Algorithm Evaluation

Let us design a simple classification algorithm

Purse vs Laptop Bag: Design a classifier








Laptop bag vs. Purse: Design a classifier

- Features:
  - Width
  - Height

- Classifier: threshold

Evaluation Metrics

Let the problem statement be: classifying purse and bags.
Purses are labeled as positive class and bags are labeled as negative class

		Predicted Class	
		Negative	**Positive**
Actual Class	**Negative**	A (true negative)	C (false positive)
	Positive	D (false negative)	B (true positive)

Term	Meaning	Example
True positive	Correct classification	Purse identified as purse
False positive	Incorrect classification	Bag identified as purse
True negative	Correct classification	Bag identified as bag
False negative	Incorrect classification	Purse identified as bag

http://www.cs.rpi.edu/~leen/misc-publications/SomeStatDefs.html

Evaluation Metrics

		Predicted Class	
		Negative	**Positive**
Actual Class	**Negative**	A (true negative)	C (false positive)
	Positive	D (false negative)	B (true positive)

Metric	Formula
Average classification accuracy	TN / (TN + FP) + TP / (TP + FN)
Type I error (false positive rate)	FP / (TN + FP)
Type II error (false negative rate)	FN / (FN + TP)
True positive rate	TP / (TP + FN)
True negative rate	TN / (TN + FP)

http://www.cs.rpi.edu/~leen/misc-publications/SomeStatDefs.html

Evaluation Metrics

Metric	Formula
Average classification accuracy	(TN + TP) / (TN+TP+FN+FP)
Type I error (false positive rate)	FP / (TN + FP)
Type II error (false negative rate)	FN / (FN + TP)
True positive rate	TP / (TP + FN)
True negative rate	TN / (TN + FP)

- Type I error or false positive rate: The chance of incorrectly classifying a (randomly selected) sample as positive
- Type II error or false negative rate: The chance of incorrectly classification a (randomly selected) sample as negative

Evaluation Metrics

Metric	Formula
Average classification accuracy	(TN + TP) / (TN+TP+FN+FP)
Type I error (false positive rate)	FP / (TN + FP)
Type II error (false negative rate)	FN / (FN + TP)
True positive rate	TP / (TP + FN)
True negative rate	TN / (TN + FP)

- Type I error or false positive rate: The chance of incorrectly classifying a (randomly selected) sample as positiv
- Type II error or false negative rate: The cha classification a (randomly selected) samp

Prevalent in computer vision and image processing related classification problems

Evaluation Metrics

Metric	Formula
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)

Precision: Fraction of retrieved instances that are relevant

Recall: Fraction of relevant instances that are retrieved

Evaluation Metrics

Metric	Formula
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)

Precision: Probability that a (randomly selected) retrieved document in relevant

Recall: Probability that a (randomly selected) relevant document is retrieved in a search

Evaluation Metrics

Metric	Formula
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)

Precision: Probability that a (randomly selected) re
document in relevant

Recall: Probability that a (randomly selected) re
document is retrieved in a search

More prevalent in information retrieval domain

Evaluation Metrics

Metric	Formula
Sensitivity	TP / (TP + FN)
Specificity	TN / (TN + FP)
Predictive value for a positive result (PV+)	TP / (TP + FP)
Predictive value for a negative result (PV-)	TN / (TN + FN)

Sensitivity: Proportion of actual positives which are correctly identified

Specificity: Proportion of actual negatives which are correctly identified

Evaluation Metrics

Metric	Formula
Sensitivity	TP / (TP + FN)
Specificity	TN / (TN + FP)
Predictive value for a positive result (PV+)	TP / (TP + FP)
Predictive value for a negative result (PV-)	TN / (TN + FN)

Sensitivity: The chance of correctly identifying positive samples. A sensitive test helps rule out disease (when the result is negative)

Specificity: The chance of correctly classifying negative samples. A very specific test rules in disease with a higher degree of confidence.

Evaluation Metrics

Metric	Formula
Sensitivity	TP / (TP + FN)
Specificity	TN / (TN + FP)
Predictive value for a positive result (PV+)	TP / (TP + FP)
Predictive value for a negative result (PV-)	TN / (TN + FN)

Sensitivity: Proportion of actual positives which are correctly identified

Specificity: Proportion of actual negatives which are correctly identified

What does the sensitivity score of 1.0 mean?

What does the specificity score of 1.0 mean?

F1 Score

- The F1 score is the harmonic mean of precision and recall:

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

- Unlike regular mean, harmonic mean gives more weight to low values.
- Therefore, the classifier's F1 score is only high if both recall and precision are high.

