Pattern Recognition and Machine Learning

Richa Singh and Pratik Mazumder

Moodle

Slides are prepared from several information sources on the web and books

About the instructors

- Richa Singh:
  - Professor and HoD - CSE, IIT Jodhpur
  - Faculty, IIIT Delhi (2009 – 2019)
  - home.iitj.ac.in/~richa
- Dr. Pratik Mazumder
  - Assistant Professor, CSE
  - https://sites.google.com/view/pratikmazumder/

Teaching Assistants

Neelu Verma
Sandeep Khanna
Shilajit Banerjee
Dattatreyo Roy
Harshil Bhatia
Manav Kapoor
Jayesh Khandelwal
Ram Khandelwal
Susim Mukul Roy
Atharva Pandey
Stuti Aswani

Course Objectives

- To understand various key paradigms for pattern classification and machine learning approaches
- To familiarize with the mathematical and statistical techniques used in pattern recognition and machine learning.
- To understand and differentiate among various pattern recognition and machine learning techniques.

Learning Outcomes

- The students are expected to have the ability to:
  - To formulate a machine learning problem
  - Select an appropriate pattern analysis tool for analyzing data in a given feature space.
  - Apply pattern recognition and machine learning techniques such as classification and feature selection to practical applications and detect patterns in the data.

Course Content

- **Introduction**: Definitions, Datasets for Pattern Recognition, Different Paradigms of Pattern Recognition and Machine Learning, Data Normalization, Hypothesis Evaluation, VC-Dimensions and Distribution, Bias Variance Tradeoff, Regression (Linear) (8 Lectures)
- **Discriminative Methods**: Distance-based methods, Linear Discriminant Functions, Decision Tree, Random Decision Forest and Boosting (5 Lectures)
- **Bayes Decision Theory**: Bayes decision rule, Minimum error rate classification, Normal density and discriminant functions, Bayesian networks (7 Lectures)
- **Parameter Estimation**: Maximum Likelihood and Bayesian Parameter Estimation (3 Lectures)
- **Feature Selection and Dimensionality Reduction**: PCA, LDA, ICA, SFFS, SBFS (4 Lectures) Artificial Neural Networks: MLP, Backprop, and RBF-Net (4 Lectures)
- **Kernel Machines**: Kernel Tricks, Support Vector Machines (primal and dual forms), K-SVR, K-PCA (6 Lectures)
- **Clustering**: k-means clustering, Gaussian Mixture Modeling, EM-algorithm (5 Lectures)

Reading Resources

- Textbooks:
  - Pattern Classification, 2nd Edition, Richard O. Duda, Peter E. Hart, David G. Stork, Wiley

- Reference Books:
  - Tom Mitchell, Machine Learning
  - C. Bishop, Pattern Recognition and Machine Learning, Springer
  - K. Murphy, Machine Learning: a Probabilistic Perspective, MIT Press
  - Shalev-Shwartz,S., Ben-David,S., (2014), Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press

Evaluation Components

- Grading
  - Assignments (labs, programming and written): 30%
  - Exams: 35%
  - Project: 15%
  - Quiz : 10%
  - Kaggle Challenges: 10%
- Project team size: 2 students
  - Predefined project topics: you have to select one
- Labs and Assignments: individually

Collaboration Policy

- Discussion with friends and colleagues is good… but
  - the objective should be to improve understanding and learning
  - Not getting answers
- If you have discussed with anyone, you should acknowledge who helped you – from the class or outside the class

Plagiarism Policy

- Cheating in assignments/quizzes/projects/
  - First offence: Zero in the evaluation component
  - Second offense: Grade reduction/F grade
- Cheating in exam: F grade
- Misbehavior: Institute guidelines

Class Organization

- Lectures: 3 days
- Labs: 2 groups
  - Time will be decided this week

Any questions regarding administrative guidelines?

What is Pattern Recognition and Machine Learning?

Machine Learning

- What do we understand by learning?
  – Learning is any process by which a system improves performance from experience.”

  – Herbert Simon (1950)

- Machine Learning is the study of algorithms that
  – improve their performance P
  – at some task T
  – with experience E.

  A well-defined learning task is given by <P, T, E>.

  - Tom Mitchell (1998)

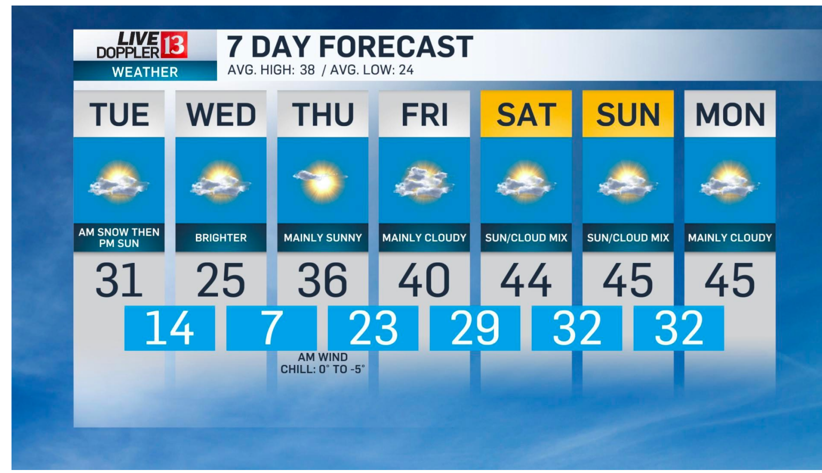
How to design a learning system

- Understand the problem statement
- Choose exactly what is to be learned
- Choose how to represent the target function
- Choose a learning algorithm to infer the target function from the experience

Based on slide by Ray Mooney

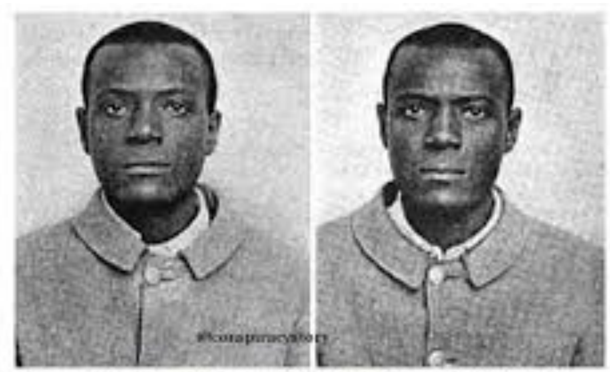
Applications of ML

What are the facial expressions?



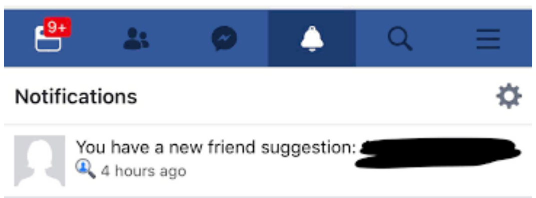
Weather forecast

What are these letters?



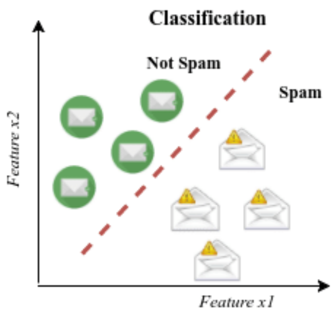
Are they same or different?

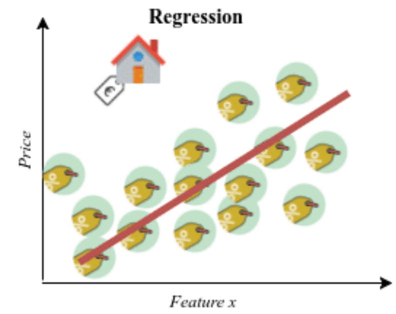
Applications of ML

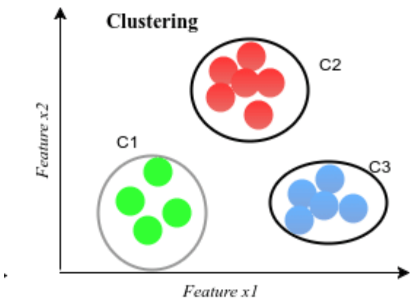


Task, T

- Classification
- Regression
- Ranking
- Recommendation
- Clustering
- Density estimation
- ....







