Nuts & Bolts of Machine Learning

Ashis Kumer Biswas, Ph.D.

Part b : Components of Learning
- b.3 : Model evaluation

Outlines

- Components of Machine Learning

Learning Components

- A Task
- Training examples
- Evaluation metric(s)

Task: classify fruits, find groups, find the leader in the group, etc.

Training example: the dataset , the data set should reflect on the task
- set of fruits
- a population of humans/animals
- set of groups

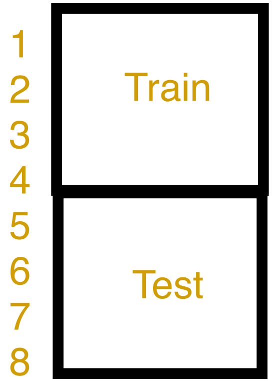
Model Evaluation

classification
↳ supervised learning

- Confusion Matrix
- Performance metrics:
  - Accuracy
  - Precision
  - Recall
  - F1 score
  - …
  - ROC
  - AUC

Accuracy is the number one metric we want to go for

Evaluation metrics



Data set is composed of:
- training set
- test set
- only working with the training set

Test set

0: Negative
1 : Positive

Acurracy = TP + TN/ TP + TN + FP + FN
Inaccuracy = FP + FN/ TP + TN + FP + FN

Sample ID	Actual Label		Predicted Label
Sample 1	1	correct	1
Sample 2	0	correct	0
Sample 3	1	failed	0
Sample 4	1	failed	0
Sample 5	0	failed	1
Sample 6	0	failed	1
Sample 7	1	failed	0
Sample 8	1	correct	1
Sample 9	0	correct	0
Sample 10	1	correct	1

TP	TN	FP	FN
1	0	0	0
0	1.	0.	0
0.	0.	0.	1
0.	0.	0.	1
0.	0.	1.	0
0.	0.	1.	0
0.	0.	0.	1
1	0.	0.	0
0.	1.	0.	0
1.	0.	0.	0
3	2	2	3

All add up to 10/# of samples

5 correct classifications/ 10

Accuracy: 5/10

True positive: predict 1 and actual is 1
True Negative: predict 0 and actual is 0
False pos: predict 1 but actual is 0
False neg: predict 0 but actual is 1

Metrics for (binary) classification performance evaluation

- Focus on the predictive capability of a model:
    - Rather than how fast it takes to classify or building the model, scales, etc.

- First prepare the confusion matrix:

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

    - TP = number of true positives
    - TN = number of true negatives
    - FP = number of false positives
    - FN = number of false negatives

Just a pen-paper exercise

Sample ID	Actual Label	Predicted Label
Sample 1	1	1
Sample 2	0	0
Sample 3	1	0
Sample 4	1	0
Sample 5	0	1
Sample 6	0	1
Sample 7	1	0
Sample 8	1	1
Sample 9	0	0
Sample 10	1	1

Metrics for classification performance evaluation

confusion matrix

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Most widely used performance metric:

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

Limitation of Accuracy

- Consider a 2-class problem (class A and class B):
  - Number of class A examples $= 9990$
  - Number of class B examples $= 10$

Class distribution

- If model predicts everything to be class A, the accuracy is $\frac{9990}{10000} = 99.9\%$
  - Accuracy is misleading because model does not detect any class B example.
  - So, when the class sizes are not even, accuracy is not a reliable performance measure.

Cost matrix

- Cost matrix is similar to the confusion matrix, except the fact that we will be calculating the cost of wrong predictions and/or right predictions.

T	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	**Class=Yes** = 0	**Class=No** miss-classify = 100 +
	Class=Yes	C(Yes|Yes)	$\boxed{\text{C(No|Yes)}}$
	Class=No	C(Yes|No) = 10	C(No|No) = 0

Cost Matrix

	+	-
+	TP	FN
-	FP	TN

Confusion Matrix

$C(i|j)$ = cost of mis-classifying class j sample as class i.

