Model Evaluation and Metrics

Jay Urbain, PhD

Topics

- Regression
  - Assessing the accuracy of model coefficients
  - RMSE – Root Mean Squared Error
- Classification
  - Confusion matrix
  - ROC Curve

Review: Assessing the accuracy of model coefficients

Linear regression with residual term. Represents what we can't explain with our model.

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

RSS measures the amount of variability that is left unexplained after performing the regression

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

TSS (Total sum of squares) measures the total variance when measuring the response *y*.

$$\mathrm{TSS} = \sum (y_i - \bar{y})^2$$

R^2 amount of variance explained by our model

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

The *RSE* is an estimate of the standard deviation of ε. It is basically the average amount that the response will deviate from the true regression line.

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

Root Mean Squared Error

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- Used for regression problems
- Square root of the mean of the squared errors
- Easily interpretable (in the "y" units)
- "Punishes" larger errors
- Other: *absolute error*

Root Mean Squared Error

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Example:
y_true = [100, 50, 30]
y_preds = [90, 50, 50]
RMSE = np.sqrt((10**2 + 0**2 + 20**2) / 3) = 12.88

Confusion Matrix: table to describe the performance of a classifier

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Example: Test for presence of disease
NO = negative test = False = 0
YES = positive test = True = 1

- How many classes are there?
- How many patients?
- How many times is disease predicted?
- How many patients actually have the disease?

Confusion Matrix: table to describe the performance of a classifier

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Basic Terminology:

- True Positives (TP)
- True Negatives (TN)
- False Positives (FP) - Type I Error
- False Negatives (FN) - Type II Error

Accuracy:

- Overall, how often is it **correct**?
- (TP + TN) / total = 150/165 = 0.91

Misclassification Rate (Error Rate):

- Overall, how often is it **wrong**?
- (FP + FN) / total = 15/165 = 0.09

Confusion Matrix: table to describe the performance of a classifier

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

True Positive Rate, Sensitivity, Recall:

- When actual value is **positive**, how often is prediction **correct**?
- TPR = TP / T = 100/105 = 0.95
- "True Positive Rate" or "Recall"

False Positive Rate, Fall-out:

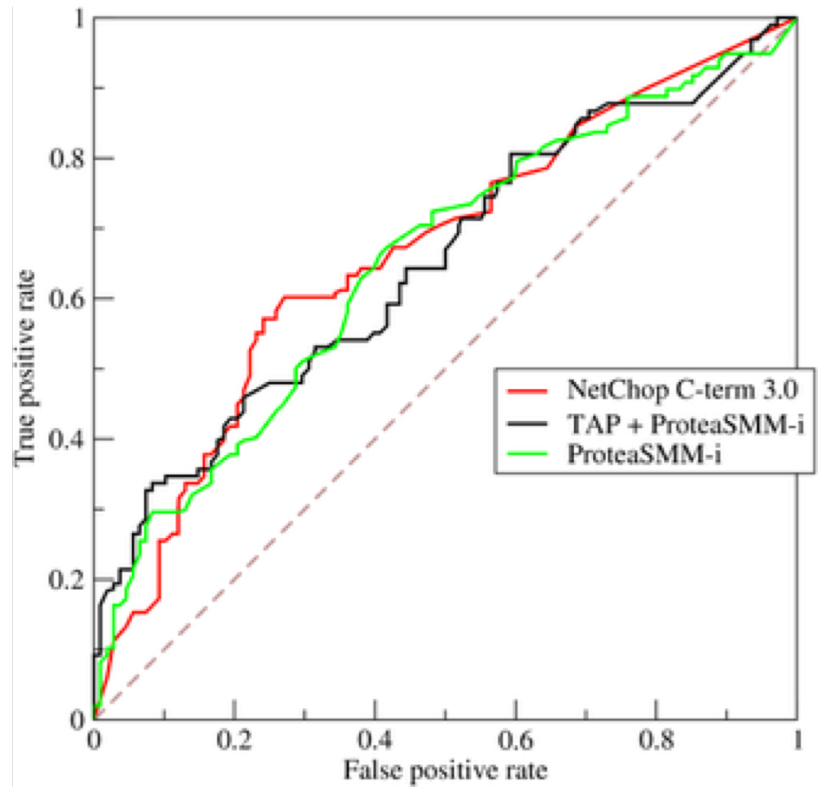
- When actual value is **negative**, how often is prediction **wrong**?
- FPR = (FP / F) = 10/60 = 0.17

