ECEN 377: Engineering Applications of AI

Dr. Mahmoud Nabil Mahmoud
mnmahmoud@ncat.edu

North Carolina A & T State University

October 13, 2024

Outline

Classification Accuracy

Accuracy: *"The ratio between the number of correctly predicted data points and the total number of data points"*

Classification Accuracy

Accuracy: *"The ratio between the number of correctly predicted data points and the total number of data points"*

- **Example:** If we evaluate a model on a test dataset of 1000 samples, and the model predicted the correct label 875 times:

Classification Accuracy

Accuracy: *"The ratio between the number of correctly predicted data points and the total number of data points"*

- **Example:** If we evaluate a model on a test dataset of 1000 samples, and the model predicted the correct label 875 times:
  - Accuracy $= \frac{875}{1000}$ = 0.875 or 87.5%

Classification Accuracy

Accuracy: *"The ratio between the number of correctly predicted data points and the total number of data points"*

- **Example:** If we evaluate a model on a test dataset of 1000 samples, and the model predicted the correct label 875 times:
  - Accuracy $= \frac{875}{1000} = 0.875$ or 87.5%
- Ordering models by accuracy (best to worst):

Classification Accuracy

Accuracy: *"The ratio between the number of correctly predicted data points and the total number of data points"*

- **Example:** If we evaluate a model on a test dataset of 1000 samples, and the model predicted the correct label 875 times:
  - Accuracy $= \frac{875}{1000} = 0.875$ or 87.5%
- Ordering models by accuracy (best to worst):
  1. 70% accuracy

Classification Accuracy

Accuracy: *"The ratio between the number of correctly predicted data points and the total number of data points"*

- **Example:** If we evaluate a model on a test dataset of 1000 samples, and the model predicted the correct label 875 times:
  - Accuracy $= \frac{875}{1000} = 0.875$ or 87.5%
- Ordering models by accuracy (best to worst):
  1. 70% accuracy
  2. 50% accuracy

Classification Accuracy

Accuracy: *"The ratio between the number of correctly predicted data points and the total number of data points"*

- **Example:** If we evaluate a model on a test dataset of 1000 samples, and the model predicted the correct label 875 times:
  - Accuracy $= \frac{875}{1000} = 0.875$ or 87.5%
- Ordering models by accuracy (best to worst):
  1. 70% accuracy
  2. 50% accuracy
  3. 15% accuracy

Classification Accuracy

Accuracy: *"The ratio between the number of correctly predicted data points and the total number of data points"*

- **Example:** If we evaluate a model on a test dataset of 1000 samples, and the model predicted the correct label 875 times:
  - Accuracy $= \frac{875}{1000} = 0.875$ or 87.5%
- Ordering models by accuracy (best to worst):
  1. 70% accuracy
  2. 50% accuracy
  3. 15% accuracy
- Is accuracy always enough for model evaluation?

Classification Accuracy

Accuracy: *"The ratio between the number of correctly predicted data points and the total number of data points"*

- **Example:** If we evaluate a model on a test dataset of 1000 samples, and the model predicted the correct label 875 times:
  - Accuracy $= \frac{875}{1000} = 0.875$ or 87.5%
- Ordering models by accuracy (best to worst):
  1. 70% accuracy
  2. 50% accuracy
  3. 15% accuracy
- Is accuracy always enough for model evaluation?
  - No, accuracy alone can be misleading, especially for imbalanced datasets.

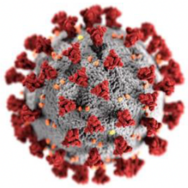
Classification Accuracy

Accuracy: *"The ratio between the number of correctly predicted data points and the total number of data points"*

- **Example:** If we evaluate a model on a test dataset of 1000 samples, and the model predicted the correct label 875 times:
  - Accuracy $= \frac{875}{1000} = 0.875$ or 87.5%
- Ordering models by accuracy (best to worst):
  1. 70% accuracy
  2. 50% accuracy
  3. 15% accuracy
- Is accuracy always enough for model evaluation?
  - No, accuracy alone can be misleading, especially for imbalanced datasets.
  - Other metrics like precision, recall, and F1-score are often necessary.