What is the cost of predicting misclassifying classes?
- predicting a negative sample as positive
- [predicting a positive sample as negative
- add more costs to mis-classfifying classes

Intuition behind it... Let's think about it

Subject id	Actual Cancer	Predicted Cancer
Subject 1	0	0
Subject 2	0	1
Subject 3	0	0
Subject 4	1	1
Subject 5	1	0
...	...	...

- What do you think what will be the cost of missclassifications?
    - For Subject 2?
    - For Subject 5?

Yet another example to promote your thought process

Confusion Matrix

Let's evaluate two infection prediction models: A and B.

Predicted: A

		+	-
Actual	+	150	170
	-	50	630

Predicted: B

		+	-
Actual	+	150	20
	-	200	630

Some costs:

1. Tests for an infection \$2,000 Everyone has to do this first
2. Sanitizing a room and moving a patient to a new room: \$5,000 Move to new room if infection is detected early or late
3. Treating an infection early: \$20,000 Failed to detect an infection early
4. Treating an infection late: \$30,000 Failed to detect an infection late

Yet another example to promote your thought process

Let's evaluate two infection prediction models: A and B.

Predicted: A

		+	-
Actual	+	150	170
	-	50	630

Predicted: B

		+	-
Actual	+	150	20
	-	200	630

1. If a patient is predicted to not have an infection and truly does not, then there is no cost. $cost(TN) = \$0$
2. If a patient is predicted to have an infection and does not then, $cost(FP) = \$2,000$, i.e., cost of the test only.
3. If a patient is predicted to not have an infection, but does, then $cost(FN) = \$37,000$. Can you deduce it? Treating an infection late
4. If a patient is predicted to have an infection and does, then $cost(TP) = \$27,000$. Can you deduce it? Treating an infection early

Computing cost of classification

Cost Matrix	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	+	-
	+	-1	100
	-	1	0

model keeps predicting false negatives

Model M_1	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	150	40
	-	60	250

Accuracy = 80%

Cost = 3910

Model M_2	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	250	45
	-	5	200

Accuracy = 90%

Cost = 4255

Cost sensitive performance measures

- Precision, $p = \frac{TP}{TP+FP} = 1 - FDR$
    - Also known as the Positive Predictive Value, PPV.
    - Proportion of predicted positive samples that belongs to the ground true positive samples.
    - It is biased towards $C(+|+)$ and $C(+|-)$.

Cost sensitive performance measures

Confusion Matrix

TP	FP
FN	TN

- Recall, $r = \frac{TP}{TP+FN}$
  - Also known as sensitivity, True Positive Rate (TPR)
  - Proportion of the ground true positive samples that are predicted.
  - It is biased towards $C(+|+)$ and $C(-|+)$

Recall emphasizes on positive predictions

TP + FP = models positive predictions

Cost sensitive performance measures

- Specificity, $Sp = \frac{TN}{N} = \frac{TN}{TN+FP}$
  - Also known as selectivity, True Negative Rate (TNR)

Cost sensitive performance measures

Precision = TP/ TP + FP

Recall = TP/P

P = TP + FN

F1 = harmonic mean of precision and recall

- F_1 measure, $F_1 = 2 \cdot \frac{pr}{p+r} = \dfrac{2TP}{2TP + FN + FP}$
  - It is biased towards all except $C(-|-)$.
  - When, TP=0, $F_1 = 0$
  - When TP=FN=FP=0, then F_1 is undefined.
  - When FP=FN $= 0$, then F_1 is 1.

0 = worst
1 = best

the higher the precision and recall, the better

Cost sensitive performance measures

Predic.

TP FN

FP TN

Bigger FDR = worse

- False Discovery Rate, $FDR = \dfrac{FP}{FP + TP} = 1 - precision$
  - It is the proportion of false discoveries (i.e., False positives) among the total discoveries (i.e., all positive predictions).

Cost sensitive performance measures

- Matthews's Correlation Coefficient, MCC measure,

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

  - When $MCC = -1$, there is a perfect disagreement between actual and predictions, and when $MCC = +1$, there is a perfect agreement.
  - When $MCC = 0$, the prediction may as well be regarded similar to a random prediction.
  - If any of the 4 sums in the denominator is zero, the denominator can be arbitrarlity set to 1 which will make MCC $= 0$.
  - Think! when we have a very negative MCC (i.e., very lose to -1).