Specificity, True Negative Rate:

- When actual value is **negative**, how often is prediction **correct**?
- TNR = (TN / F) = 50/60 = 0.83

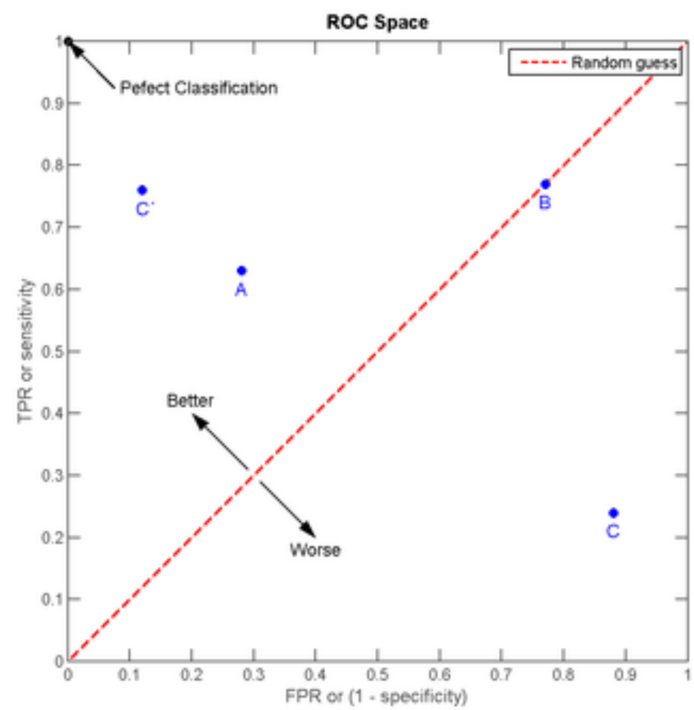
Receiver operating characteristic (ROC) Curve

- The ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.
- The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.



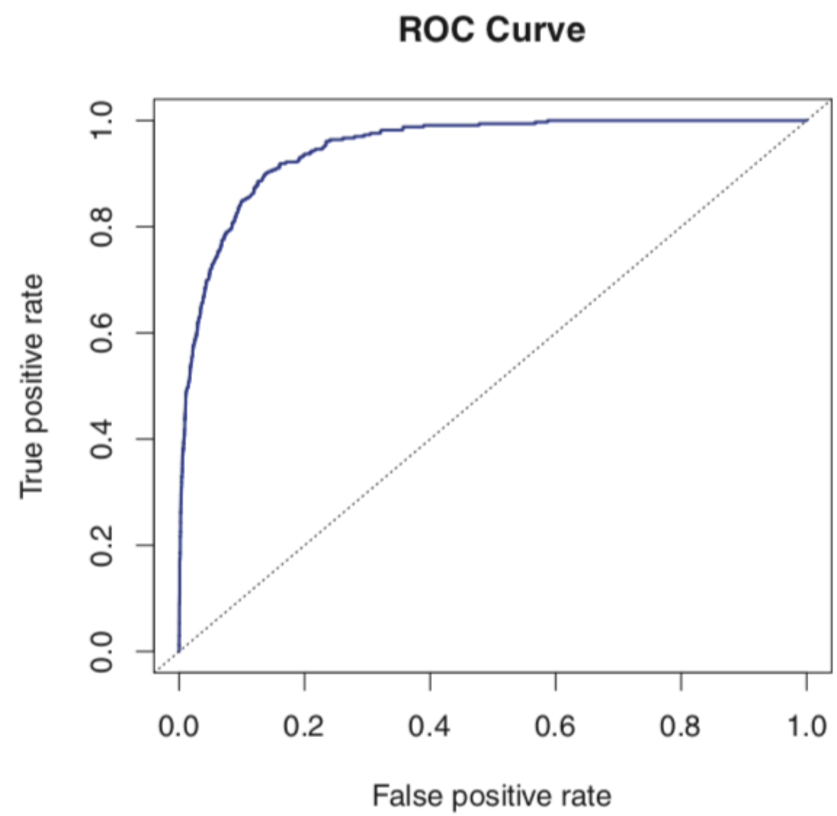
Receiver operating characteristic (ROC) Curve

- To draw a ROC curve, only the true positive rate (TPR) and false positive rate (FPR) are needed (as functions of some classifier parameter).
- A ROC space is defined by FPR and TPR as *x* and *y* axes, respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs).
- The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives).
- The (0,1) point is also called a *perfect classification*.