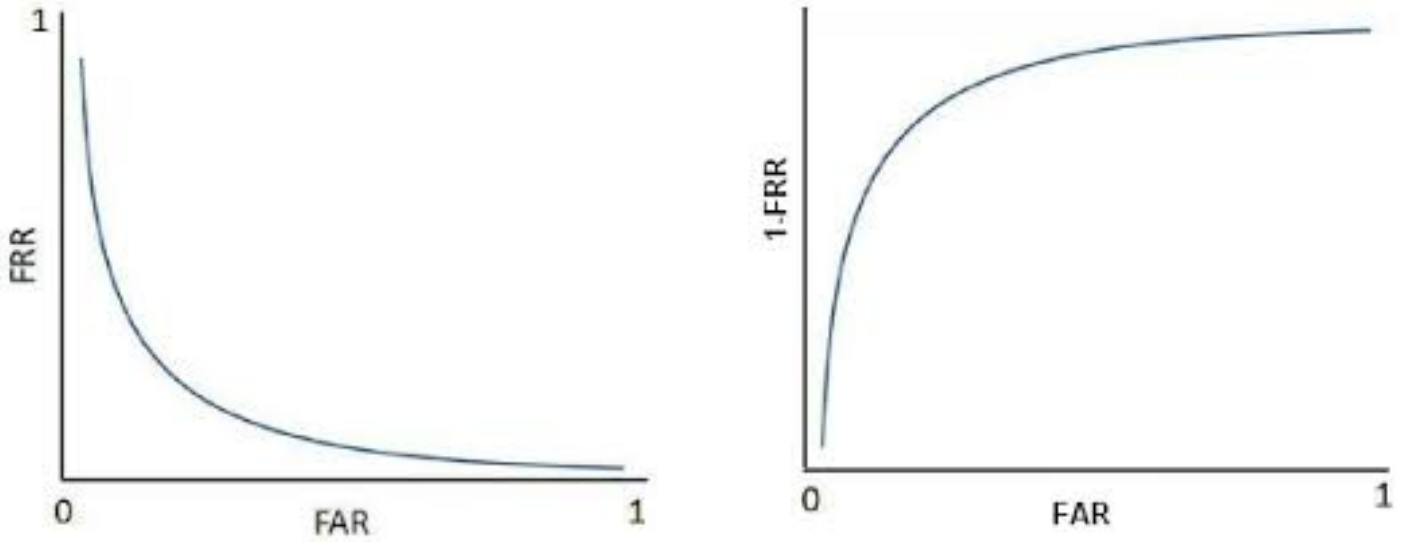
Performance Evaluation

- Classification is of two types:
  - Authentication / verification (1:1 Matching)
    - Is she Richa?
    - Is this an image of a helicopter?
  - Identification (1:n matching)
    - Who's photo is this?
    - This image belongs to which class?

Performance Evaluation

- Receiver operating characteristics (ROC) curve
  - For authentication/verification
  - False positive rate vs true positive rate
- Detection error-tradeoff (DET) curve
  - False positive rate vs false negative rate
- Cumulative match curve (CMC)
  - Rank vs identification accuracy

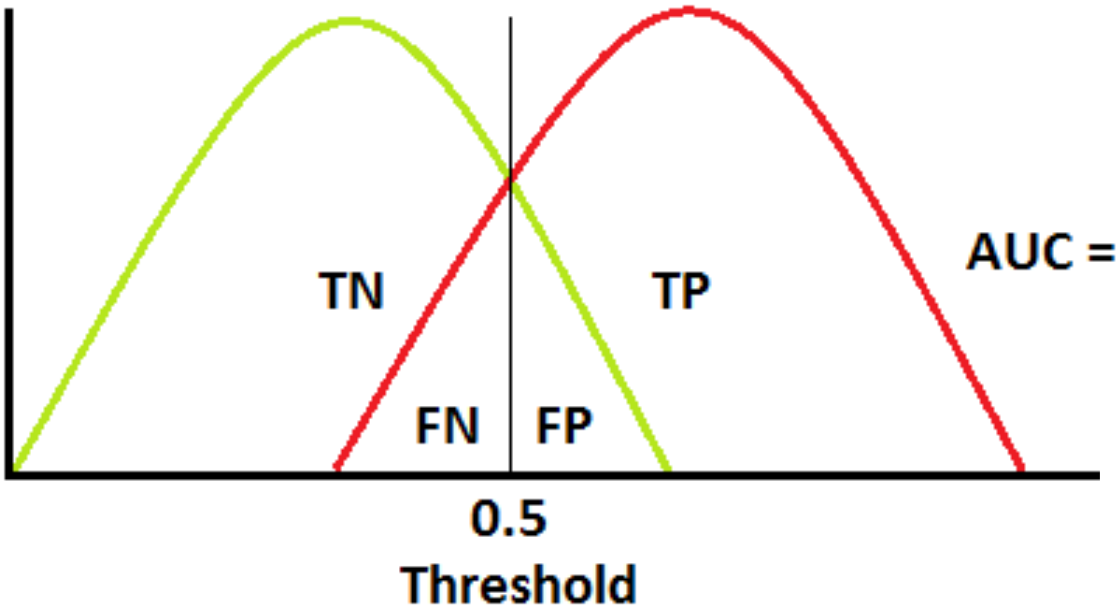
ROC Curve



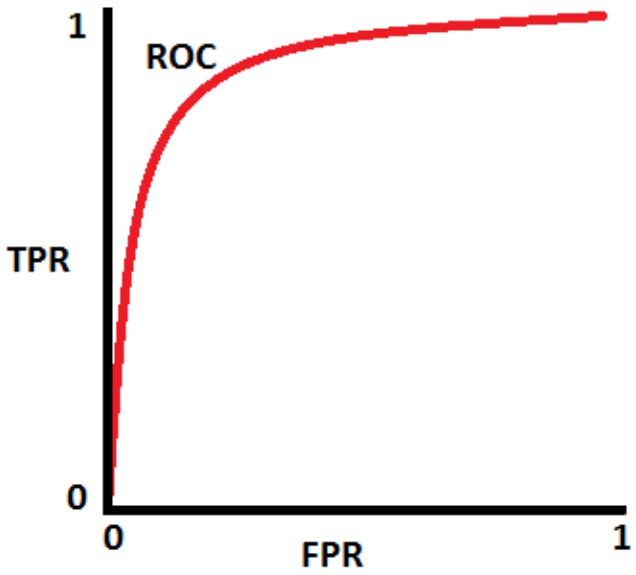
Laptop bag vs. Purse: Design a classifier