Performance, P

- Metric used to evaluate the performance of T
  - Classification: error rate or accuracy
  - Regression: mean squared error
  - Density estimation: probability assigned to samples

ML in Practice

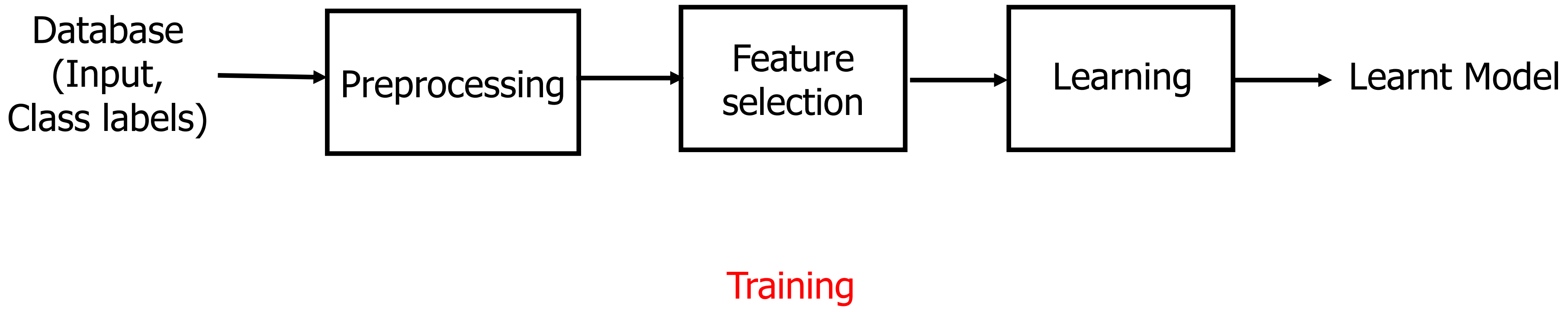
- Understand domain, prior knowledge, and goals
- Data integration, selection, cleaning, pre-processing, etc.
- Learn models
- Interpret results
- Consolidate and deploy discovered knowledge