Example Datasets

- **Medical Dataset**
  - A set of patient diagnosed with coronavirus
  - A medical dataset with 1000 persons
  - 10 diagnosed as "sick" with coronavirus
  - 990 diagnosed as "healthy"



- **Email Dataset**
  - A set of emails labeled spam or ham
  - A dataset of 100 emails
  - 40 are "spam"
  - 60 are "ham"

Limitations of Accuracy

- Let's revisit our question: **"Is accuracy always enough for model evaluation?"**
- Consider this scenario:
  - "I have developed a classifier for coronavirus dataset that":
    - takes not much time to run
    - doesn't require any examinations
    - has an accuracy of 99%!"
  - What do you think?

Limitations of Accuracy

- Let's revisit our question: **"Is accuracy always enough for model evaluation?"**
- Consider this scenario:
  - "I have developed a classifier for coronavirus dataset that":
    - takes not much time to run
    - doesn't require any examinations
    - has an accuracy of 99%!"
  - What do you think?
- On our coronavirus dataset:
  - If we classify all samples as healthy, our model accuracy is 99%!
  - This demonstrates how accuracy can be misleading for imbalanced datasets.

Limitations of Accuracy

- Let's revisit our question: **"Is accuracy always enough for model evaluation?"**
- Consider this scenario:
  - "I have developed a classifier for coronavirus dataset that":
    - takes not much time to run
    - doesn't require any examinations
    - has an accuracy of 99%!"
  - What do you think?
- On our coronavirus dataset:
  - If we classify all samples as healthy, our model accuracy is 99%!
  - This demonstrates how accuracy can be misleading for imbalanced datasets.

Limitations of Accuracy

- Let's revisit our question: **"Is accuracy always enough for model evaluation?"**
- Consider this scenario:
  - "I have developed a classifier for coronavirus dataset that":
    - takes not much time to run
    - doesn't require any examinations
    - has an accuracy of 99%!"
  - What do you think?
- On our coronavirus dataset:
  - If we classify all samples as healthy, our model accuracy is 99%!
  - This demonstrates how accuracy can be misleading for imbalanced datasets.

Thresholding

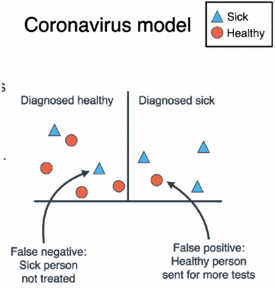
- In order to map a logistic regression value to a binary category, we must define a classification threshold.
- Classification threshold is problem-dependent.
- It does not have to be 0.5.
- Classification metrics are used to define the classification threshold.

Outline

False Positives & False Negatives

For coronavirus dataset:

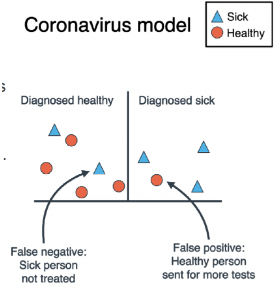
- **True positive:**
  - A sick person that gets diagnosed as sick.
- **True negative:**
  - A healthy person that gets diagnosed as healthy.
- **False positive:**
  - A healthy person that gets incorrectly diagnosed as sick.
- **False negative:**
  - A sick person that gets incorrectly diagnosed as healthy.



False Positives & False Negatives

For coronavirus dataset:

- **True positive:**
  - A sick person that gets diagnosed as sick.
- **True negative:**
  - A healthy person that gets diagnosed as healthy.
- **False positive:**
  - A healthy person that gets incorrectly diagnosed as sick.
- **False negative:**
  - A sick person that gets incorrectly diagnosed as healthy.