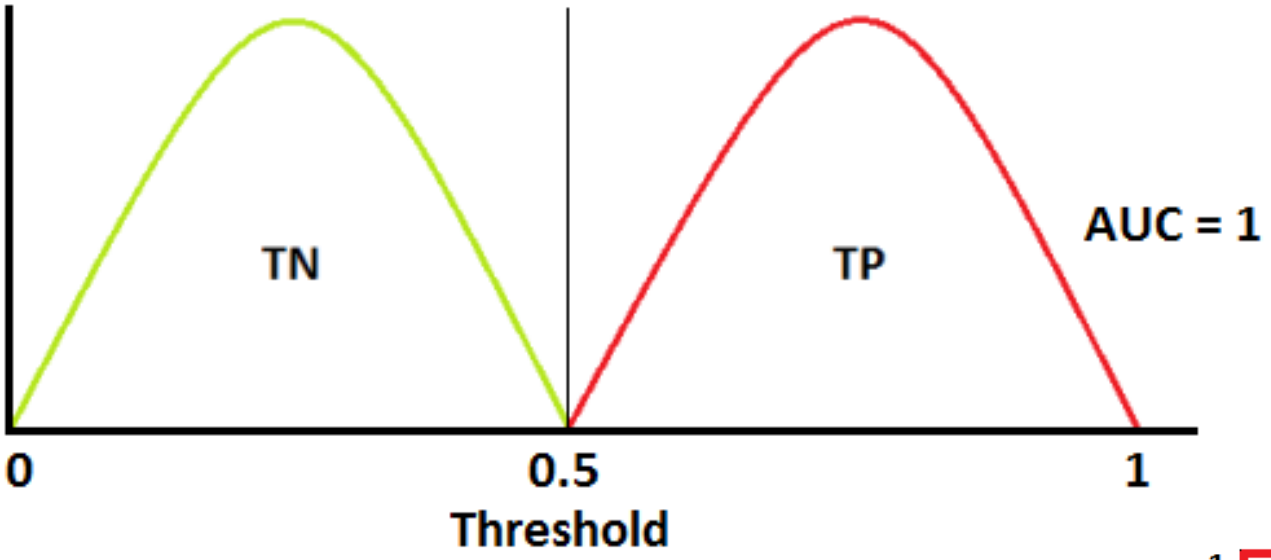
- Features:
  - Width
  - Height
  - Weight