MCC = 0 is undesirable
- equal to a random prediction

Outlines

More evaluation metrics of a classifier

Given the confusion matrix:

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

- Precision, $p = \frac{TP}{TP+FP}$

More evaluation metrics of a classifier

Given the confusion matrix:

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

- Precision, $p = \frac{TP}{TP+FP}$
  - Proportion of predicted positive samples (TP) out of all predicted positives ($TP + FP$).

More evaluation metrics of a classifier

Given the confusion matrix:

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

- Precision, $p = \frac{TP}{TP+FP}$
  - Proportion of predicted positive samples (TP) out of all predicted positives ($TP + FP$).
  - Also known as Positive Predictive Value, PPV

More evaluation metrics of a classifier

Given the confusion matrix:

		Prediction	
		Positive	**Negative**
Actual	**Positive**	TP	FN
	Negative	FP	TN

- Recall, $r = \frac{TP}{TP+FN} = \frac{TP}{P}$
  - Proportion of successfully predicted positive samples (TP) to total number of actual positives ($TP + FN = P$).

More evaluation metrics of a classifier

Given the confusion matrix:

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

- Recall, $r = \frac{TP}{TP+FN} = \frac{TP}{P}$
  - Proportion of successfully predicted positive samples (TP) to total number of actual positives ($TP + FN = P$).
  - also known as, True Positive Rate, TPR

More evaluation metrics of a classifier

Given the confusion matrix:

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

- Recall, $r = \frac{TP}{TP+FN} = \frac{TP}{P}$
  - Proportion of successfully predicted positive samples (TP) to total number of actual positives ($TP + FN = P$).
  - also known as, True Positive Rate, TPR
  - also known as, Sensitivity, Sn

More evaluation metrics of a classifier

Given the confusion matrix:

		Prediction	
		Positive	**Negative**
Actual	**Positive**	TP	FN
	Negative	FP	TN

- True Negative Rate, $TNR = \frac{TN}{TN+FP}$
  - Proportion of predicted negative samples (TN) that are actually negative ($TN + FP$).

More evaluation metrics of a classifier

Given the confusion matrix:

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

- True Negative Rate, $TNR = \frac{TN}{TN+FP}$
  - Proportion of predicted negative samples (TN) that are actually negative ($TN + FP$).
  - also known as, Specificity, Sp

More evaluation metrics of a classifier

Given the confusion matrix:

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

- False positive Rate, $FPR = \frac{FP}{TN+FP} = \frac{FP}{N}$

More evaluation metrics of a classifier

Given the confusion matrix:

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

- False positive Rate, $FPR = \frac{FP}{TN+FP} = \frac{FP}{N}$
  - Proportion of predicted false positive samples (FP) that are actually negatives ($TN + FP = N$).

More evaluation metrics of a classifier

Given the confusion matrix:

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

- False positive Rate, $FPR = \frac{FP}{TN+FP} = \frac{FP}{N}$
  - Proportion of predicted false positive samples (FP) that are actually negatives ($TN + FP = N$).
  - $$\begin{aligned} FPR &= 1 - Specificity \\ &= 1 - \frac{TN}{TN + FP} \\ &= \frac{\cancel{TN} + FP - \cancel{TN}}{TN + FP} \\ &= \frac{FP}{TN + FP} \end{aligned}$$

True Positive rate vs. False Positive rate

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

$\sum = P$

$\sum = N$

- True Positive rate is $TPR = \dfrac{TP}{P}$
  - TPR = # of correctly predicted positives / # of positives in the test data