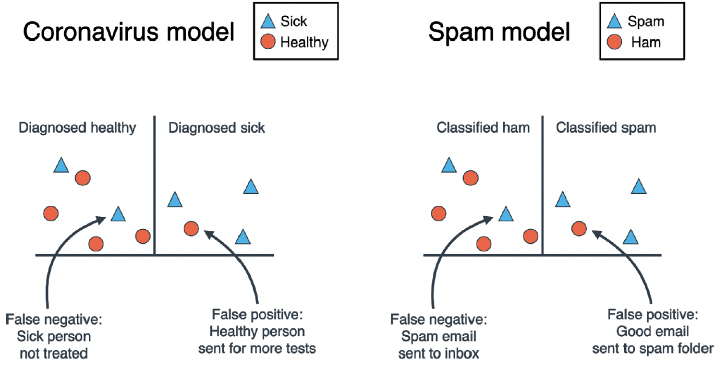


Example:

- 3 true positives
- 4 true negatives
- 1 false positive
- 2 false negatives

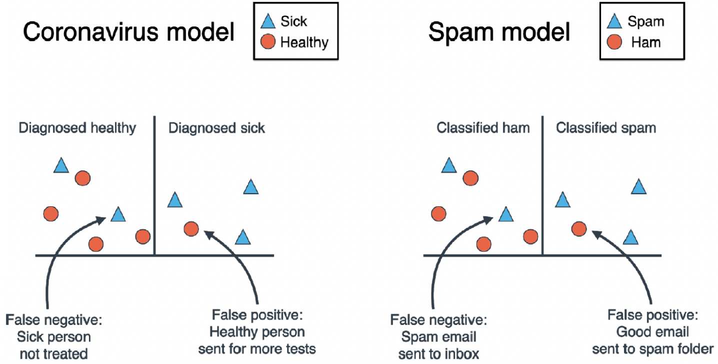
False positives & False negatives

Which is important for each dataset False positives or False negatives?



False positives & False negatives

Which is important for each dataset False positives or False negatives?



- In coronavirus it is more important to not have undetected sick people So false negative is more important
- In spam model it is more important to not have ham mails in junk box So false positive is more important

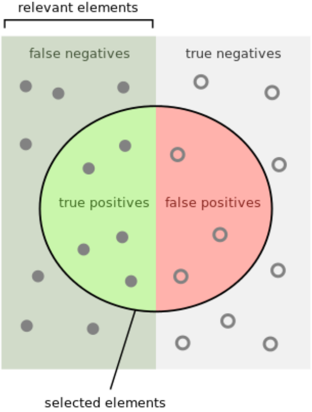
Outline

Confusion Matrix

- Our model is confused between 2 classes.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)



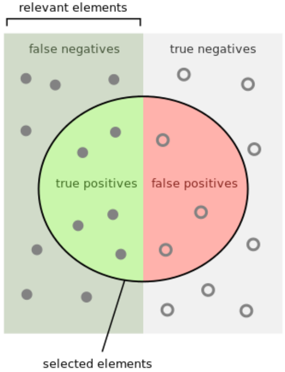
We want large diagonal, small FP, FN

Outline

Accuracy and Error

- **Accuracy** is the total correct prediction

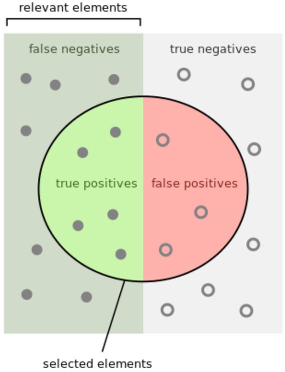
	Actually Positive (1)	Actually Negative (0)	
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)	P'
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)	N'
	P	N	



Accuracy and Error

- **Accuracy** is the total correct prediction
  - $\frac{TP+TN}{TP+TN+FP+FN}$

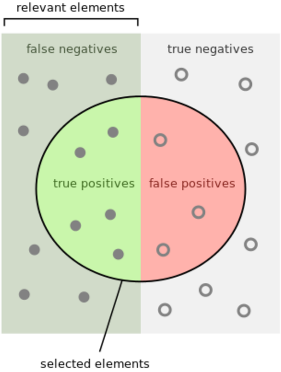
	Actually Positive (1)	Actually Negative (0)	
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)	P'
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)	N'
	P	N	



Accuracy and Error

	Actually Positive (1)	Actually Negative (0)	
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)	P'
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)	N'
	P	N	

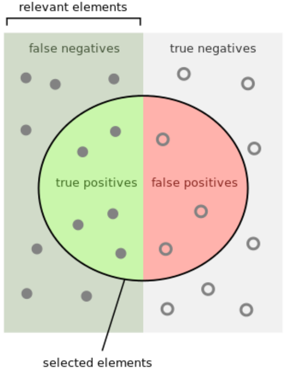
- **Accuracy** is the total correct prediction
    - $\frac{TP+TN}{TP+TN+FP+FN}$
- **Error** is the total false prediction