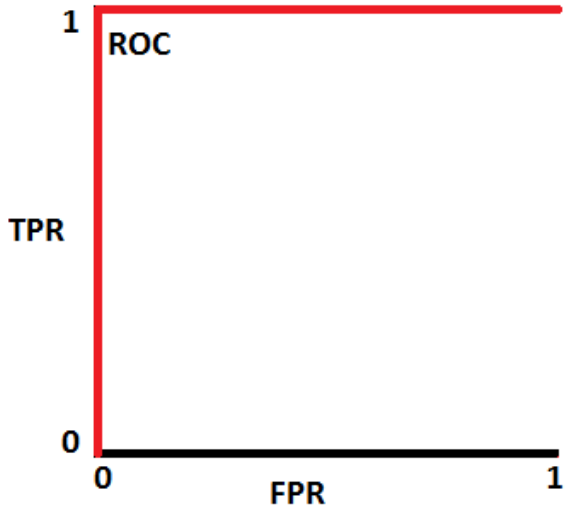
- Classifier: threshold

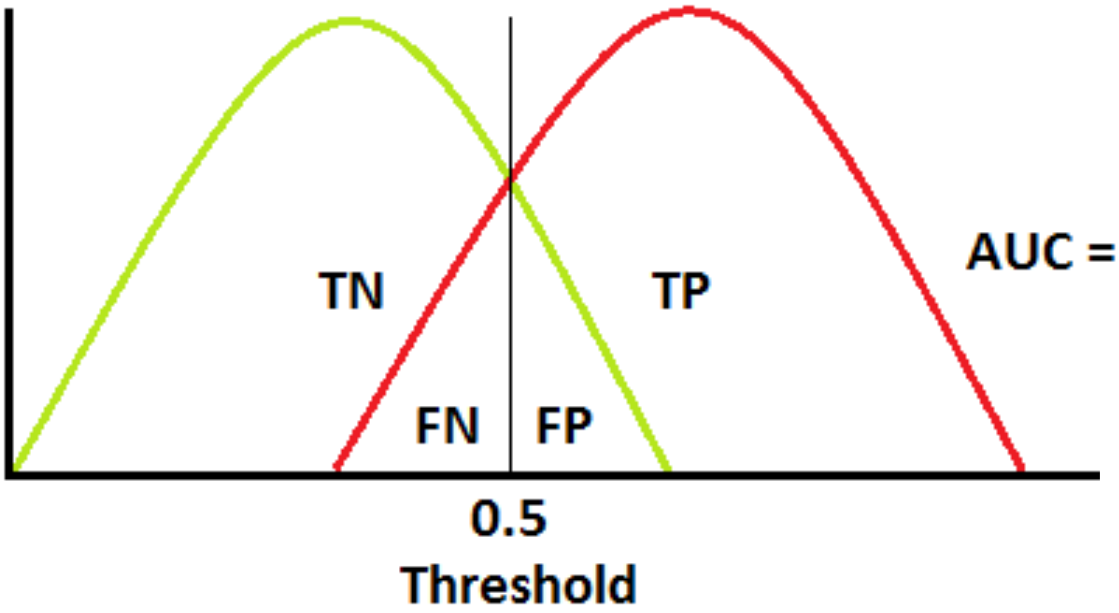


AUC = 0.7

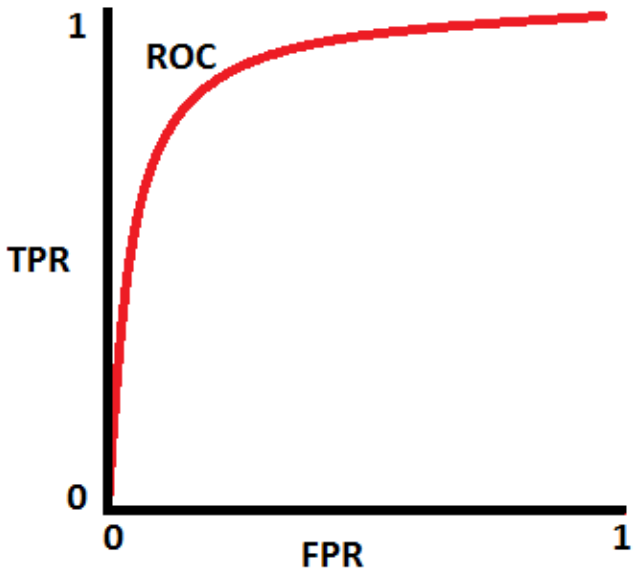


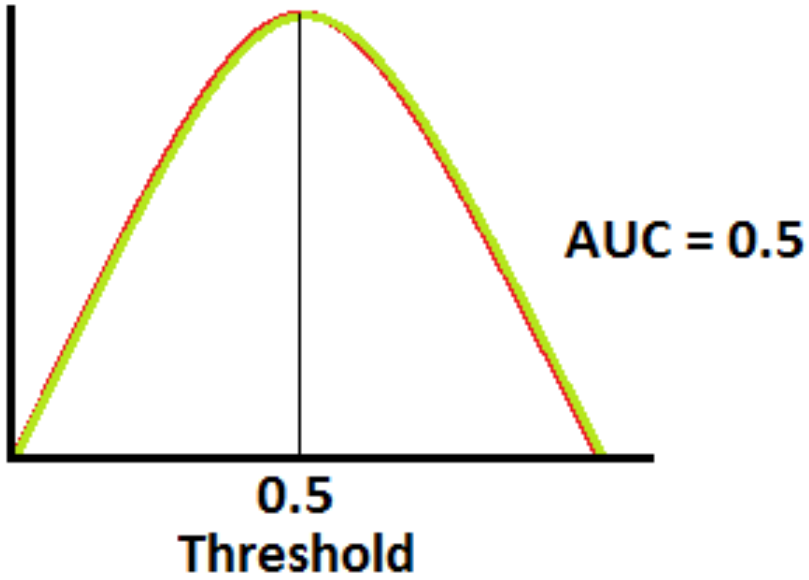

TN
TP
AUC = 1
0
0.5
1
Threshold

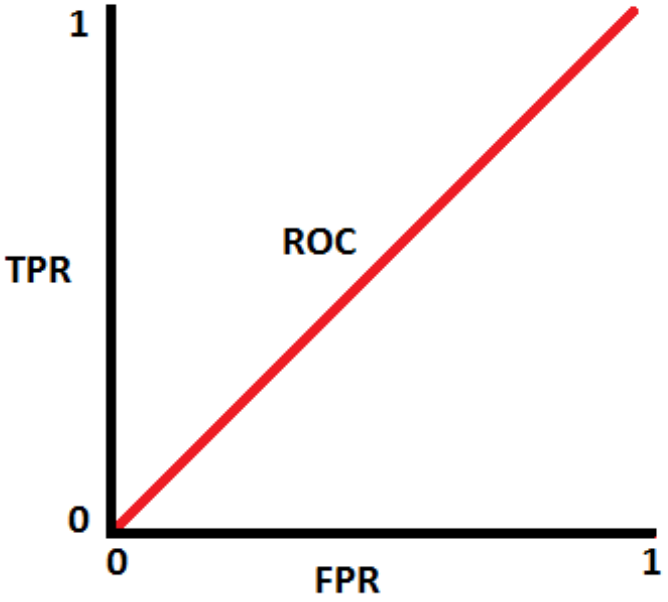

1
ROC
TPR
0
0
1
FPR

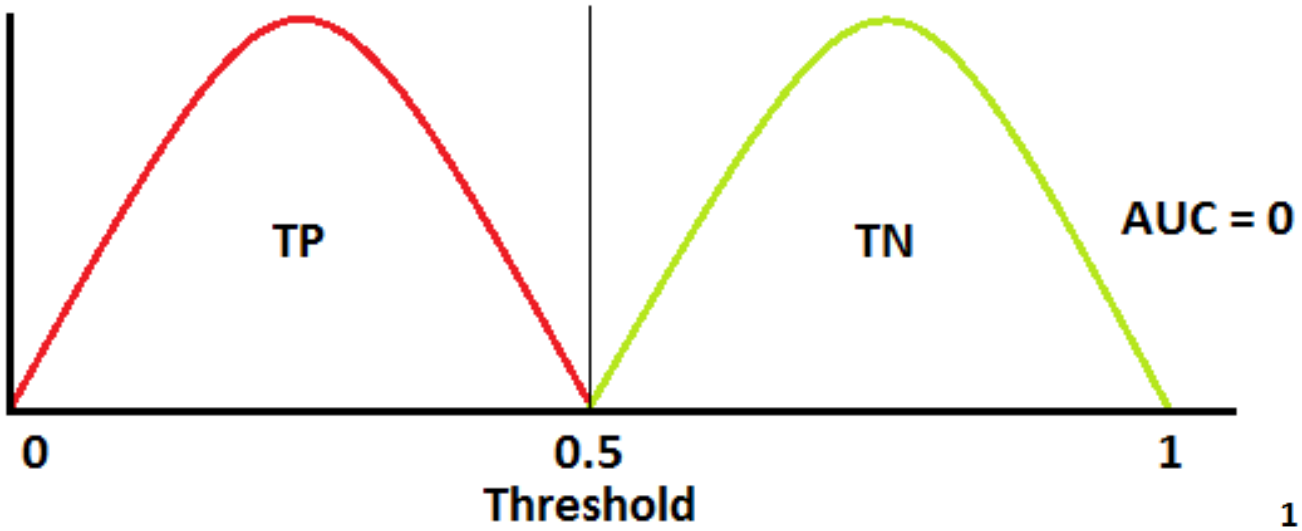