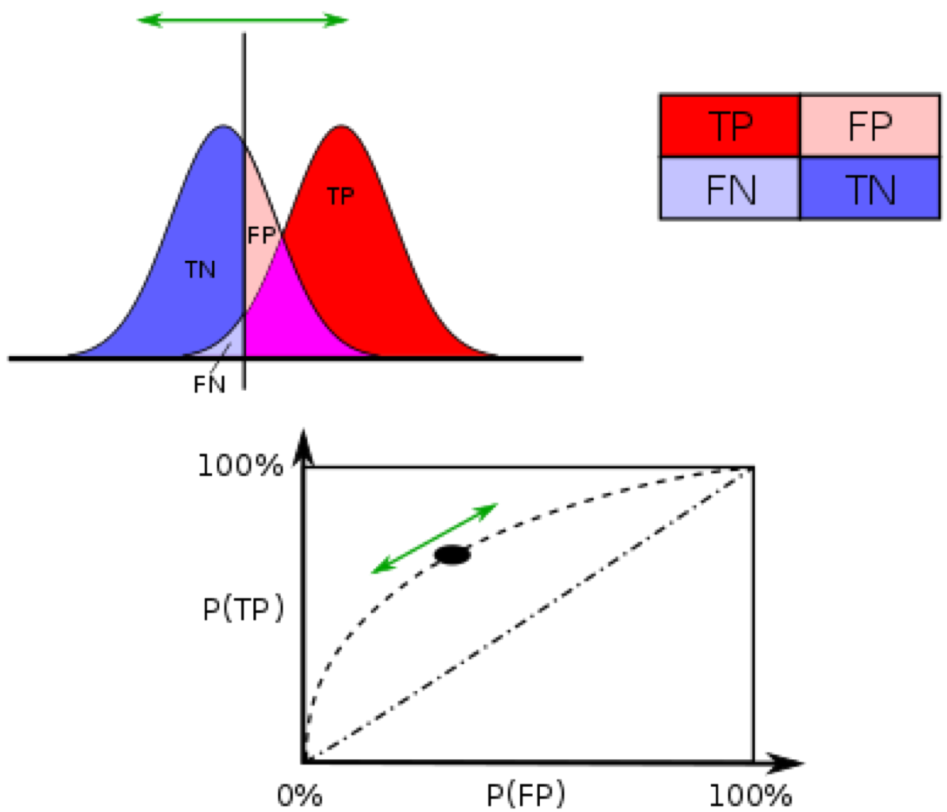


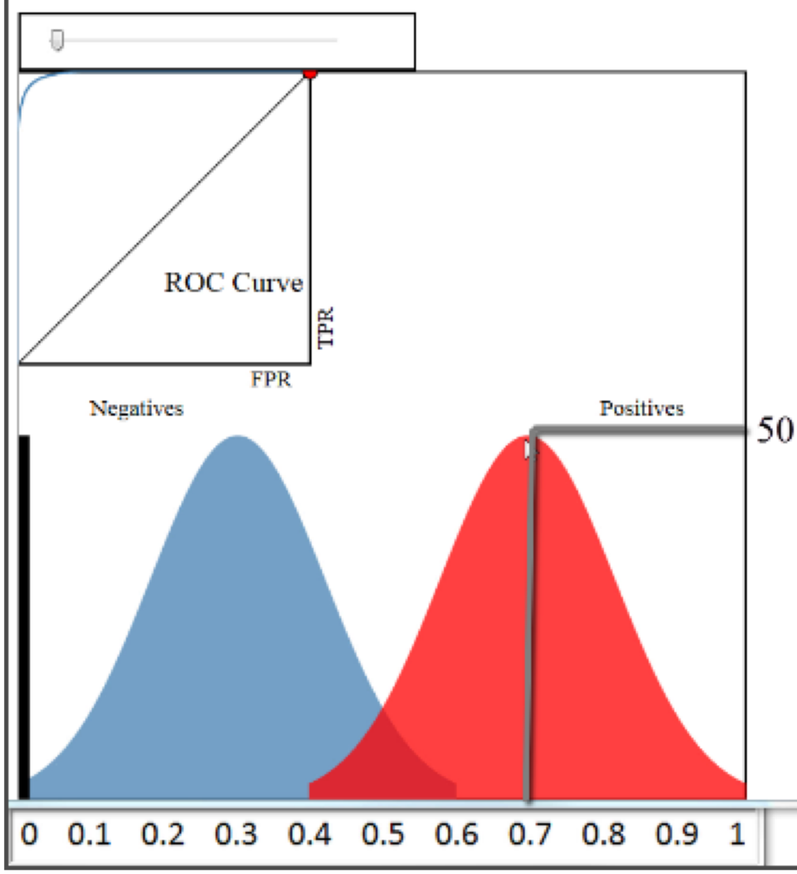
ROC Curves

- The overall performance of a classifier summarized over all possible thresholds, is given by the area under the ROC curve (AUC).
- An ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier.
- A classifier performing not better than chance would have an AUC of 0.5