Accuracy and Error

	Actually Positive (1)	Actually Negative (0)	
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)	P'
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)	N'
	P	N	

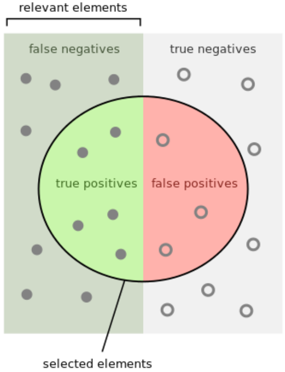
- **Accuracy** is the total correct prediction
  - $\frac{TP+TN}{TP+TN+FP+FN}$
- **Error** is the total false prediction
  - $\frac{FP+FN}{TP+TN+FP+FN} = 1$ - Accuracy



Accuracy and Error

	Actually Positive (1)	Actually Negative (0)	
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)	P'
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)	N'
	P	N	

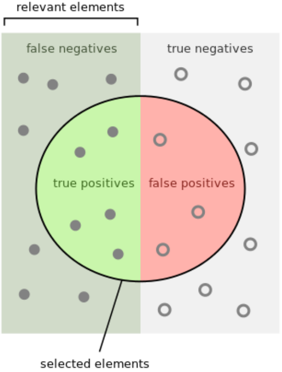
- **Accuracy** is the total correct prediction
  - $\frac{TP+TN}{TP+TN+FP+FN}$
- **Error** is the total false prediction
  - $\frac{FP+FN}{TP+TN+FP+FN} = 1$ - Accuracy
- **Problem:** cannot handle unbalanced classes



Accuracy and Error

	Actually Positive (1)	Actually Negative (0)	
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)	P'
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)	N'
	P	N	

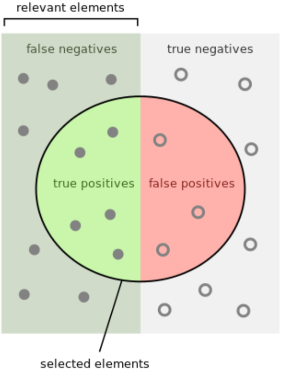
- **Accuracy** is the total correct prediction
  - $\frac{TP+TN}{TP+TN+FP+FN}$
- **Error** is the total false prediction
  - $\frac{FP+FN}{TP+TN+FP+FN} = 1$ - Accuracy
- **Problem:** cannot handle unbalanced classes
  - Predict whether an earthquake is about to happen



Accuracy and Error

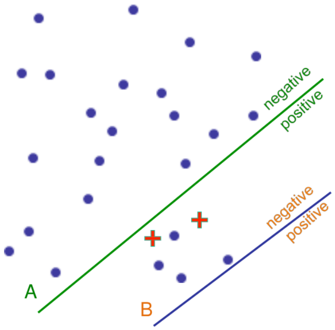
	Actually Positive (1)	Actually Negative (0)	
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)	P'
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)	N'
	P	N	

- **Accuracy** is the total correct prediction
  - $\frac{TP+TN}{TP+TN+FP+FN}$
- **Error** is the total false prediction
  - $\frac{FP+FN}{TP+TN+FP+FN} = 1$ - Accuracy
- **Problem:** cannot handle unbalanced classes
  - Predict whether an earthquake is about to happen
  - Happen very rarely, very good accuracy if always predict "No".



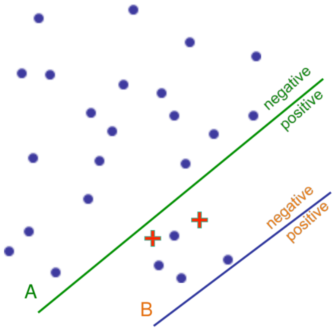
Problem with Accuracy

- You're predicting cancer possiblity (+) vs. not (•)



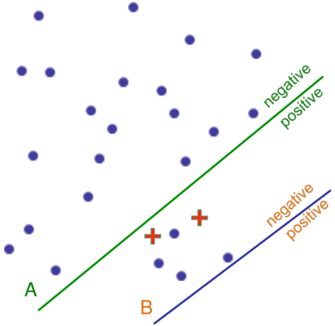
Problem with Accuracy

- You're predicting cancer possiblity ($+$) vs. not ($\bullet$)
- Accuracy will prefer classifier B (fewer errors)



Problem with Accuracy

- You're predicting cancer possiblity (+) vs. not (•)
- Accuracy will prefer classifier B (fewer errors)
- Classifier A is better though.