Based on a slide by Pedro Domingos

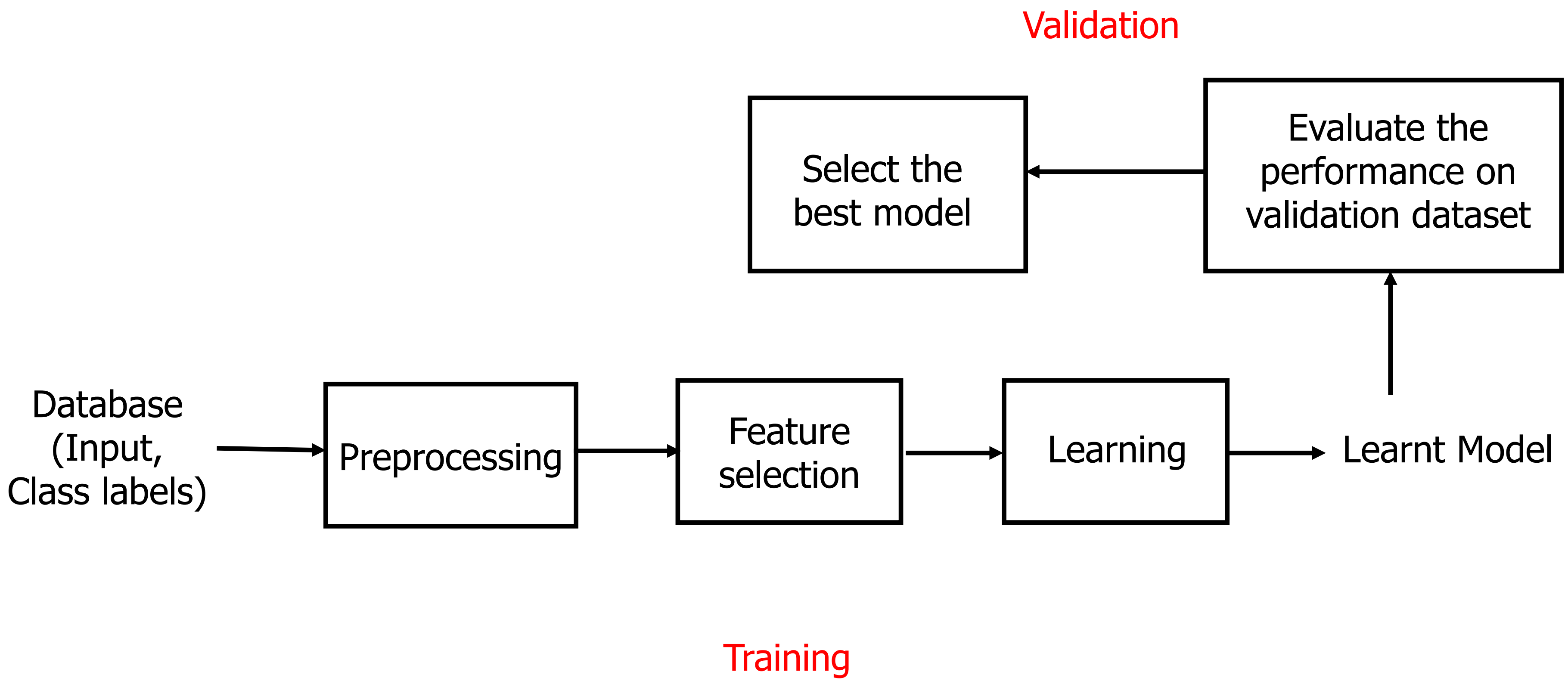
Machine Learning Pipeline

- Three steps:
  - Training
  - Validation
  - Testing

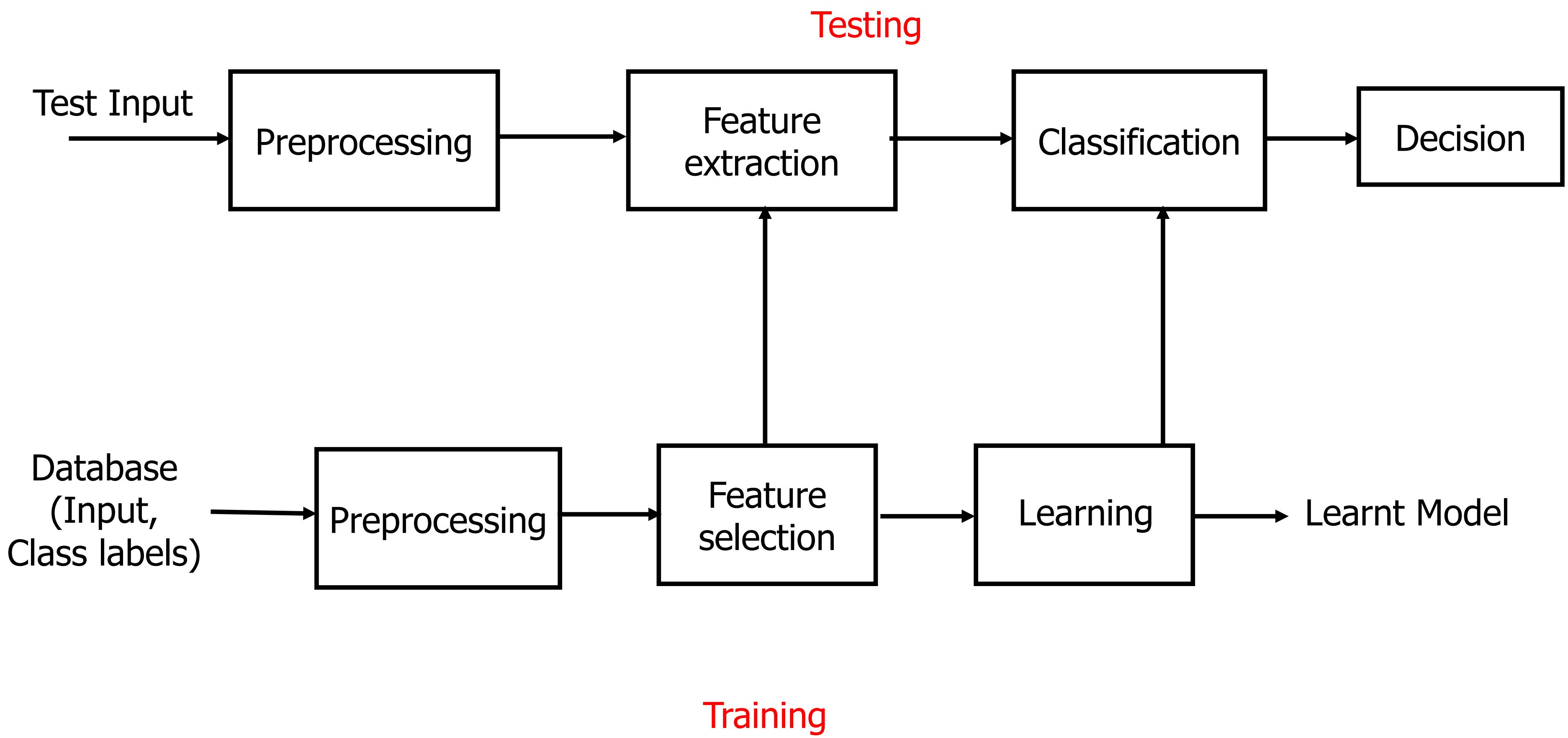
Machine Learning Pipeline



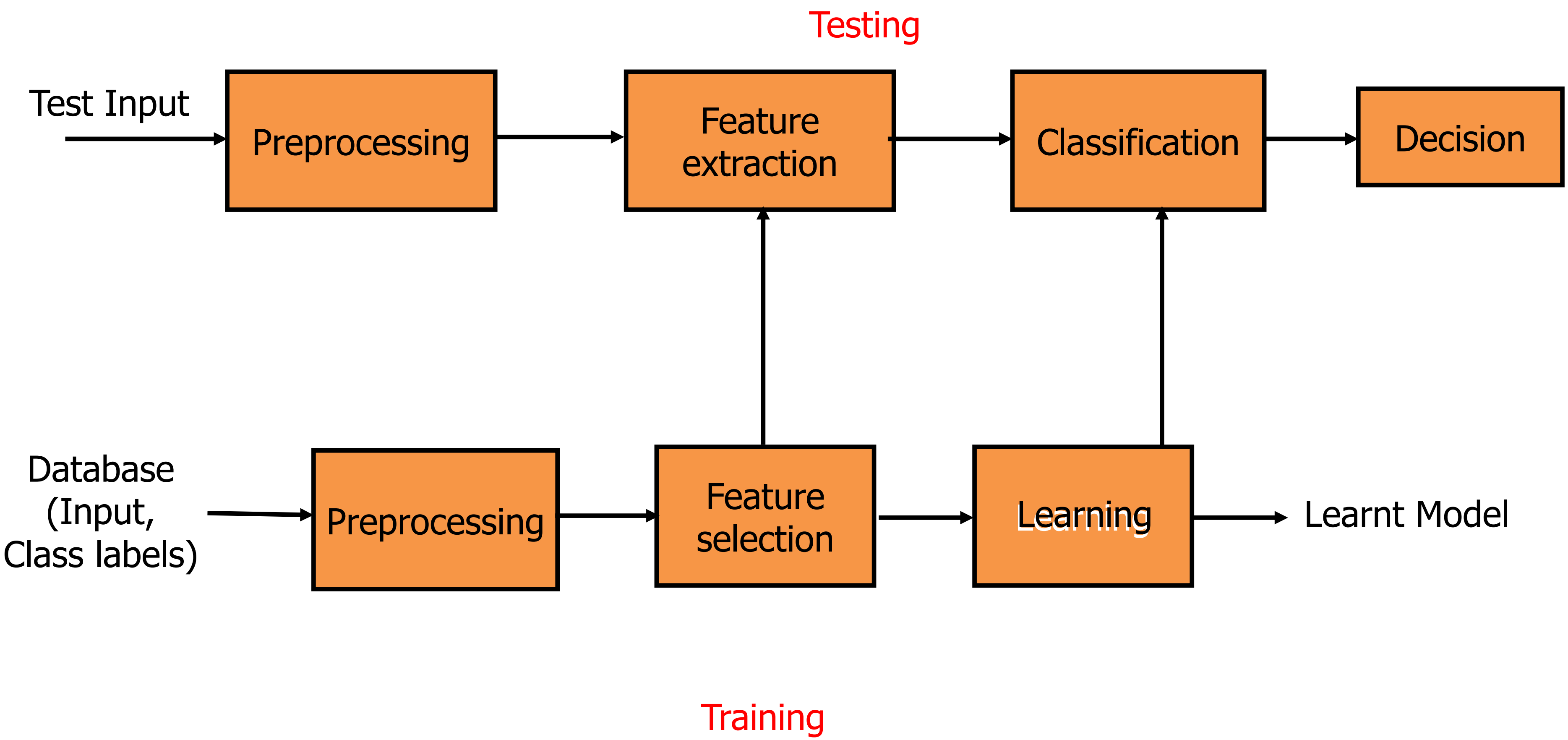
Machine Learning Pipeline



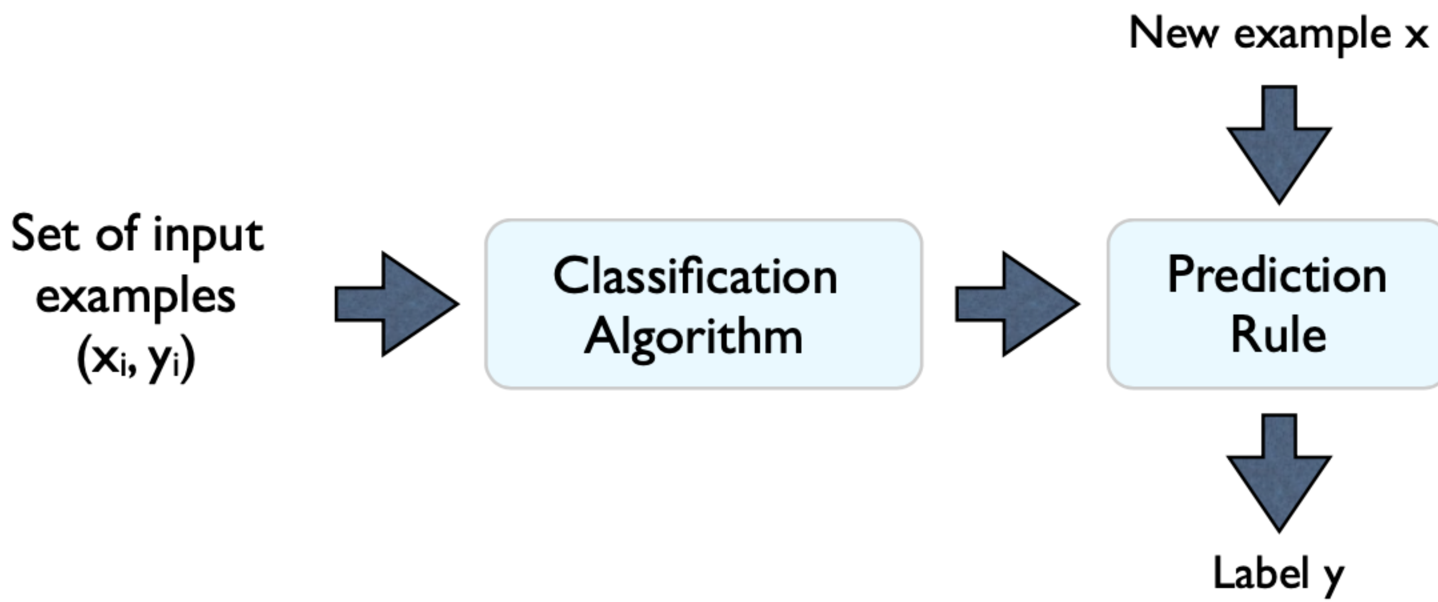
Machine Learning Pipeline



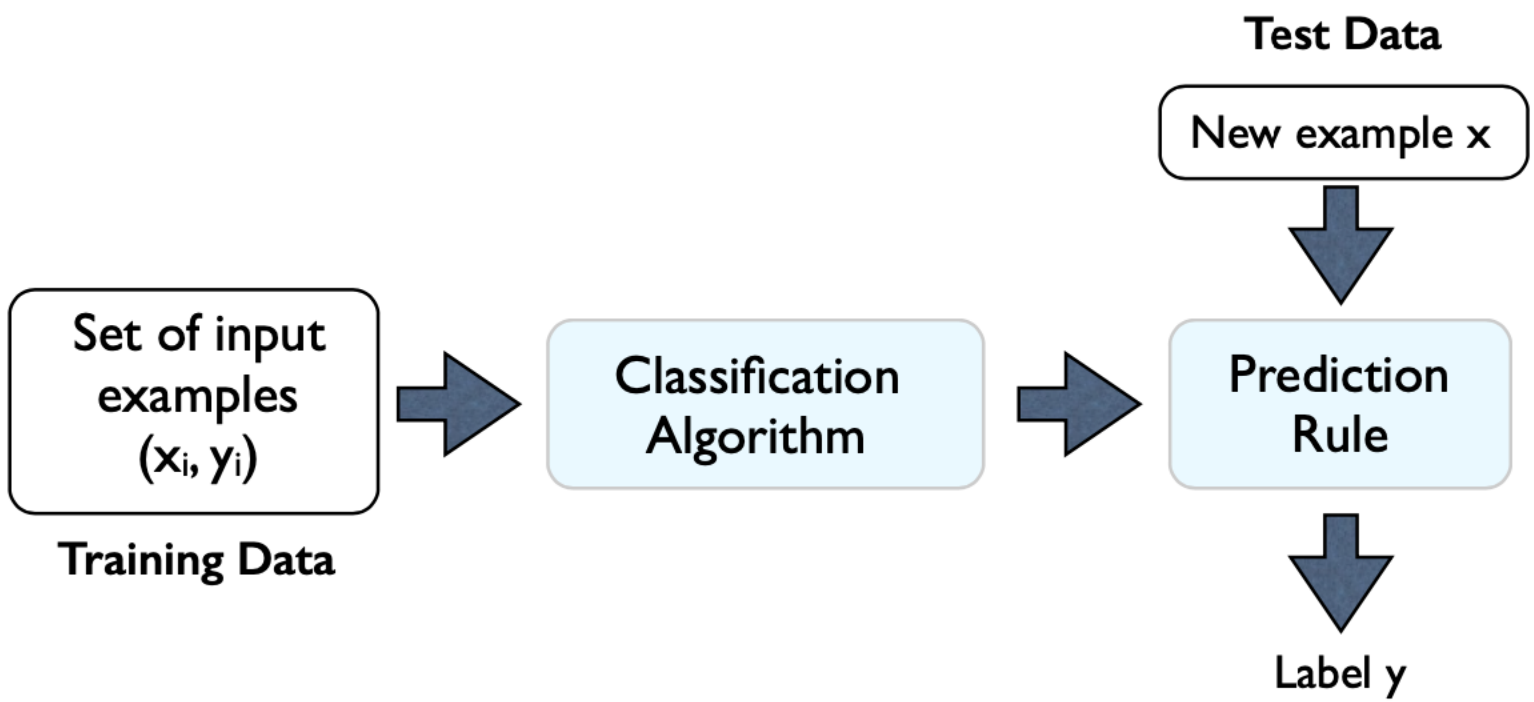
Machine Learning Pipeline



Typical Classification Algorithm



Typical Classification Algorithm



Training and test data must be **separate**!

Performance Measure:
Accuracy (or fraction of correct answers) on **test data**

ML Paradigms and Evaluation Metrics

Types of ML Paradigms

- Supervised learning
  - Classification
  - Regression
- Unsupervised learning
- Reinforcement learning

Supervised Learning

- Given: training data + desired outputs (labels)
- (x1, y1), (x2, y2), ..., (xn, yn)
- Learn a function f(x) to predict y given x

Cats

Dogs

Supervised Learning

Example 1: Predict if a **new** patient has flu or not, based on **existing** patient data

What is x and y?