ROC Space

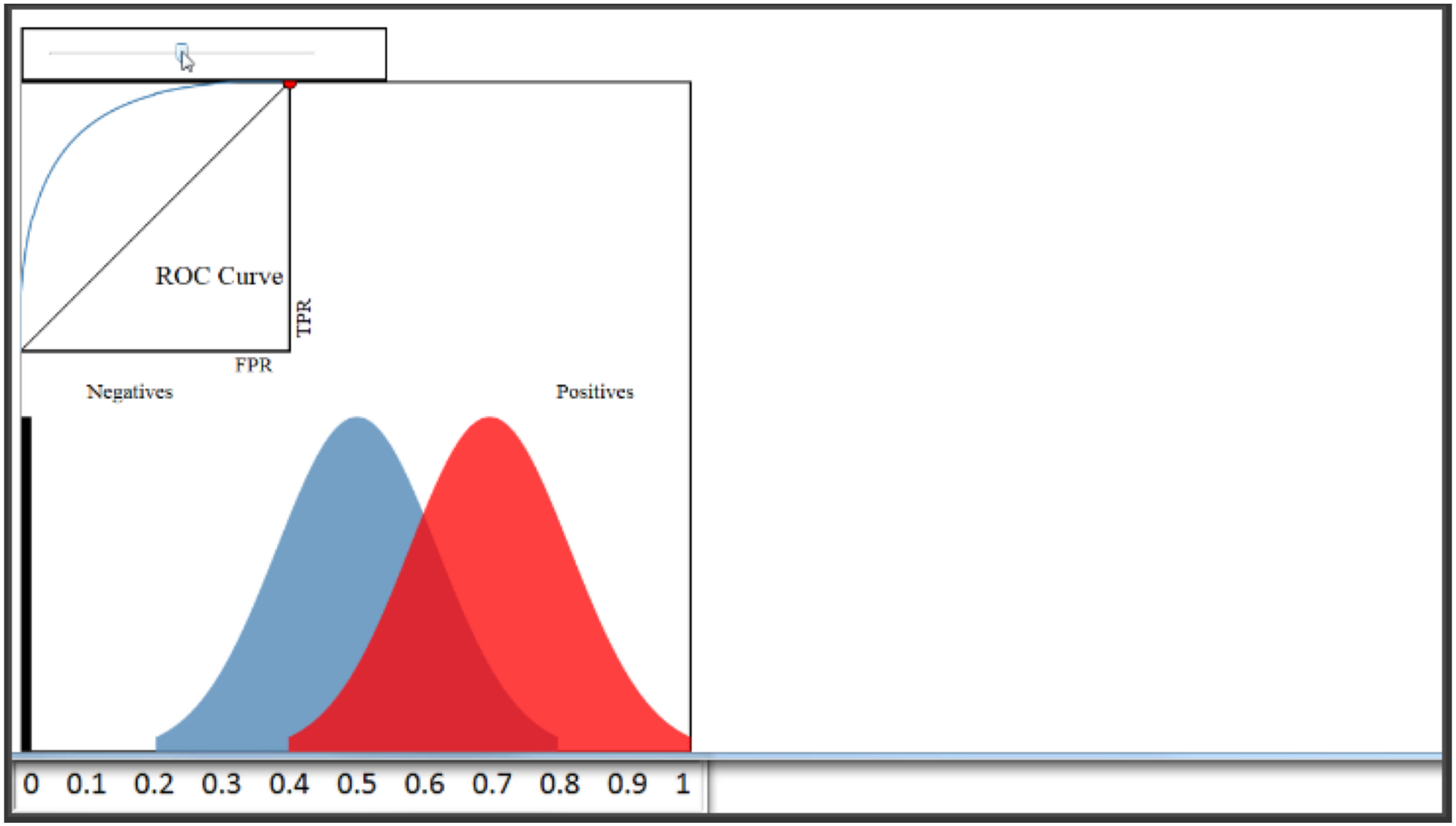




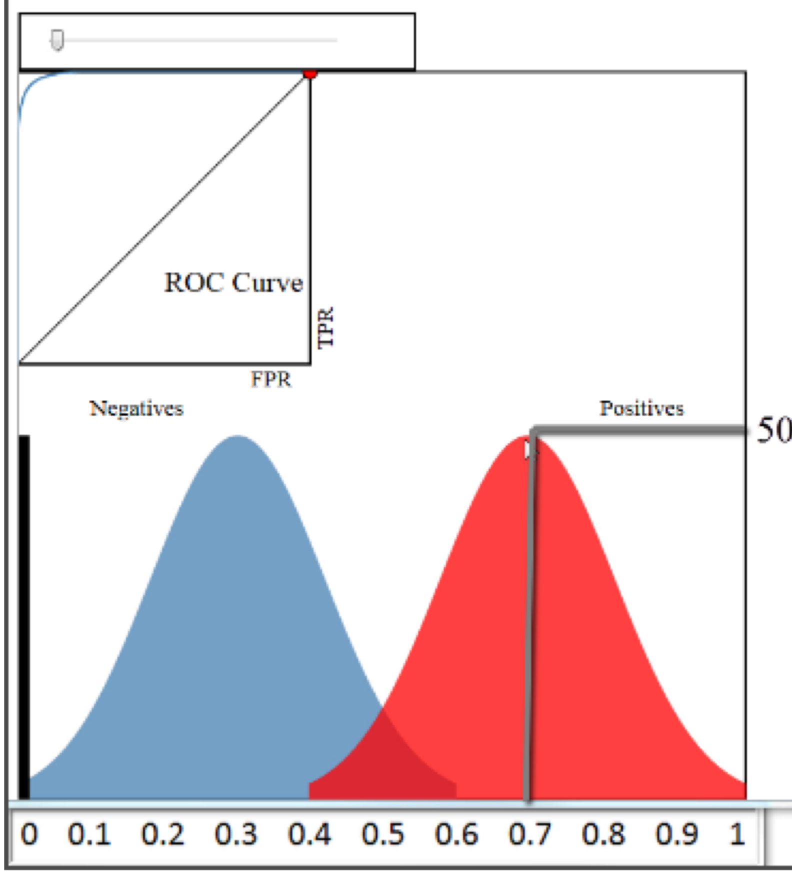
Remember:

blue pixel = "not admitted"
red pixel = "admitted"

http://www.navan.name/roc/



ROC Curve
TPR
FPR
Negatives
Positives
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1



Remember:

blue pixel = "not admitted"
red pixel = "admitted"