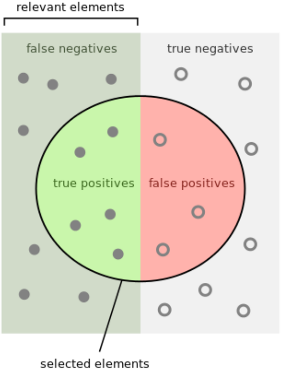
Outline

1. Classification Accuracy
2. False Positives and False Negatives
3. Confusion Matrix
   - Accuracy and Error
   - Sensitivity and Miss Rate
   - Specificity and False Alarm
   - Precision and F1-measure
4. ROC curves
5. AUC (Area Under the Curve)
6. Decision based on ROC
7. Exercise

Metrics (1/3)

	Actually Positive (1)	Actually Negative (0)	
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)	P'
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)	N'
	P	N	

- **Sensitivity** How many (+) we hit? (Hit rate = Recall = Sensitivity = True pos rate)
  - $\frac{TP}{P} = \frac{TP}{TP+FN}$
- **Miss Rate** How many (+) we miss? (Miss rate = False neg rate = false rejection = type II error rate)
  - $1 - hitrate = \frac{FN}{P} = \frac{FN}{TP+FN}$

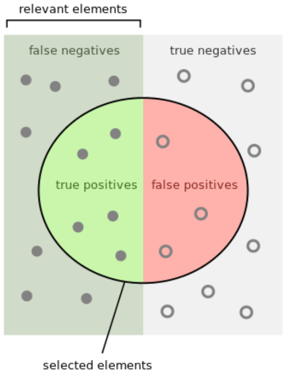


Outline

Metrics (2/3)

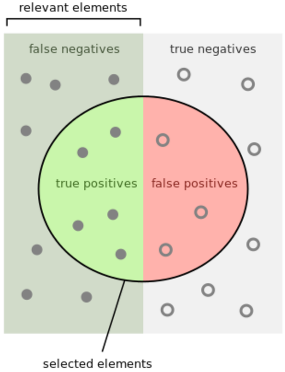
	Actually Positive (1)	Actually Negative (0)	
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)	P'
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)	N'
	P	N	

- **Specificity** How many (-) we hit? (Specificity = True neg rate)
  - $\frac{TN}{N} = \frac{TN}{FP+TN}$
- **False Alarm** How many (-) we miss OR How many (+) we falsely accepted? (False alarm = False pos rate = false acceptance = = type I error rate) How many irrelevant items are selected?
  - $1 - Specificity = \frac{FP}{FP+TN}$



Metrics (2/3)

	Actually Positive (1)	Actually Negative (0)	
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)	P'
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)	N'
	P	N	



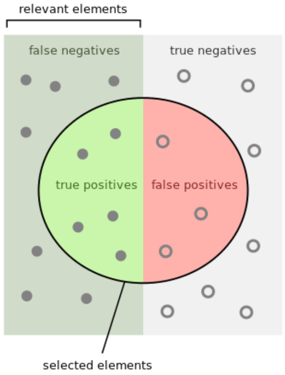
- **Specificity** How many (-) we hit? (Specificity = True neg rate)
  - $\frac{TN}{N} = \frac{TN}{FP+TN}$
- **False Alarm** How many (-) we miss OR How many (+) we falsely accepted? (False alarm = False pos rate = false acceptance = = type I error rate) How many irrelevant items are selected?
  - $1 - Specificity = \frac{FP}{FP+TN}$

Outline

Metrics (3/3)

	Actually Positive (1)	Actually Negative (0)	
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)	P'
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)	N'
	P	N	