Fever	Cold	Temperature	Flu?
Yes	No	99F	Yes
	↓		
1	0	99	+
	x		y

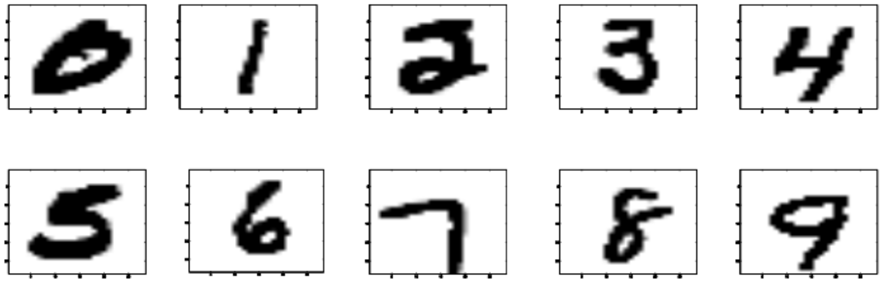
Features: Properties of patient

Label: Flu/No flu

A **binary** (two-label) classification problem

Supervised Learning

Example 2: Which digit in the image ?

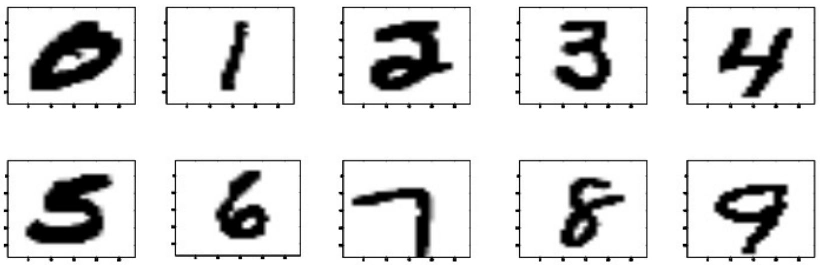


Label: 0,1,..,9

What are the features?

A **multiclass** classification problem

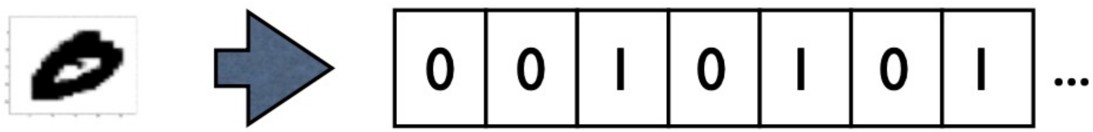
Supervised Learning



Label: 0,1,...,9

What are the features?

Option: vector of pixel colors



Image

x (0 for white, 1 for black)

There are other options too

Lesson: Choosing features is non-trivial in real applications

Supervised Learning

Example 3: Spam or not?

Email 1
From: Canadian Pharmacy
Subject: Offer ends now!

Email 2
From: Yuncong Chen
Subject: TA meeting

	Pharmacy	offer	meeting	TA	Spam?
Email 1	1	1	0	0	Yes
Email 2	0	0	1	1	No

Label: 0 (not spam), 1 (spam)

Features: Words in the email

Supervised Learning

Regression:

Given data:

$(x_i, \quad y_i) \qquad i=1,..,n$

independent variable — dependent variable

where y is **continuous**, design a rule to predict y values for unseen x

Supervised Learning

Regression: Given data (x_i, y_i)
where y is **continuous**, predict y values for unseen x

Example 1: Predict house price from properties of house

Bedrooms	Bathrooms	Area	Price
3	2	2000	600K
2	1	1200	400K
	x		y

Independent Variable: Property of house
Dependent variable: price

Supervised Learning

Child

Adult

Classification

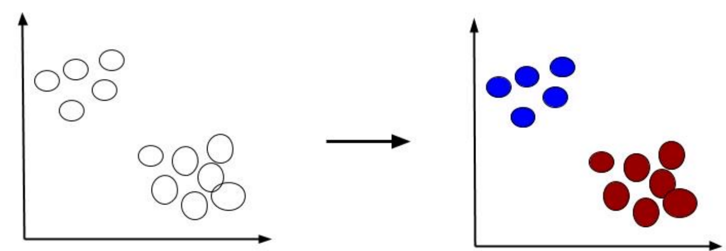
6 years

?? years

Regression

Unsupervised Learning

- Given: training data (without desired outputs)
- x1, x2, ..., xn (without labels)
- Output hidden structure behind the x's – Group them by similarity



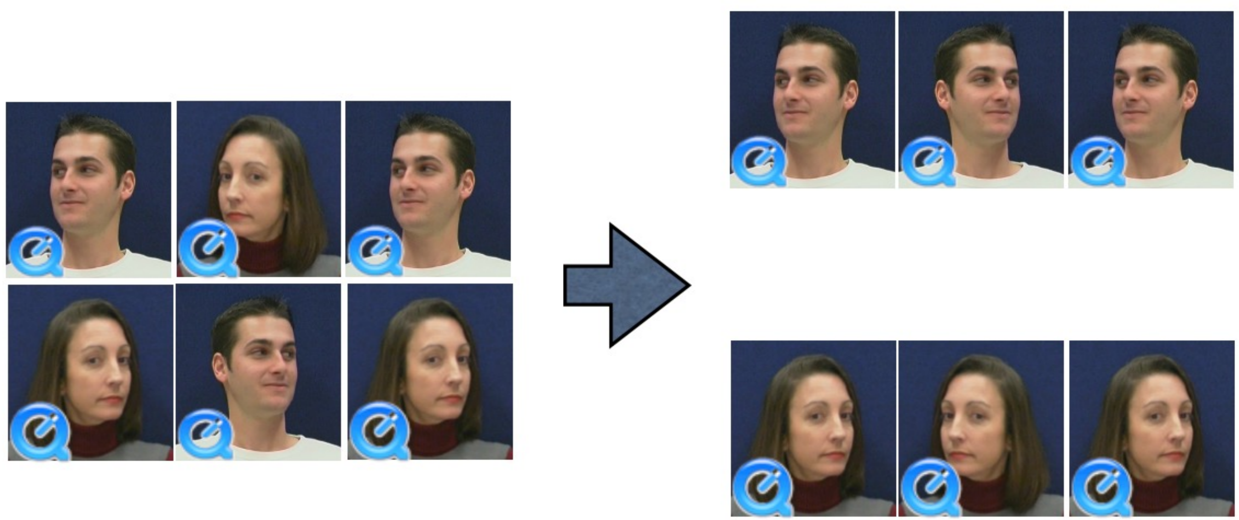
Clustering

Unsupervised Learning

Clustering

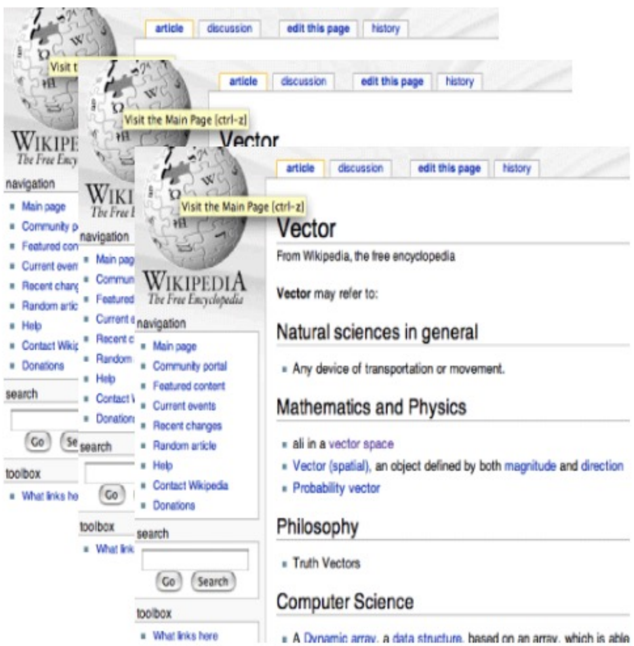
Given a set of input objects, group them to clusters by similarity

Example 1: Cluster videos by people in them



Unsupervised Learning

Example 2: Cluster documents by topic



Physics

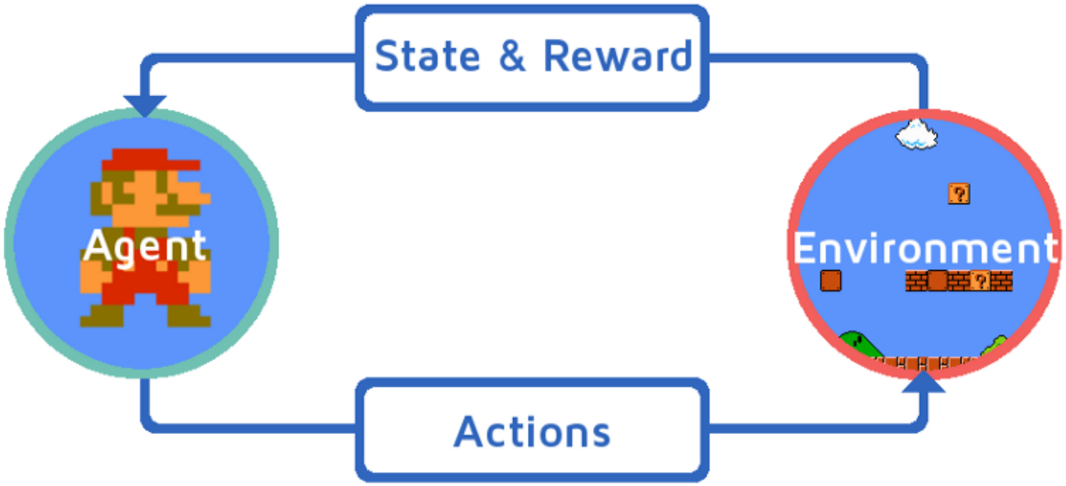
- Gravity
- Laws of Motion
- Electricity