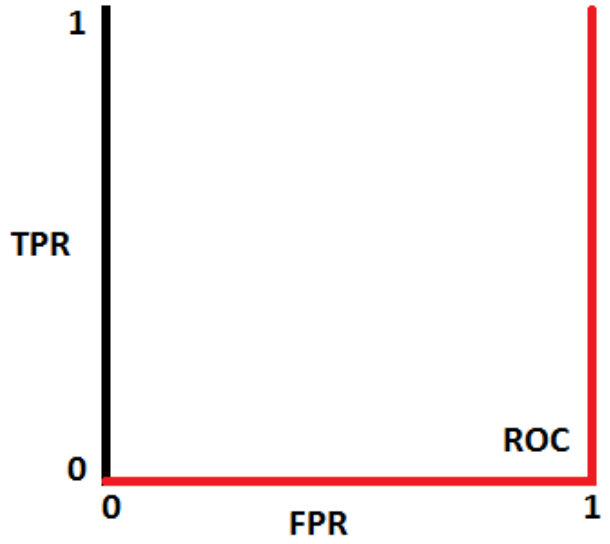
AUC = 0.7

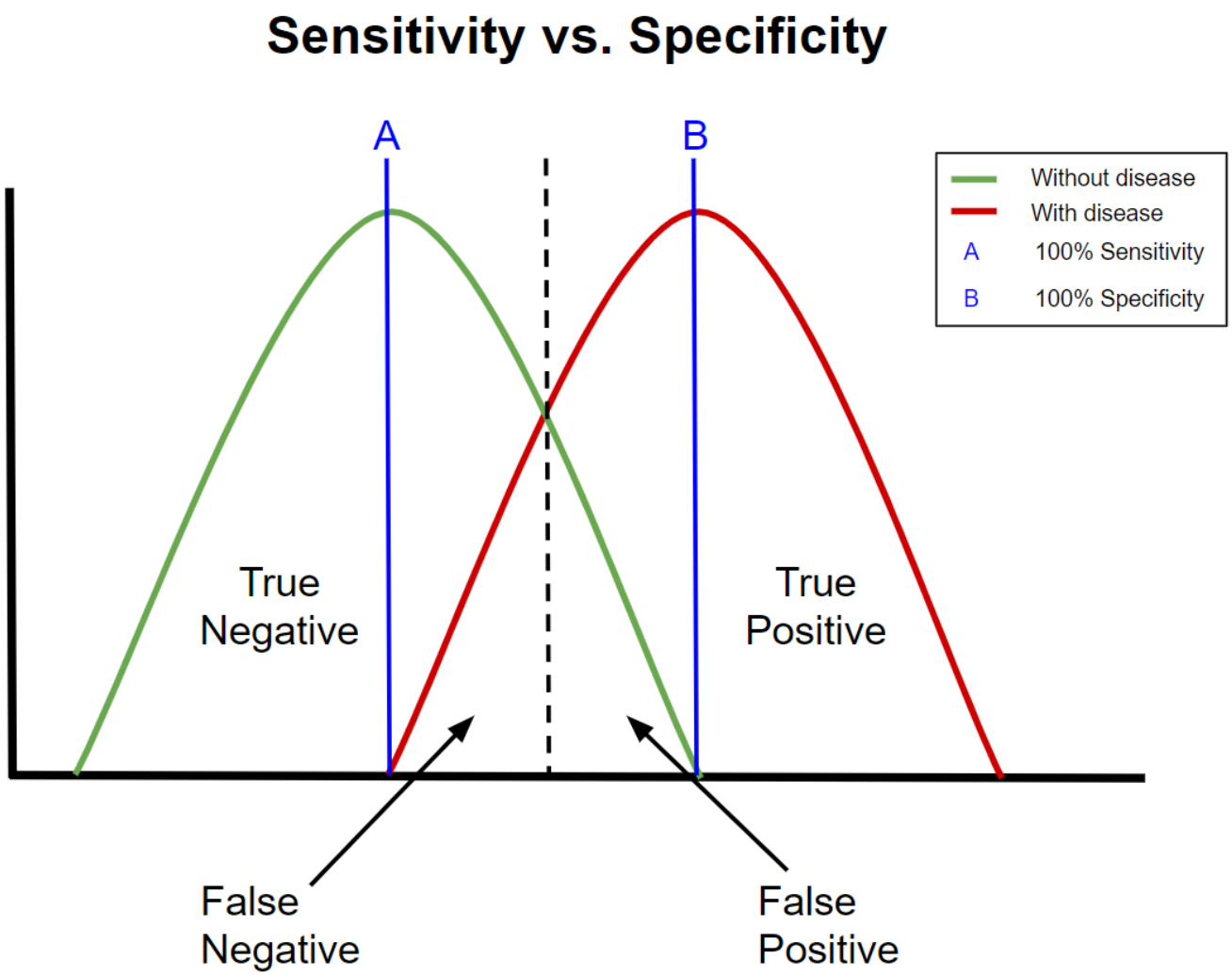



AUC = 0.5
0.5
Threshold

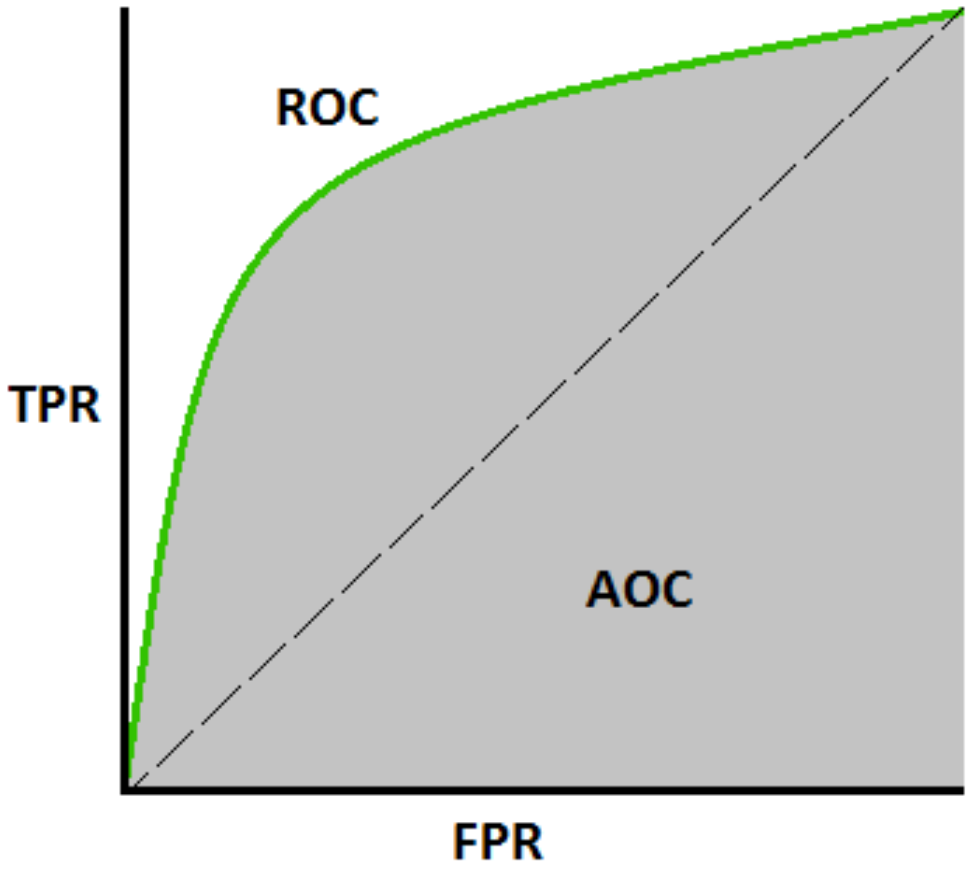

1
TPR
ROC
0
0
FPR
1


TP
TN
AUC = 0
0
0.5
1
Threshold


1
TPR
0
ROC
0
FPR
1

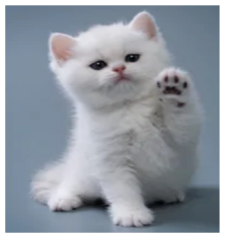

Sensitivity vs. Specificity
A
B
Without disease
With disease
A 100% Sensitivity
B 100% Specificity
True Negative
True Positive
False Negative
False Positive

Area Under the Curve



CMC

Kitten A



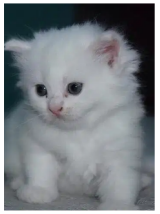
Kitten B

Kitten C

And four test queries:

1

2



3

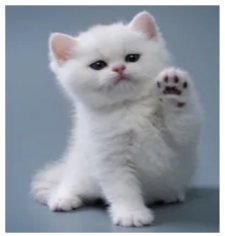
4

Query	Top result	Result 2	Result 3
1	A	B	C
2	B	C	A
3	A	B	C
4	C	B	A

What are the rank accuracies?