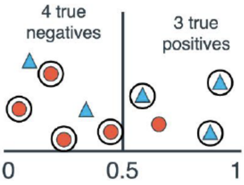
- **Precision** How many of our (+) decisions are correct?
  - $\frac{TP}{P'} = \frac{TP}{TP+FP}$
- **F1 measure** Harmonic mean of precesion and the recall
  - $2\frac{PER*REC}{PER+REC}$



Outline

ROC curves

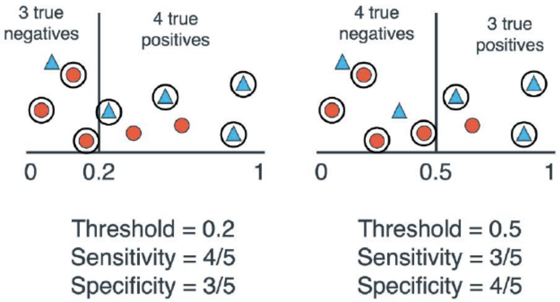
- **Classification threshold** is the point where the model decides to classify a sample as positive or negative.



Threshold = 0.5
Sensitivity = 3/5
Specificity = 4/5

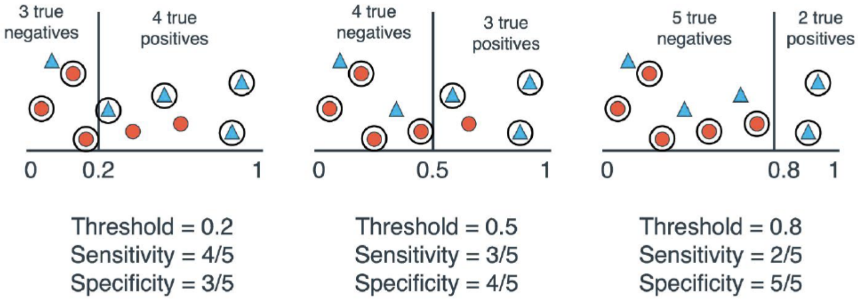
ROC curves

- **Classification threshold** is the point where the model decides to classify a sample as positive or negative.



ROC curves

- **Classification threshold** is the point where the model decides to classify a sample as positive or negative.

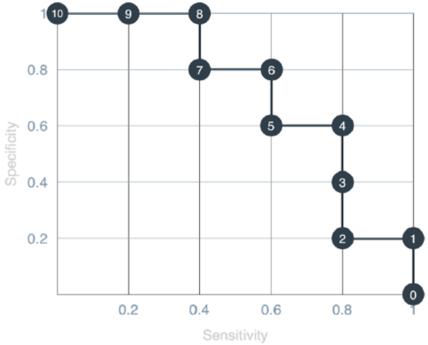


ROC curves characteristics

Timestep	Threshold	True positives	Sensitivity	True negatives	Specificity
0	0	5	1	0	0
1	0.1	5	1	1	0.2
2	0.2	4	0.8	1	0.2
3	0.3	4	0.8	2	0.4
4	0.4	4	0.8	3	0.6
5	0.5	3	0.6	3	0.6
6	0.6	3	0.6	4	0.8
7	0.7	2	0.4	4	0.8
8	0.8	2	0.4	5	1
9	0.9	1	0.2	5	1
10	1	0	0	5	1

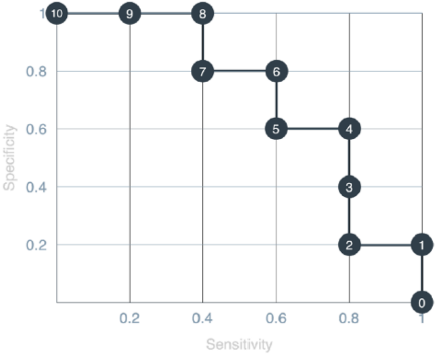
ROC curves characteristics

Timestep	Threshold	True positives	Sensitivity	True negatives	Specificity
0	0	5	1	0	0
1	0.1	5	1	1	0.2
2	0.2	4	0.8	1	0.2
3	0.3	4	0.8	2	0.4
4	0.4	4	0.8	3	0.6
5	0.5	3	0.6	3	0.6
6	0.6	3	0.6	4	0.8
7	0.7	2	0.4	4	0.8
8	0.8	2	0.4	5	1
9	0.9	1	0.2	5	1
10	1	0	0	5	1