- Note note:
  TP + FN = P
  FP + TN = N

True Positive rate vs. False Positive rate

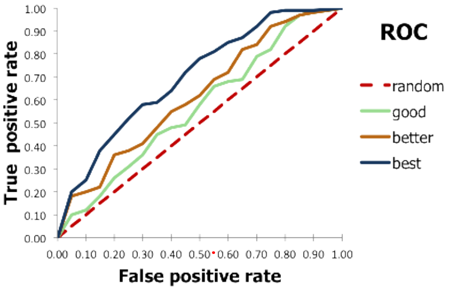
		Prediction	
		Positive	**Negative**
Actual	**Positive**	TP	FN
	Negative	FP	TN

$$\sum = P$$

$$\sum = N$$

- True Positive rate is $TPR = \dfrac{TP}{P}$
  - TPR = # of correctly predicted positives / # of positives in the test data
- False Positive rate is $FPR = \dfrac{FP}{N}$
  - TPR = # of incorrectly predicted positives / # of negatives in the test data
- Note note:
  TP + FN = P
  FP + TN = N

Receiver Operating Characteristics (ROC) curve



- It is a classifier performance plotting method.
- Used to compare the relative performance among different classifiers.
- ROC is a 2-dimensional graph plotting *TPR* against the *FPR*.
- It depicts relative trade-offs between –
  - benefits (true positive rate) and
  - costs (false positive rate)

ROC space

ROC for the classifiers that predicts only class label (e.g, Decision trees), without the thresholding:

- Each of these classifier has only a single (FPR,TPR) pair that we plot as a single point in the ROC space.

		Prediction by A Pos	Prediction by A Neg	
Actual	Pos	$TP = 63$	$FN = 37$	100
Actual	Neg	$FP = 28$	$TN = 72$	100
	Total	91	109	200

- FPR $= 0.28$
- TPR $= 0.63$
- Accuracy $= 0.68$

ROC space

ROC for the classifiers that predicts only class label (e.g, Decision trees), without the thresholding:

- Each of these classifier has only a single (FPR,TPR) pair that we plot as a single point in the ROC space.

		Prediction by B		
		Pos	Neg	
Actual	Pos	$TP = 77$	$FN = 23$	100
	Neg	$FP = 77$	$TN = 23$	100
	Total	154	46	200

- FPR $= 0.77$
- TPR $= 0.77$
- Accuracy $= 0.50$

ROC space

ROC for the classifiers that predicts only class label (e.g, Decision trees), without the thresholding:

- Each of these classifier has only a single (FPR,TPR) pair that we plot as a single point in the ROC space.

		Prediction by C Pos	Neg	
Actual	Pos	$TP = 24$	$FN = 76$	100
	Neg	$FP = 88$	$TN = 12$	100
	Total	112	88	200

- FPR $= 0.88$
- TPR $= 0.24$
- Accuracy $= 0.18$

ROC space

ROC for the classifiers that predicts only class label (e.g, Decision trees), without the thresholding:

- Each of these classifier has only a single (FPR,TPR) pair that we plot as a single point in the ROC space.

		Prediction by C		
		Pos	Neg	
Actual	Pos	$TP = 24$	$FN = 76$	100
	Neg	$FP = 88$	$TN = 12$	100
	Total	112	88	200

- FPR $= 0.88$
- TPR $= 0.24$
- Accuracy $= 0.18$
- It is so bad! Ohhh!!! wait a minute...

ROC space

ROC for the classifiers that predicts only class label (e.g, Decision trees), without the thresholding:

- Each of these classifier has only a single (FPR,TPR) pair that we plot as a single point in the ROC space.

Prediction by C

		Pos	Neg	
Actual	Pos	$TP = 24$	$FN = 76$	100
	Neg	$FP = 88$	$TN = 12$	100
	Total	112	88	200

- FPR $= 0.88$
- TPR $= 0.24$
- Accuracy $= 0.18$
- It is so bad! Ohhh!!! wait a minute...
  - Let's flip all your predictions – say all "Yes"s to "No"s, and all "No"s to "Yes"s.

ROC space

ROC for the classifiers that predicts only class label (e.g, Decision trees), without the thresholding:

- Each of these classifier has only a single (FPR,TPR) pair that we plot as a single point in the ROC space.