Recap: CMC


Kitten A

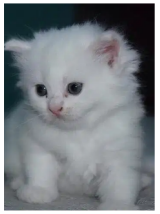

Kitten B


Kitten C

And four test queries:


1


2


3

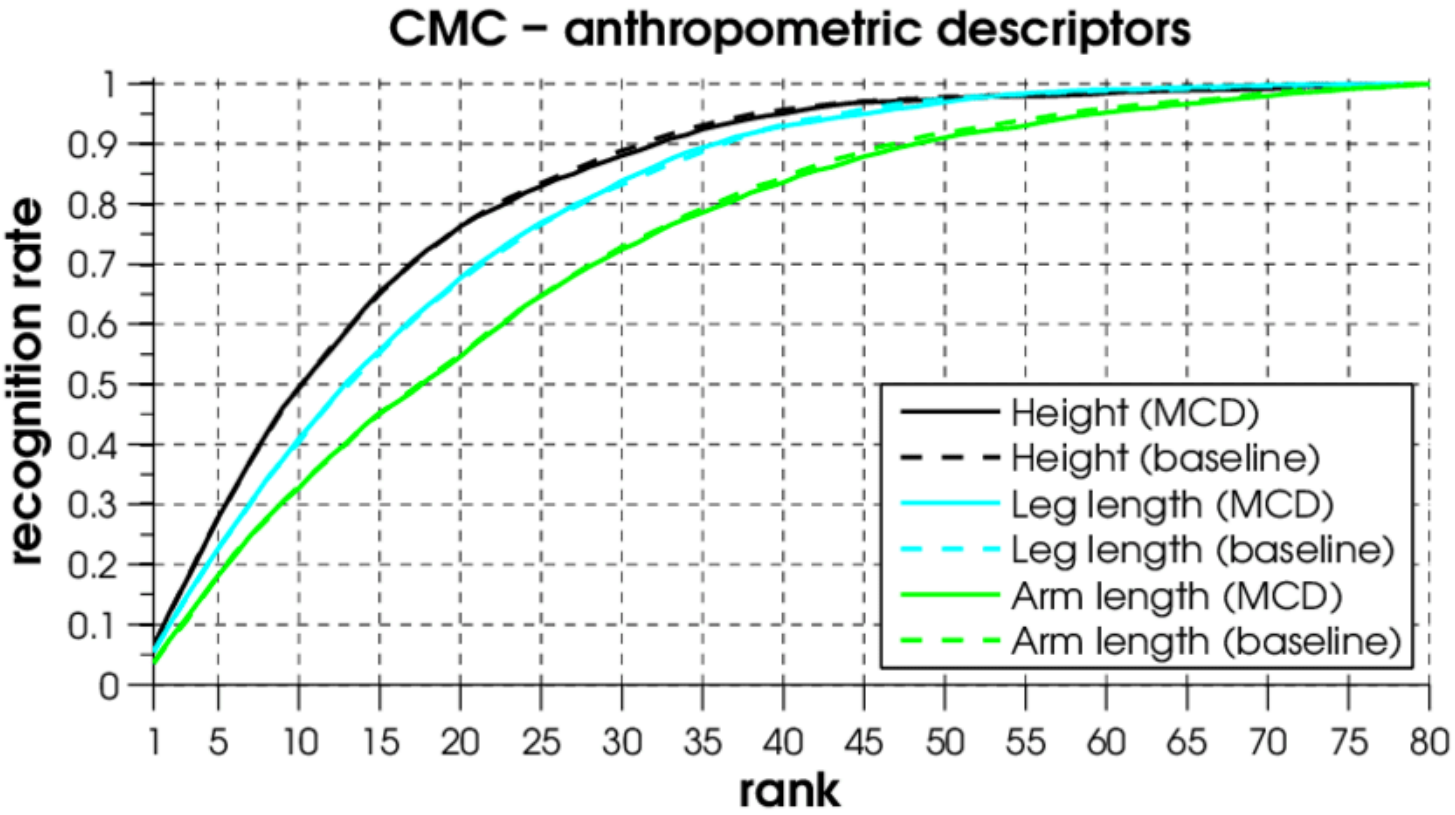

4

Query	Top result	Result 2	Result 3
1	A	B	C
2	B	C	A
3	A	B	C
4	C	B	A

What are the rank accuracies?

Rank 1: 1/4 predicted correctly: 25%
Rank 2: 3/4 : 75%
Rank 3: 4/4 : 100%

CMC Curve

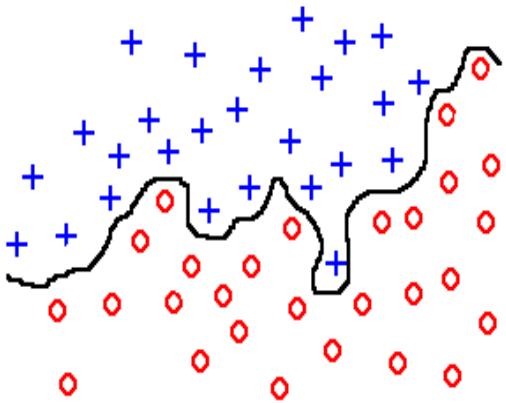

CMC – anthropometric descriptors
recognition rate
rank
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
Height (MCD)
Height (baseline)
Leg length (MCD)
Leg length (baseline)
Arm length (MCD)
Arm length (baseline)

Evaluating ML Systems

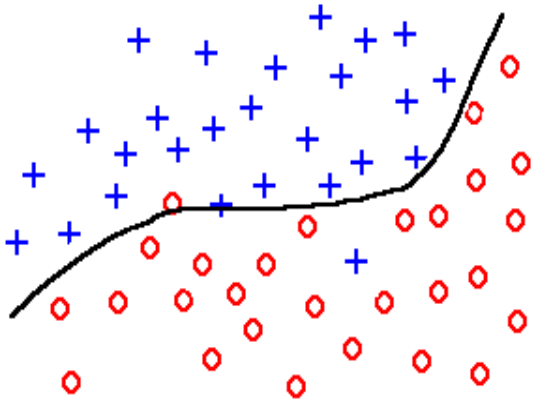
- Assumption: Building a model for population
- Reality: Population is not available
- We work with a sample database – not necessarily true representation of the population