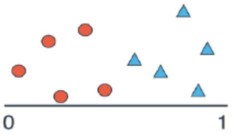
ROC Curves Benefits

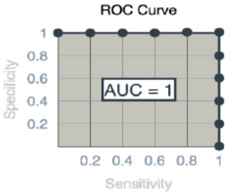
- Plot Sensitivity vs. Specificity as classification threshold (t) varies from 0 to 1
- RoC summarizes all the confusion matrices for all possible thresholds.
- Each point on the RoC is for a different classification threshold.
- (1,1) point is all (+) threshold.
- (0,0) point is all (-) threshold.



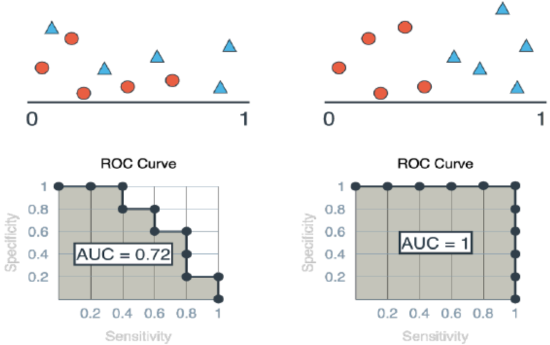
Outline

Area Under the Curve

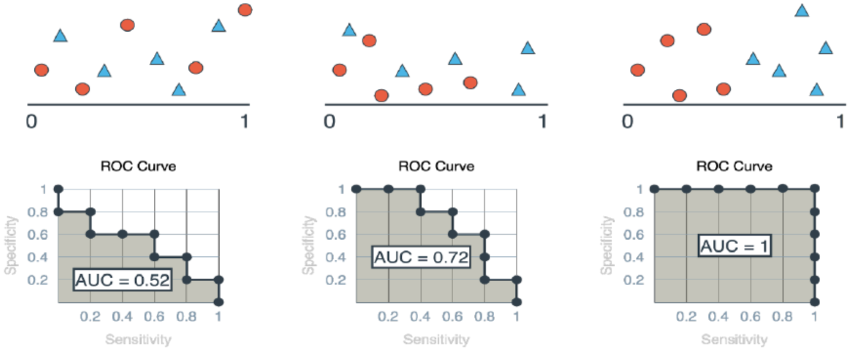




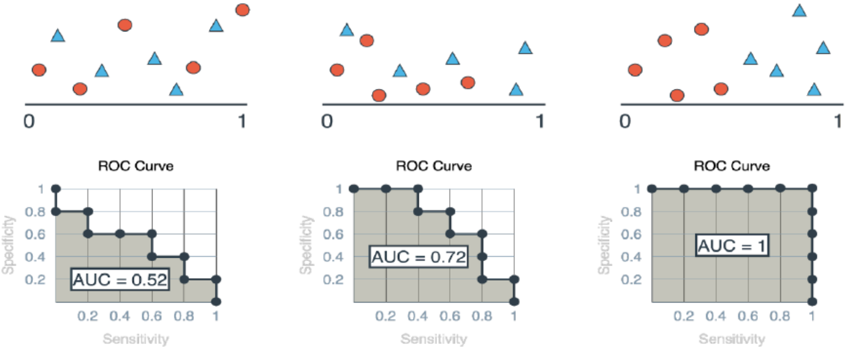
Area Under the Curve



Area Under the Curve



Area Under the Curve

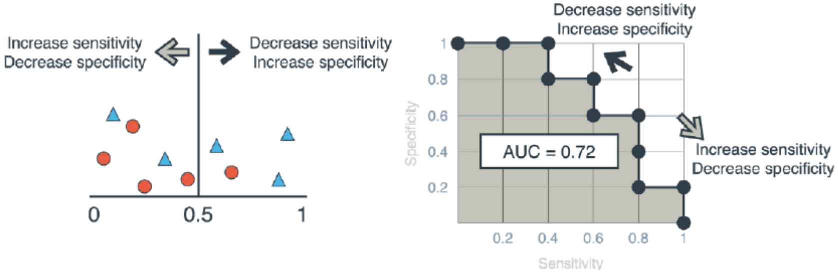


Area Under the Curve tells us how much our model separate the classes.