Math

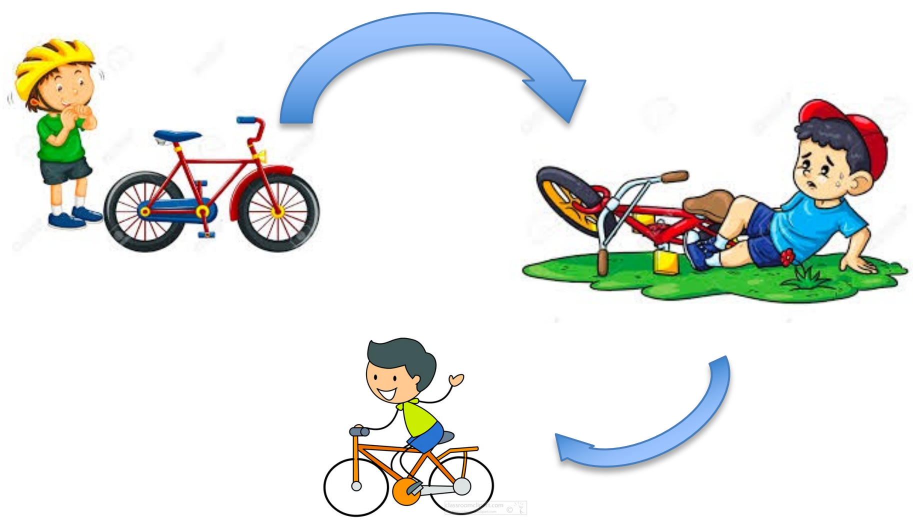
- Geometry
- Algebra

Features: Words in the document

Reinforcement Learning



Rewards from sequence of actions



Reinforcement Learning

- Playing games: Atari, Chess, Checkers
- Robot navigation
- Talent acquisition

Applications of Learning Paradigms

Supervised

- Person identification
- Object recognition
- Stock prediction

Reinforcement

- Game playing
- Credit assignment

Unsupervised

- Social network analysis
- Dimensionality reduction
- Market segmentation

More Recently…

- Combination of these paradigms are being explored:
  – Semi-supervised learning
  – Self supervised learning
  – Supervised + reinforcement
  – ….

History of ML

- 1950s
  - Samuel's checker player
  - Selfridge's Pandemonium
- 1960s:
  - Neural networks: Perceptron
  - Pattern recognition
  - Learning in the limit theory
  - Minsky and Papert prove limitations of Perceptron
- 1970s:
  - Symbolic concept induction
  - Winston's arch learner
  - Expert systems and the knowledge acquisition bottleneck
  - Quinlan's ID3
  - Michalski's AQ and soybean diagnosis
  - Scientific discovery with BACON
  - Mathematical discovery with AM

History of ML

- 2000s
  - Support vector machines & kernel methods
  - Graphical models
  - Statistical relational learning
  - Transfer learning
  - Sequence labeling
  - Collective classification and structured outputs
  - Computer Systems Applications (Compilers, Debugging, Graphics, Security)
  - E-mail management
  - Personalized assistants that learn
  - Learning in robotics and vision
- 2010s
  - Deep learning systems
  - Learning for big data
  - Bayesian methods
  - Multi-task & lifelong learning
  - Applications to vision, speech, social networks, learning to read, etc.
- 2020s
  - Deep learning
  - ???

Based on slide by Ray Mooney

Reading Material

- Steps Towards Artificial Intelligence: https://web.media.mit.edu/~minsky/papers/steps.html

Evaluation Metrics:

Let us design a simple classification algorithm

Purse vs Laptop Bag: Design a classifier








Laptop bag vs. Purse: Design a classifier

- Features:
  - Width
  - Height
  - Weight

- 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 - 50	A (true negative) - 40	C (false positive) - 10
	Positive - 50	D (false negative) - 20	B (true positive) - 30

		Negative	Positive
Actual Class	Negative - 95	A (true negative) - 90	C (false positive) - 5
	Positive - 5	D (false negative) - 3	B (true positive) - 2

Metric	Formula
Average classification accuracy	[(TN + TP) / (TN + FP + TP + FN)]
Class-wise classification accuracy	[TN / (TN + FP) + TP / (TP + FN)]/2
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

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: 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.

More prevalent in research related to medical sciences

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)

Predictive value of a positive result: If the test is positive, what is the probability that the patient actually has the disease

Predictive value of a negative result: If the test is negative, what is the probability that the patient does not have the disease

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.

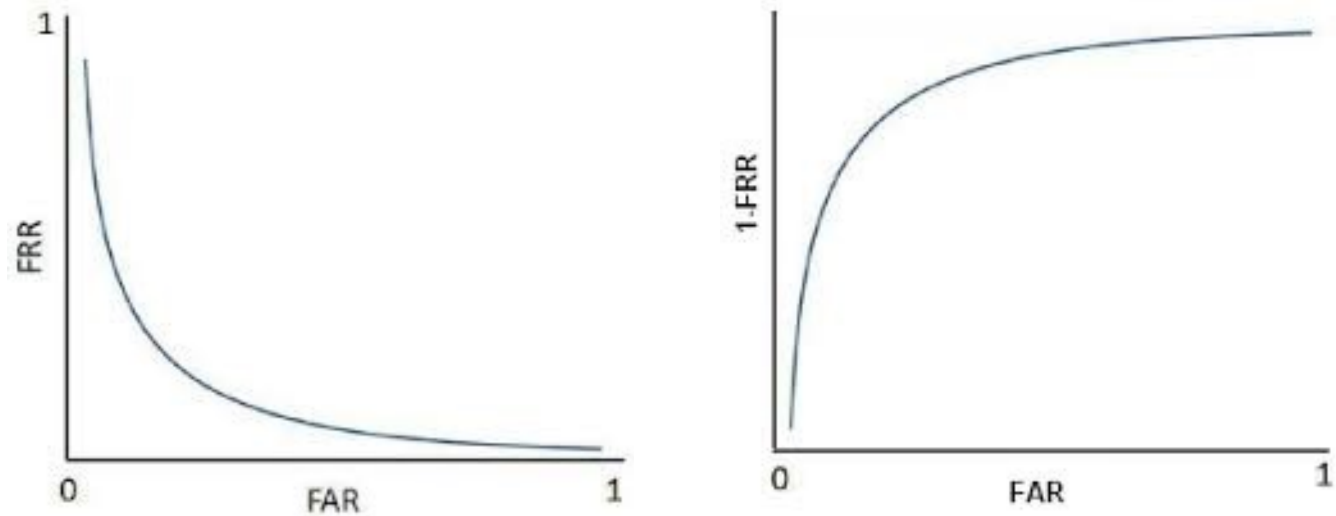
Performance Evaluation

- 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
  - False accept rate vs true accept rate
- Detection error-tradeoff (DET) curve
  - False positive rate vs false negative rate
  - False accept rate vs false reject rate
- Cumulative match curve (CMC)
  - Rank vs identification accuracy