Prediction by C_{rev}

		Pos	Neg	
Actual	Pos	$TP = 76$	$FN = 24$	100
	Neg	$FP = 12$	$TN = 88$	100
	Total	88	112	200

- FPR $= 0.12$
- TPR $= 0.76$
- Accuracy $= 0.82$

ROC space

ROC for the classifiers that predicts only class label (e.g, Decision trees), without the thresholding:

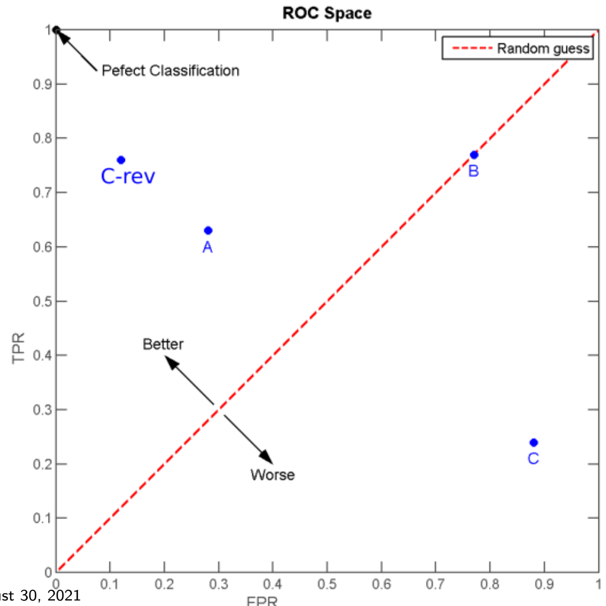
- Each of these classifier has only a single (FPR,TPR) pair that we plot as a single point in the ROC space.

Prediction by C_{rev}

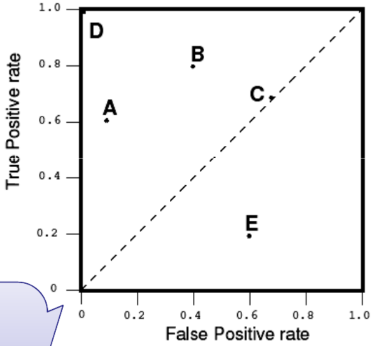
		Pos	Neg	
Actual	Pos	$TP = 76$	$FN = 24$	100
	Neg	$FP = 12$	$TN = 88$	100
	Total	88	112	200

- FPR $= 0.12$
- TPR $= 0.76$
- Accuracy $= 0.82$
- It became an awesome classifier!!!

ROC plot of the four classifiers, A, B, C, C_{rev}



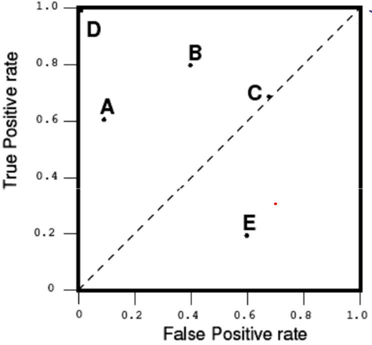
Lower Left point (0,0)



Never issue a positive classification!

such a classifier commits
no false positive errors
but also gains
no true positives.

Upper Right point (1,1)



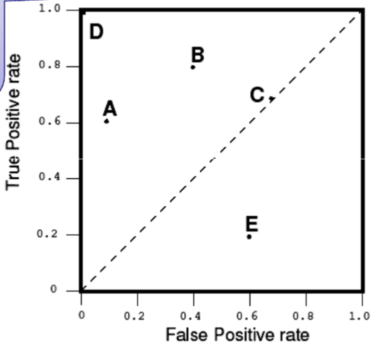
Unconditionally issue positive classification!

such a classier predicts
all positive instances correctly
but at the cost of predicting
all negative instances wrongly

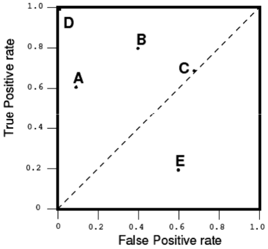
Point at (0,1)

Get everything perfect!

this perfect classier commits
no false positive errors
and gets
all true positives