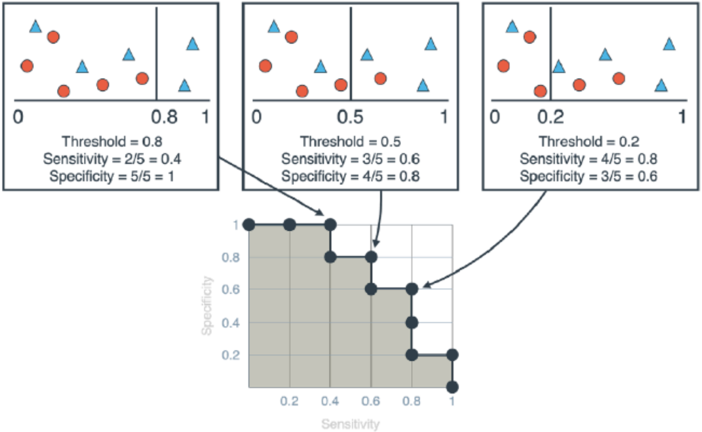
Outline

Decision based on ROC

As we increase or decrease the threshold, we change the sensitivity and specificity of the model, and this change is illustrated by moving in the ROC curve.



Decision based on ROC

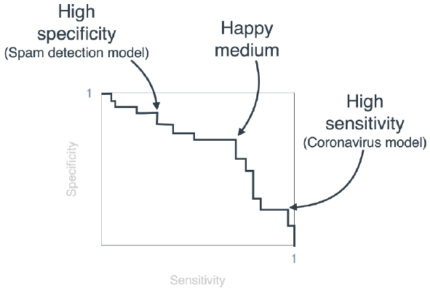


Which point is better for coronavirus detection and which point is better for spam detection?

Decision based on ROC

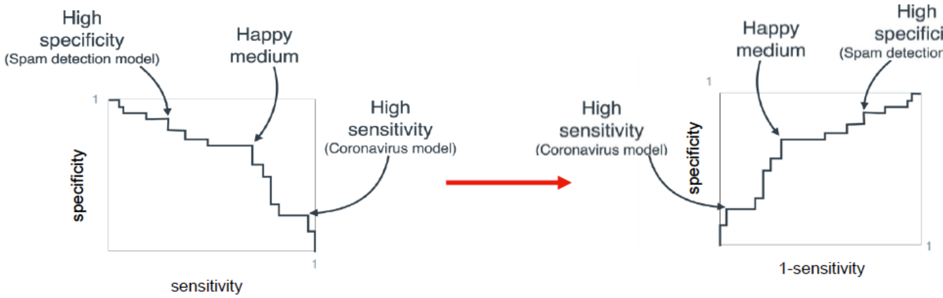
For problems that need **high sensitivity** (like coronavirus model), we use ROC to choose a threshold that achieves that.

For problems that need **high specificity** (like spam detector model), we use ROC to choose a threshold that achieves that.



One minus Specificity

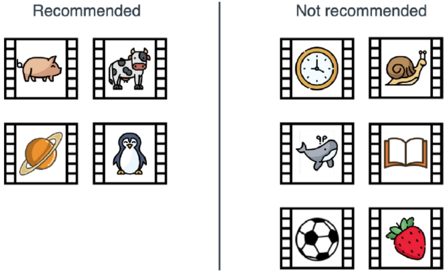
Usually we plot specificity VS 1-sensitivity



Outline

Exercise 1

A video site has established that a particular user likes animal videos and absolutely nothing else. You can see the recommendations that this user got when logging into the site.



a) **What is the accuracy of the model?**
b) **What is the recall of the model?**
c) **What is the precision of the model?**
d) **What is the F1 score of the model?**
e) **Would you say that this is a good recommendation model?**

Exercise 2

Find the sensitivity and specificity of the medical model with the following confusion matrix.

	Predicted Sick	Predicted Healthy
Sick	120	2
Healthy	63	795

Questions