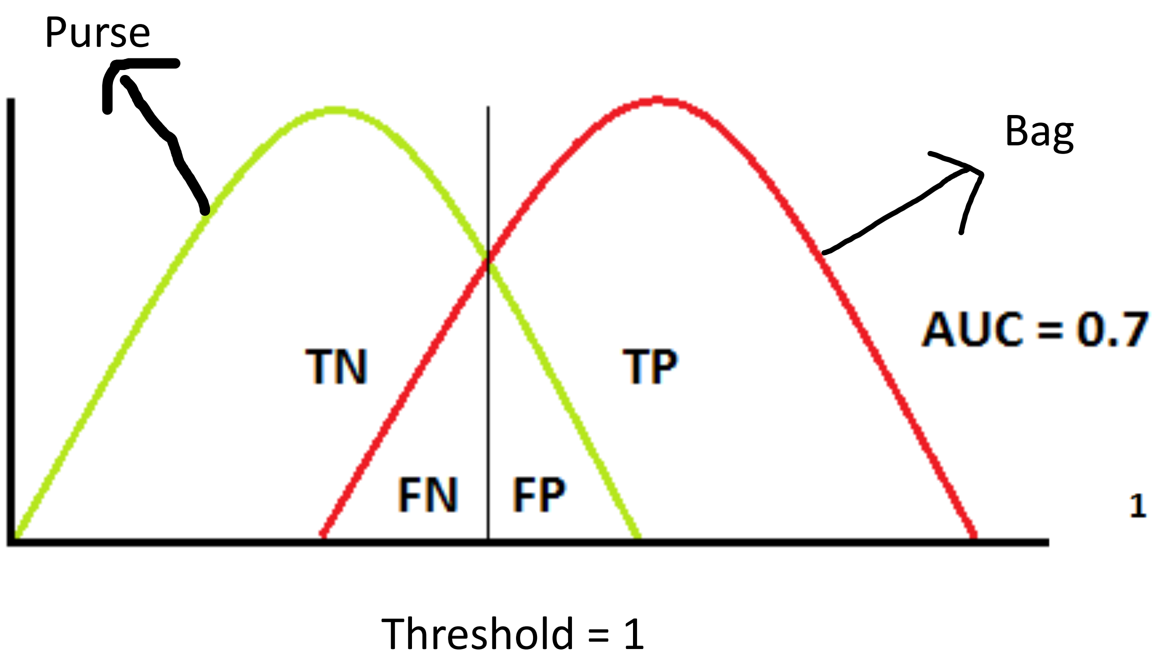
ROC Curve

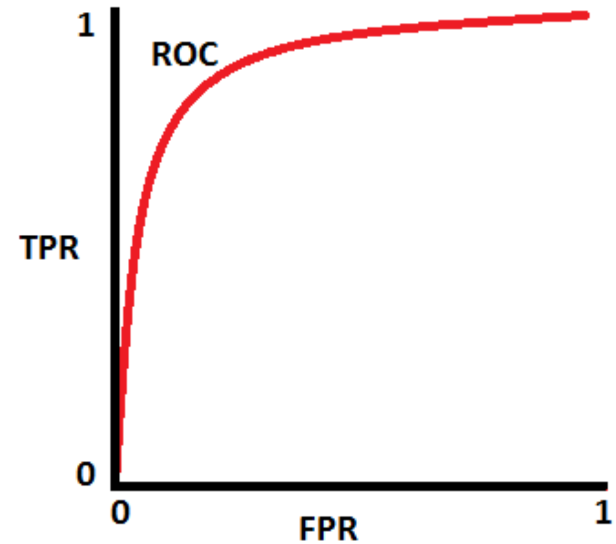


Laptop bag vs. Purse: Design a classifier

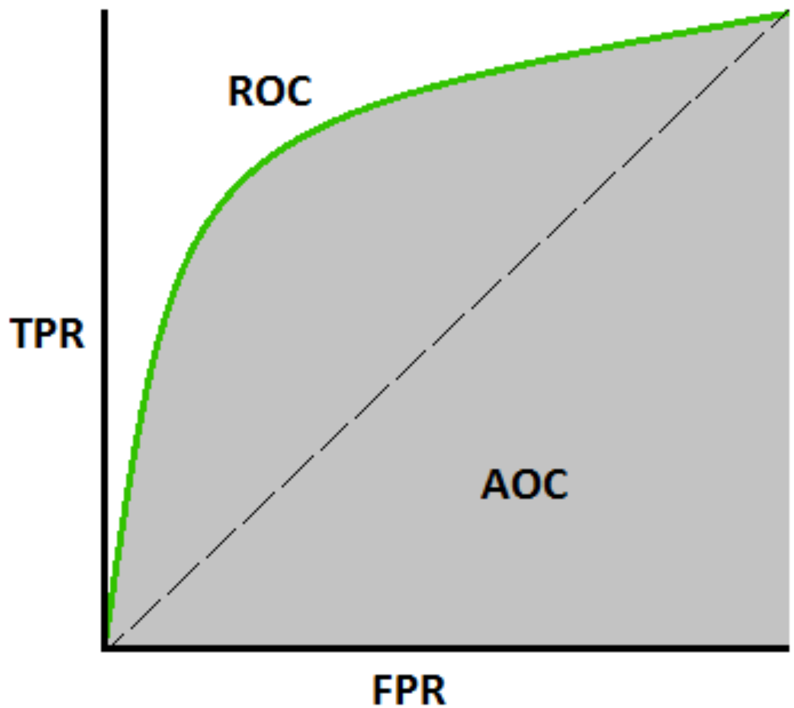
- Features:
  - Ratio of Height to Width