ROC Curve

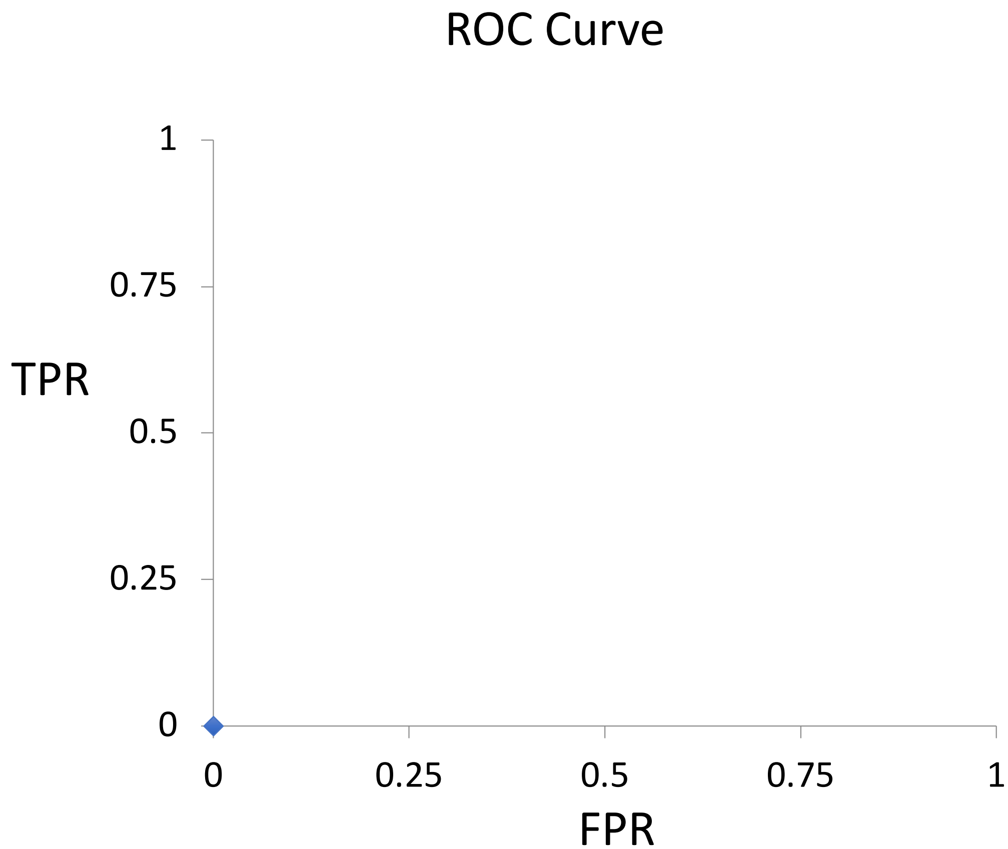
Email Number	Score	True Label
5	0.99	Spam
8	0.82	Spam
2	0.60	Spam
1	0.60	Ham
7	0.48	Spam
3	0.22	Ham
4	0.10	Ham
6	0.02	Ham

Every email is assigned a “spam” score by our classification algorithm. To actually make our predictions, we choose a numeric cutoff for classifying as spam.

An ROC Curve will help us visualize how well our classifier is doing without having to choose a cutoff!

ROC Curve

Email Number	Score	True Label
5	0.99	Spam
8	0.82	Spam
2	0.60	Spam
1	0.60	Ham
7	0.48	Spam
3	0.22	Ham
4	0.10	Ham
6	0.02	Ham



ROC Curve

https://www.dataschool.io/roc-curves-and-auc-explained/

Email Number	Score	True Label
5	0.99	Spam
8	0.82	Spam
2	0.60	Spam
1	0.60	Ham
7	0.48	Spam
3	0.22	Ham
4	0.10	Ham
6	0.02	Ham

TPR: When actual value is **spam**, how often is prediction **correct**?

FPR: When actual value is **ham**, how often is prediction **wrong**?

Cutoff	TPR (y)	FPR (x)	Cutoff	TPR (y)	FPR (x)
0			**0.50**		
0.05			**0.65**		
0.15			**0.85**		
0.25			**1**		

ROC Curve

Email Number	Score	True Label
5	0.99	Spam
8	0.82	Spam
2	0.60	Spam
1	0.60	Ham
7	0.48	Spam
3	0.22	Ham
4	0.10	Ham
6	0.02	Ham

TPR: When actual value is **spam**, how often is prediction **correct**?

FPR: When actual value is **ham**, how often is prediction **wrong**?

Cutoff	TPR (y)	FPR (x)	Cutoff	TPR (y)	FPR (x)
0	1	1	**0.50**	0.75	0.25
0.05	1	0.75	**0.65**	0.5	0
0.15	1	0.5	**0.85**	0.25	0
0.25	1	0.25	**1**	0	0

ROC Curve

Email Number	Score	True Label
5	0.99	Spam
8	0.98	Spam
2	0.97	Spam
1	0.97	Ham
7	0.96	Spam
3	0.95	Ham
4	0.94	Ham
6	0.93	Ham

Q: Would the ROC Curve (and AUC) change if the **scores** changed, but the **ordering** remained the same?

A: Not at all! The ROC Curve is only sensitive to **rank ordering** and does not require **calibrated scores**.