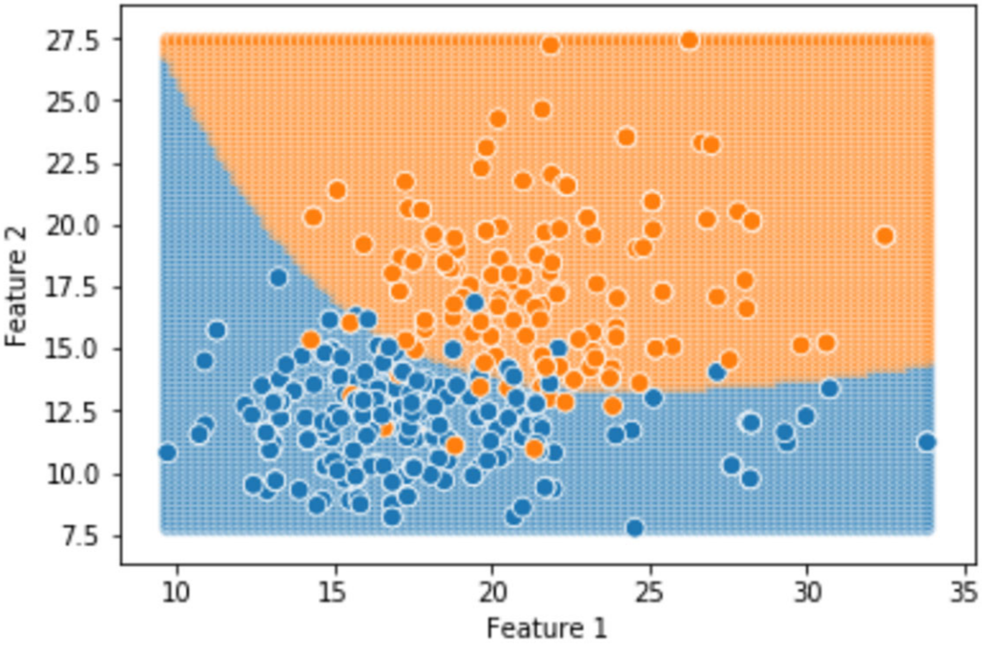
Multiple Categories

- If we have three or more categories, then we can split the classification problem into multiple problems with two categories.
- Each problem try to classify one category versus the other categories. We call the approach One-versus-Rest.



Decision Boundaries

- Remember that we can transform the features in a linear regression model to fit data with a nonlinear shape
- Similarly we can transform the features in a logistic regression model to obtain a curved decision boundary.
- Sometimes we want the decision boundary to bend around the regions containing the two categories



Summary

- Gradient Descent for Logistic Regression
- True Positive Rate and False Negative Rate
- Multiple Categories

Goals

- Understand the need for regularization in logistic regression
- Generate a ROC curve
- Use One-versus-Rest approach for classification into three or more categories