- What to do?
  - Should we use the entire available database for training the model?
    - High accuracy on the training data
    - Lower accuracy on the testing data
    - Called as overfitting

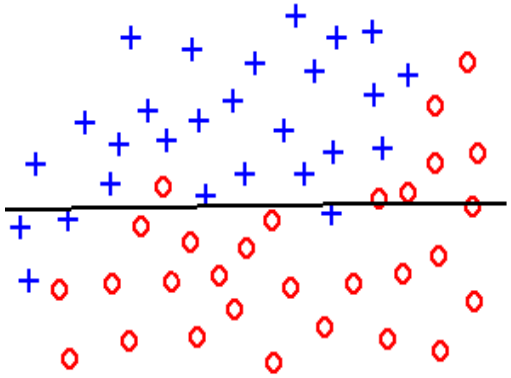
Evaluating ML Systems



Overfitting



Good fit

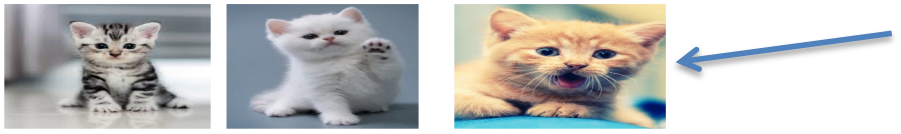


Underfitting

- Underfitting: Learning algorithm had the opportunity to learn more from training data, but didn't
- Overfitting: Learning algorithm paid too much attention to idiosyncracies of the training data; the resulting tree doesn't generalize

Cross Validation

- "Cross-Validation is a statistical method of evaluating and comparing learning algorithms."
- The data is divided into two parts:
  - Training: to learn or train a model
  - Testing: to validate the model



Kitten A Kitten B Kitten C

Training database

1 2 3 4

Testing database

Cross Validation

- It is used for
  - Performance evaluation: Evaluate the performance of a classifier using the given data
  - Model Selection: Compare the performance of two or more algorithms (DT classifier and neural network) to determine the best algorithm for the given data
  - Tuning model parameters: Compare the performance of two variants of a parametric model

Type of Cross Validation

- Resubstitution Validation
- Hold-Out Validation
- K-Fold Cross-Validation
- Leave-One-Out Cross-Validation
- Repeated K-Fold Cross-Validation

Type of Cross Validation…

- Resubstitution Validation
  - All the available data is used for training and the same data is used for testing
    - Does not provide any information about generalizability

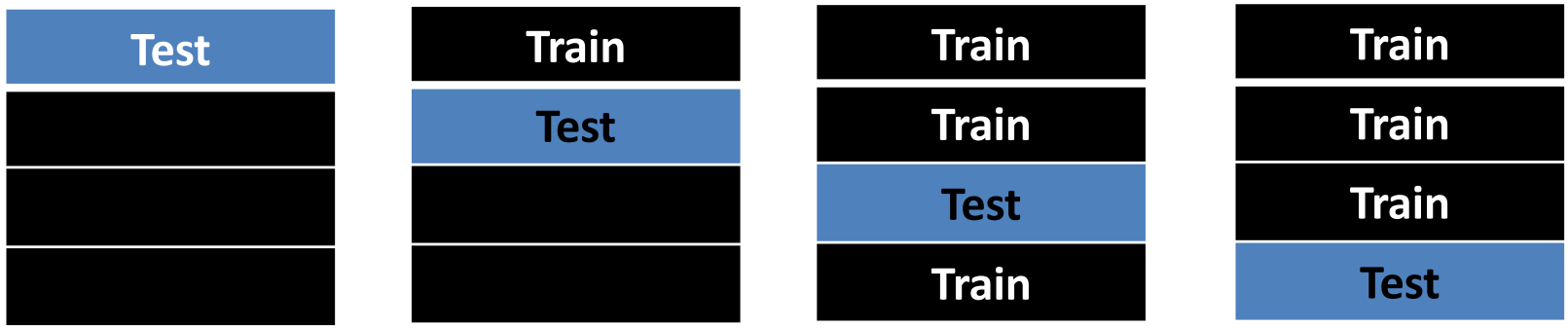
Type of Cross Validation…

- Hold-Out Validation
  - The database is partitioned into two non-overlapping parts, one for training and other for testing
    - The results depend a lot on the partition, may be skewed if the test set is too easy or too difficult

Type of Cross Validation…

- K-Fold Cross-Validation
  - Data is partitioned into k equal folds (partitions). k-1 folds are used for training and 1-fold for testing
  - The procedure is repeated k times

- Across multiple folds, report:
  - Average error or accuracy
  - Standard deviation or variance



4-fold cross validation

Type of Cross Validation…

- Repeated K-Fold Cross-Validation
  - Repeat k-fold cross validation multiple times
- Leave-One-Out Cross-Validation
  - Special case of k-fold cross validation where k=number of instances in the data
  - Testing is performed on a single instance and the remaining are used for training

- Across multiple folds, report:
  - Average error or accuracy
  - Standard deviation or variance

Comparing Cross-Validation Methods

Validation Method	Advantages	Disadvantages
Resubstitution	Simple	Overfitting
Hold-out validation	Independent training and testing sets	Reduced data for training and testing
K-fold cross validation	Accurate performance estimation	Small sample for performance estimation, underestimated performance variance or overestimated degree of freedom for comparison
Leave-one-out cross validation	Unbiased performance estimation	Very large variance
Repeated k-fold cross-validation	Large number of performance estimates	Overlapped training and test data between each round, underestimated performance variance or overestimated degree of freedom for comparison