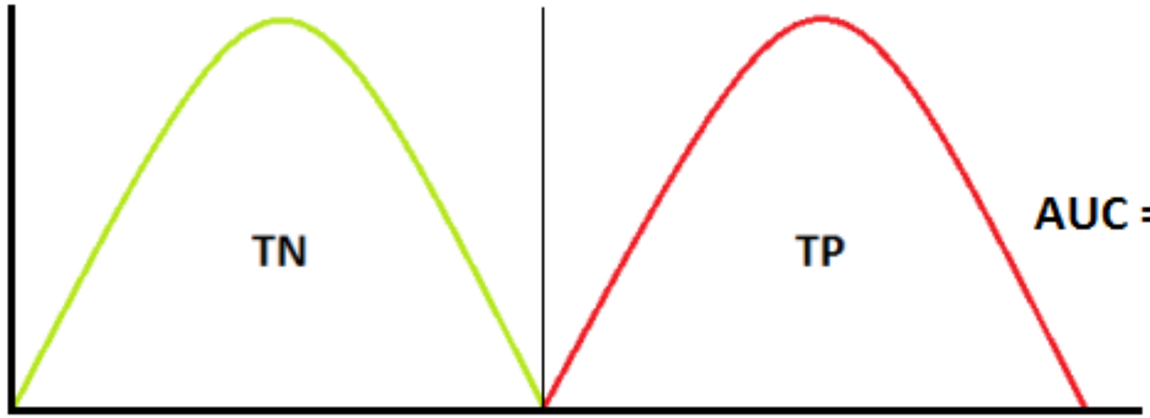
- Classifier: threshold


Purse
Bag
TN
TP
FN
FP
AUC = 0.7
Threshold = 1


1
ROC
TPR
0
0
FPR
1

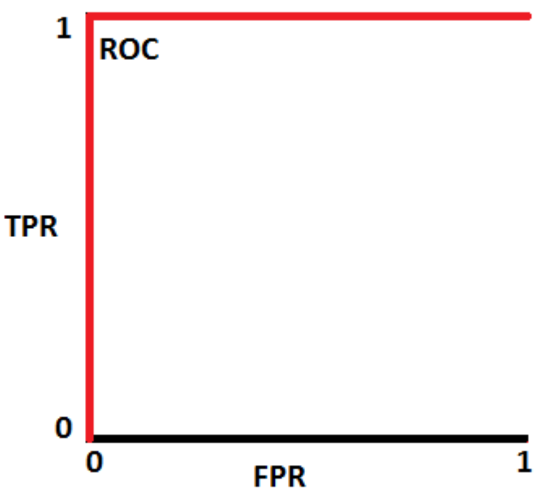
Area Under the Curve

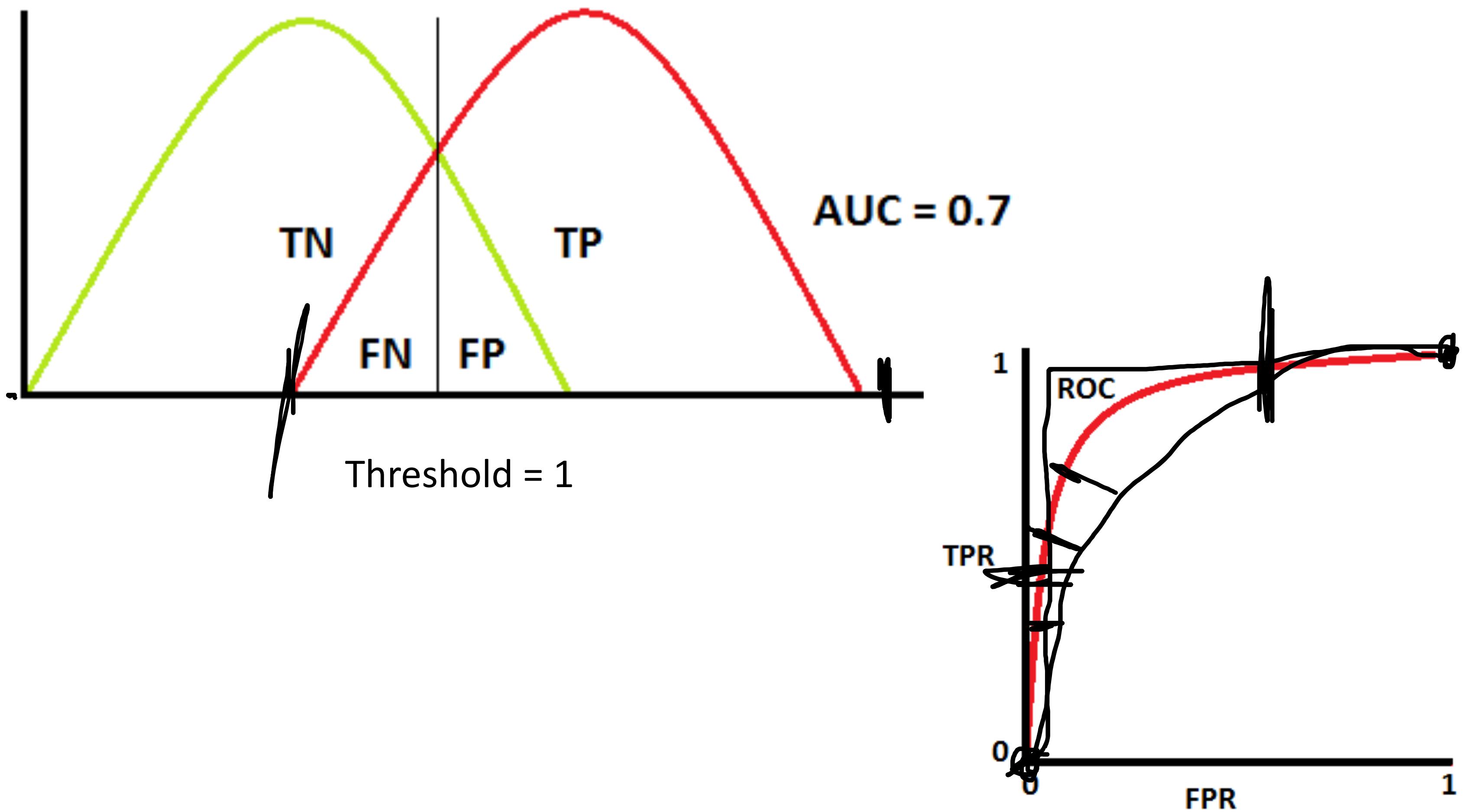




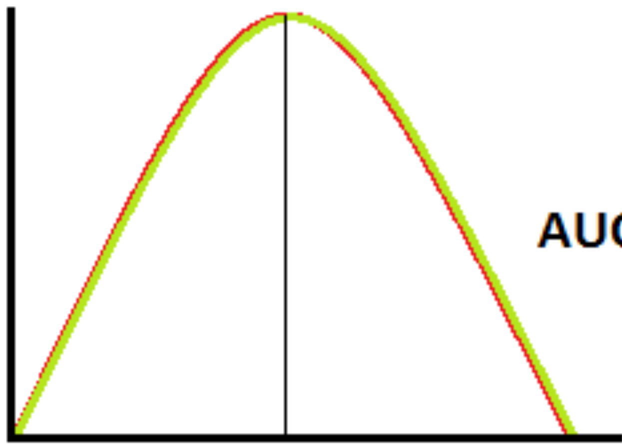
AUC = 1

Threshold = 1



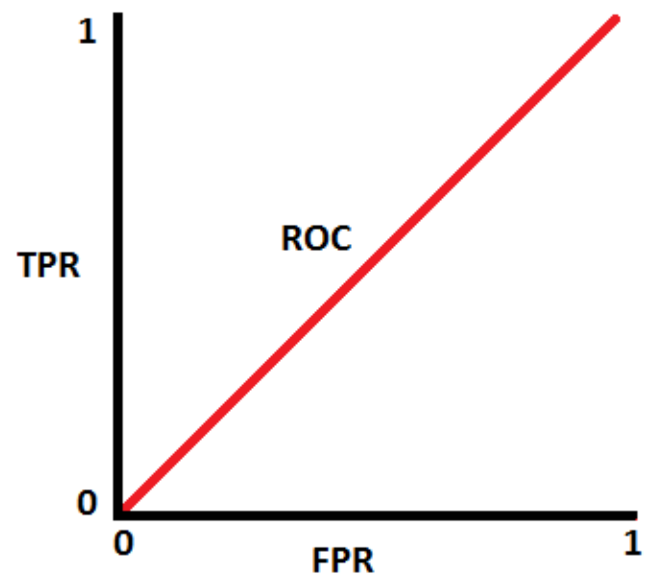


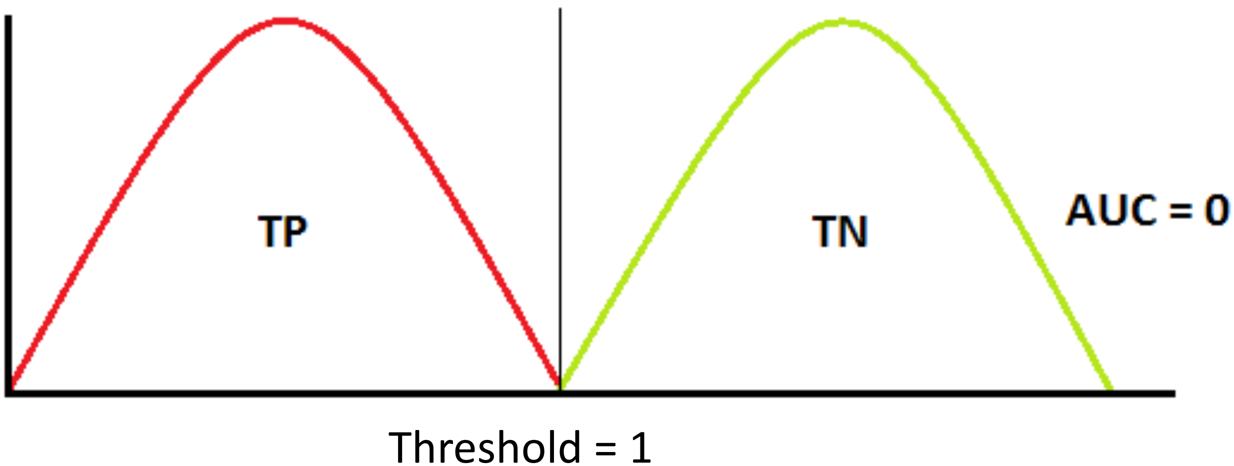
AUC = 0.7
TN
TP
FN
FP
Threshold = 1
1
ROC
TPR
0
0
FPR
1

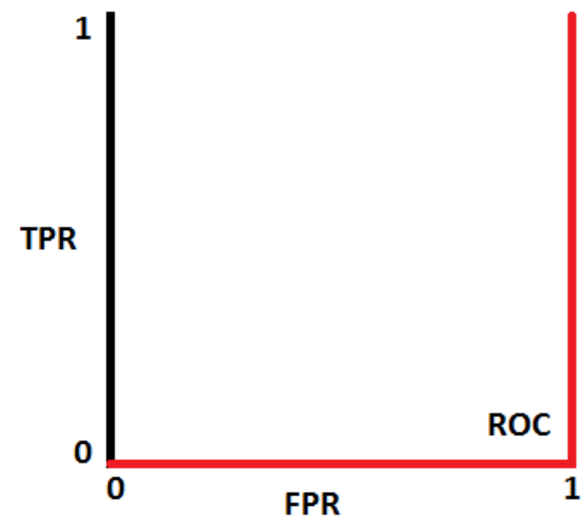


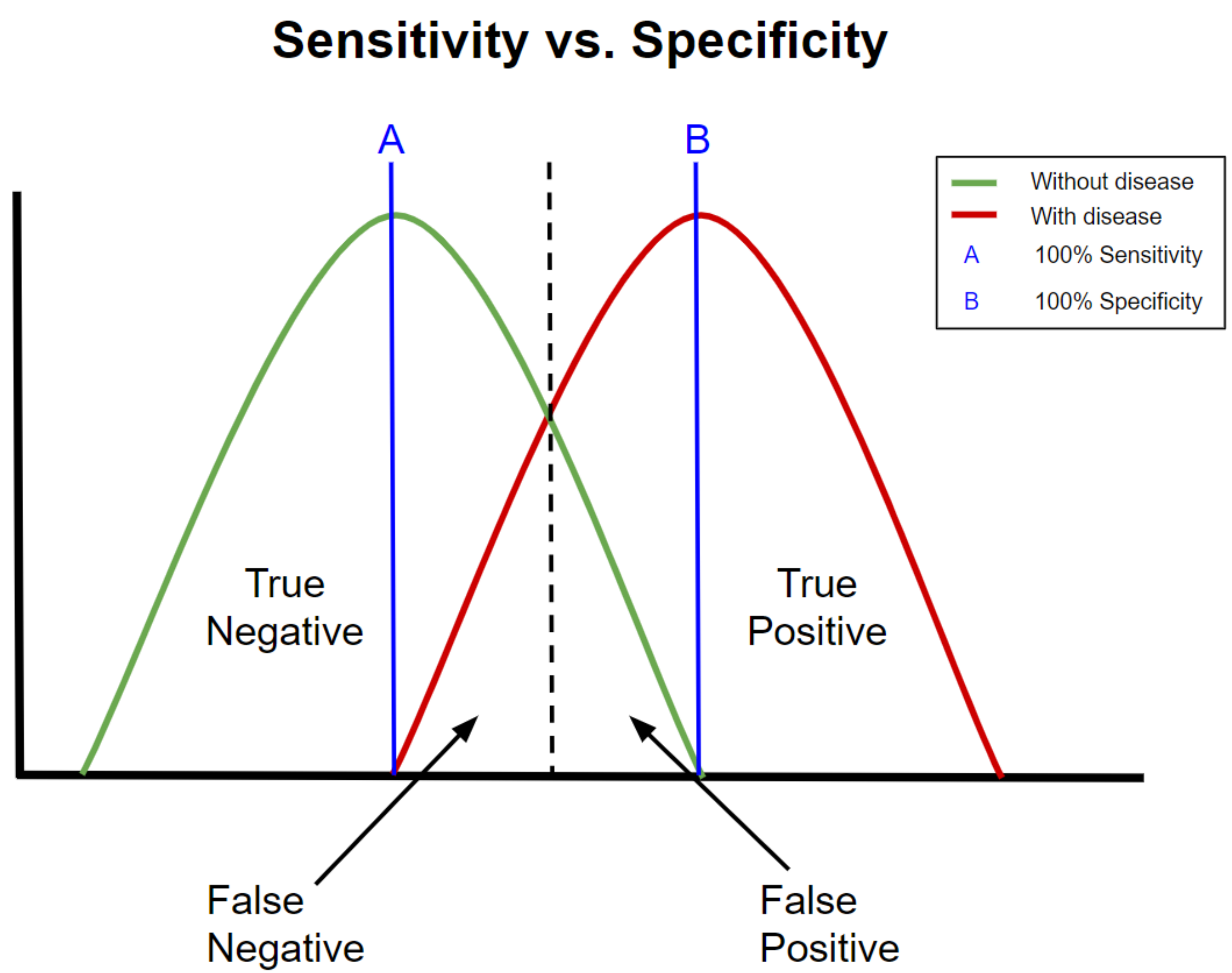
Threshold = 1

AUC = 0.5



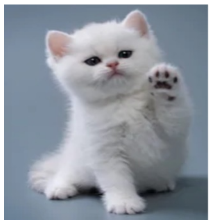

TP
TN
AUC = 0
Threshold = 1


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

CMC

Kitten A



Kitten B

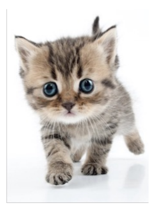
Kitten C

And four test queries:

1

2

3



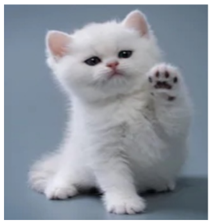
4

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

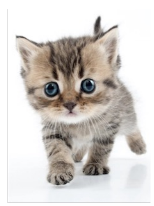
Kitten C

And four test queries:

1

2

3



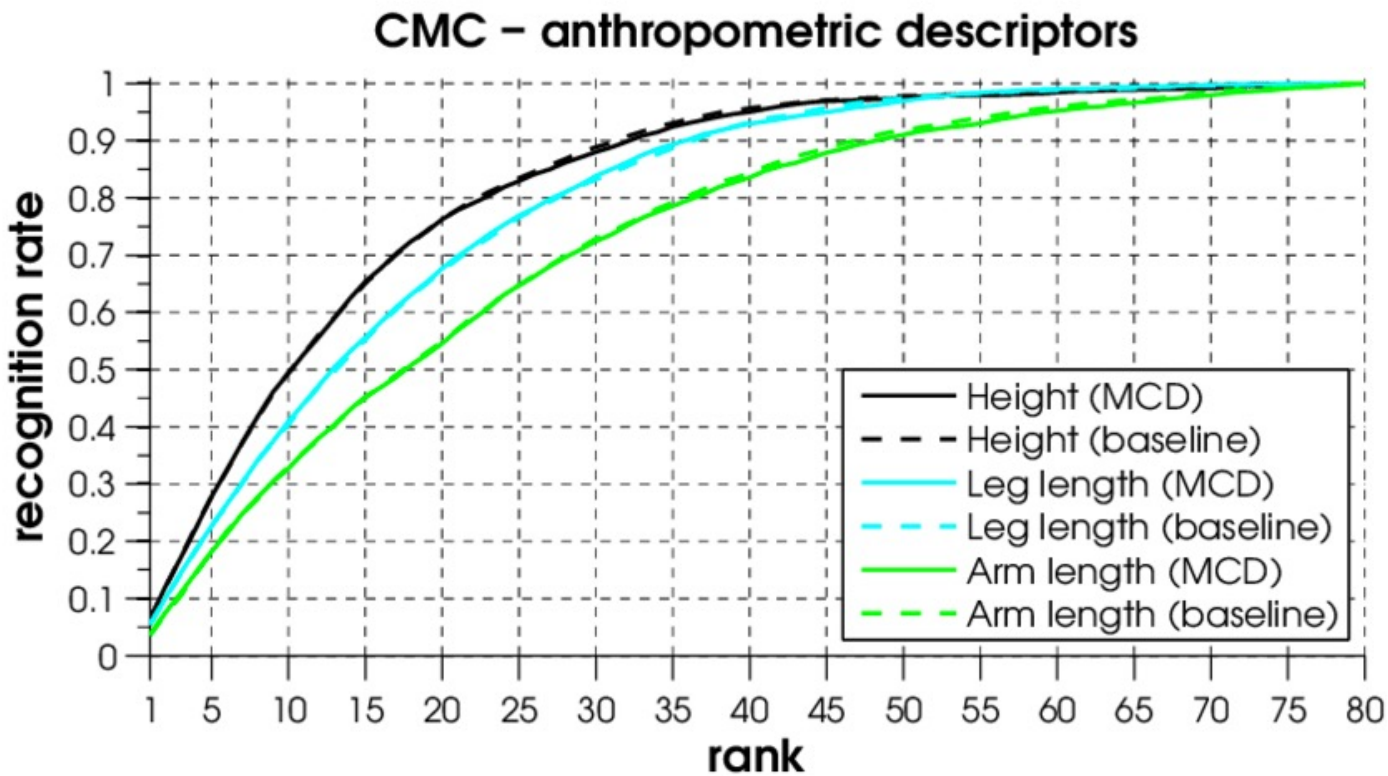
4

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



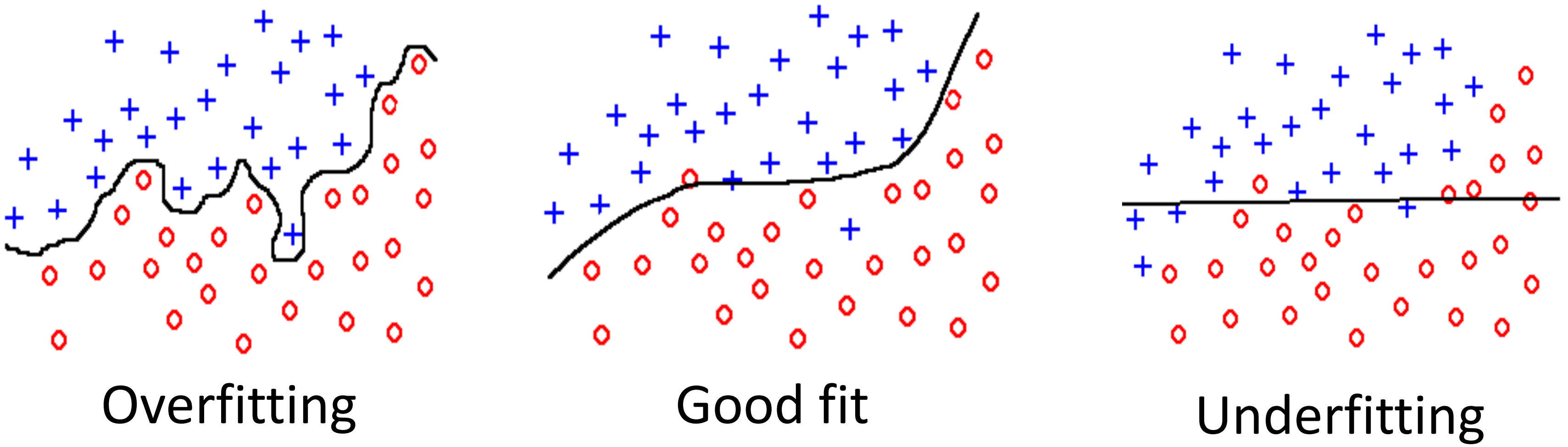
CMC – anthropometric descriptors
recognition rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
rank
1
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
Height (MCD)
Height (baseline)
Leg length (MCD)
Leg length (baseline)
Arm length (MCD)
Arm length (baseline)

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

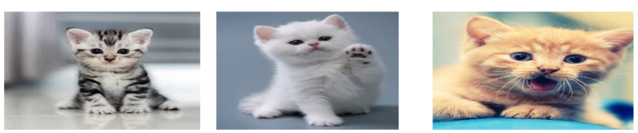
Evaluating ML Systems



- 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

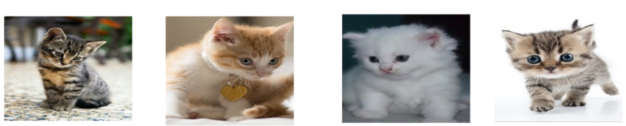
Cross Validation

- “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
  - <u>Performance evaluation</u>: Evaluate the performance of a classifier using the given data
  - <u>Model Selection</u>: Compare the performance of two or more algorithms (DT classifier and neural network) to determine the best algorithm for the given data
  - <u>Tuning model parameters</u>: 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

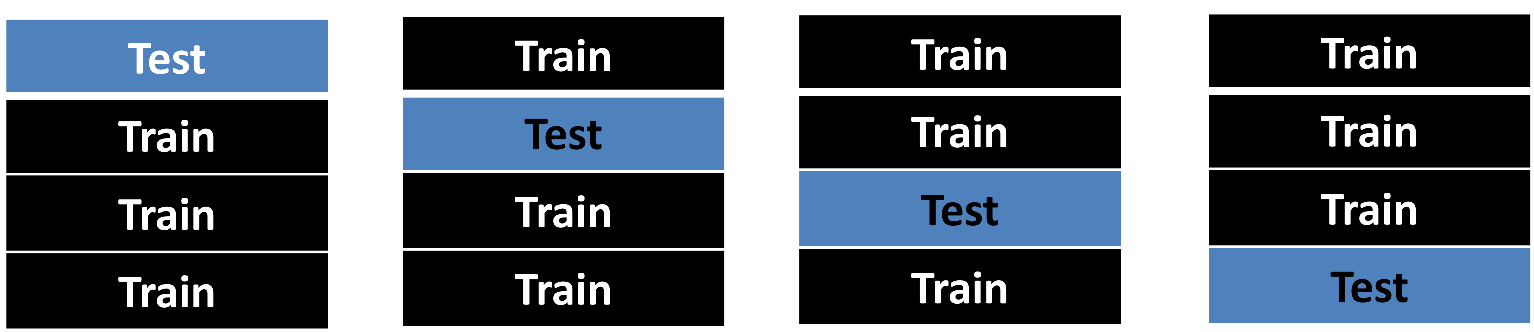
Type of Cross Validation…

- 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