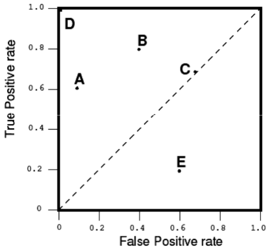


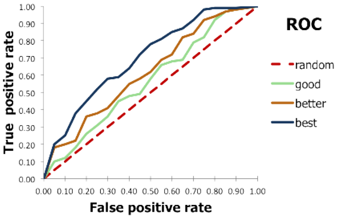
Several notes on the ROC space



- A point in ROC space is better than another if it is to the **northwest** of the other, i.e.,
  - TPR is higher.
  - FPR is lower.
- Any classifier that appears in the lower right triangle performs worse than random guessing

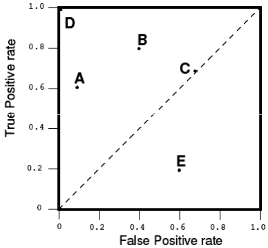
Points vs. Curves in the ROC space

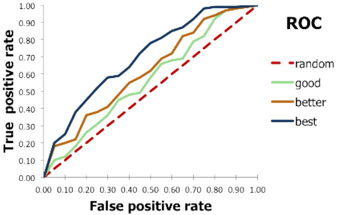




- Many classifiers are discrete classifiers, such as decision trees, kNN that are designed to produce only a target class, i.e., either Yes, or a No on each sample.
  - For such a classifier is applied on a test set, it produces a single confusion matrix, which in turn corresponds to a single **ROC point**.

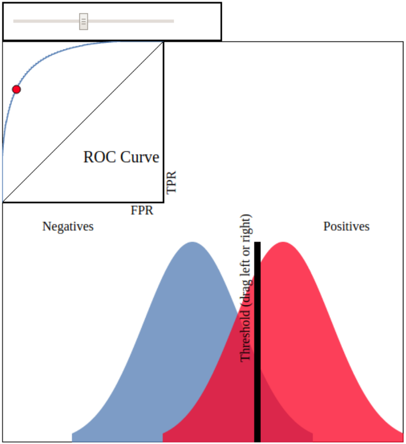
Points vs. Curves in the ROC space

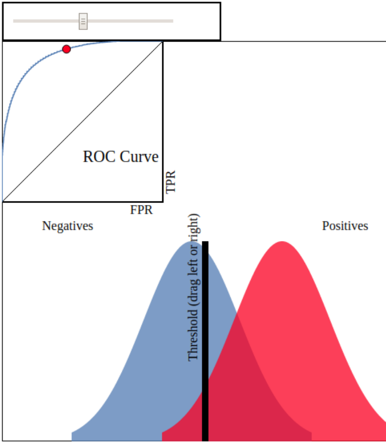




- Whereas, some classifiers, such as Naïve Bayes, Logistic regression, yield probability or some kind of scores before assigning a target class label to a sample.
  - Such a ranking or scoring classifier can be used with a threshold to produce a discrete classifier:
    - if the classifier output score is above the threshold, the classifier produces a Yes.,
    - otherwise it produces a No
  - Each different threshold value produces a different point in the ROC space (corresponding to a different confusion matrices).

Behavior of ROC curve vs distribution of classes

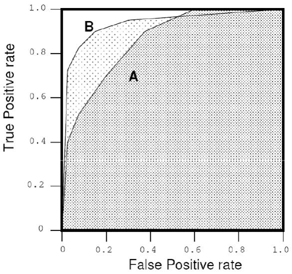


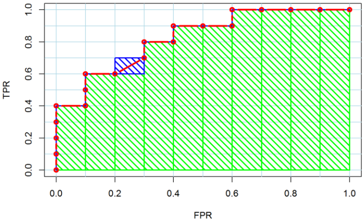


Please adjust the mean of the two distributions (positive and negative) and also the threshold here:
`http://www.navan.name/roc/` (Last checked: 08-30-2021 1:37PM MST)

Area Under an ROC Curve, AUC

- AUC is often used to compare classifiers:
  - The bigger the AUC the better.
- AUC can be computed by the "trapezoidal rule" once you have the ROC curve.
  - More on "trapezoidal rule" method – `https://en.wikipedia.org/wiki/Trapezoidal_rule`
  - `http://blog.revolutionanalytics.com/2016/11/calculating-auc.html`





Thanks
Questions?