Machine Learning

Classification using Decision Trees

Speaker: Syeda Saleha Raza

Data Science and Machine Learning Workshop, 2017
Habib University

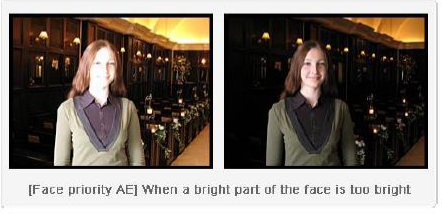
Acknowledgement

- Some slides of this lecture have been taken from:
  - lecture notes of Tan, Steinbach, Kumar Introduction to Data Mining
  - https://web.stanford.edu/class/cs46n

What is Machine Learning?

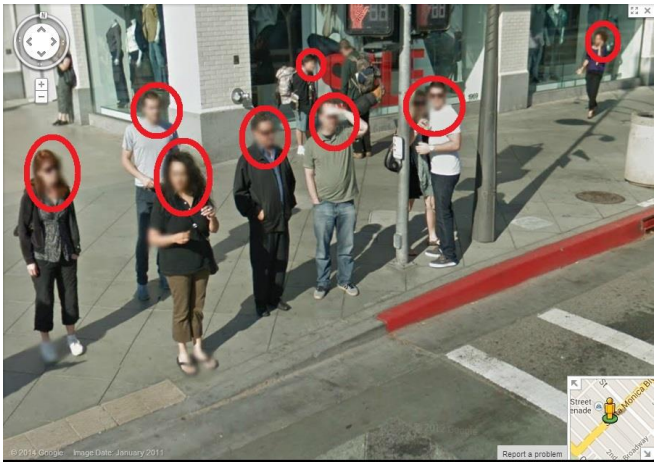
Spam Filtering

Face Detection in Cameras

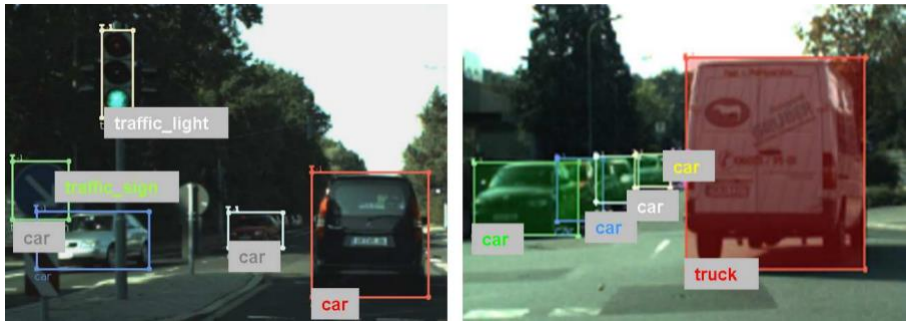


[Face priority AE] When a bright part of the face is too bright

Face Recognition



Object Detection



News Clustering

Why today's **earthquake** – 1200 km away – was felt in Delhi
The Indian Express - 26-Oct-2015
Almost exactly six months after the **Nepal earthquake** that killed nearly 10,000 people, an earthquake of similar magnitude hit north-west ...
Over 260 dead as 7.5 **earthquake** rocks Afghanistan, Pak and India
In-Depth - Hindustan Times - 26-Oct-2015
Explore in depth (4,890 more articles)

400-Plus Quakes Strike San Ramon in 2 Weeks: USGS
NBC Bay Area - 27-Oct-2015
San Ramon, California, appears to have broken a new **earthquake** record over the last two weeks: A total of 408 small ... (Published Tuesday, Oct. 27, **2015**).
Record-Breaking 408 **Earthquakes** Hit Bay Area City Over Past 2 ...
International - Live Science - 27-Oct-2015
Explore in depth (107 more articles)

Afghanistan **earthquake 2015**: Which country has the mo...
City A.M. - 26-Oct-2015
Today, the world was struck by yet another major **earthquake**. This time it was in the mountainous Kush region in northern Afghanistan, close to ...
Man clears rubble after **earthquake**
In-Depth - Economic Times - 27-Oct-2015
Explore in depth (166 more articles)

Recommendations

What is Machine Learning?

- Machine learning is the science of getting computers to act without being explicitly programmed.

Learning from Data

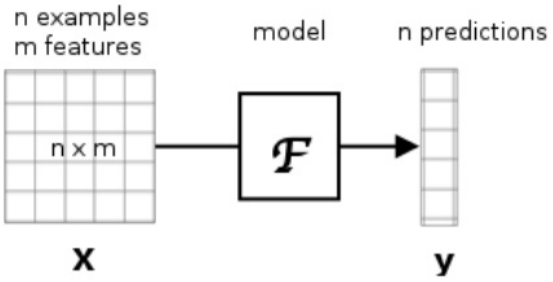
- A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data.

Type of Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning

- Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the outpu



Supervised Learning (contd.)

- **Classification**: A classification problem is when the output variable is a category, such as “Yes/No” or “blue” or “disease” and “no disease”.
- **Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Classification

Classification Example

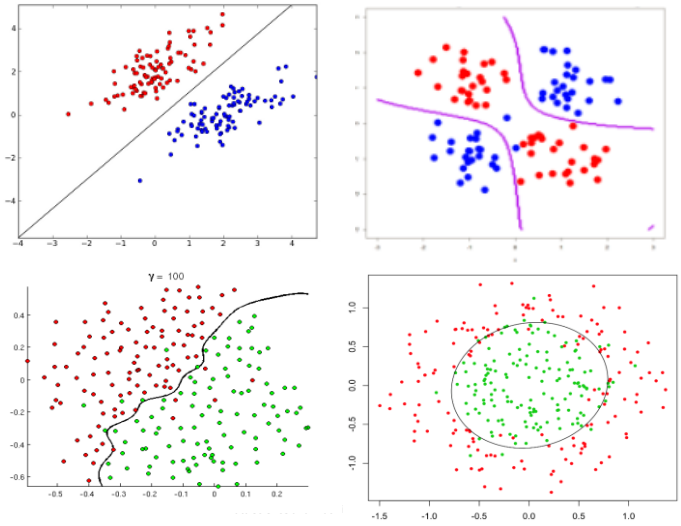
Patient checkup recommendation report

RECORD_ID	AGE	SEX	PAIN_TYPE	BLOOD_PRESSURE	CHOLESTEROL	ECG	HEART_RATE	ANGINA	OLD_PEAK	SLOPE	NUM_VESSELS	THAL	Check
120	41	f	3	112	268	2	172	y	0	1	0	3	not necessary
106	46	f	3	142	177	2	160	y	1.4	3	0	3	maybe
237	97	f	4	262	392	0	345	n	1.4	1	0	3	not necessary
75	67	f	4	106	223	0	142	n	0.3	1	2	3	not necessary
16	71	f	4	112	149	0	125	n	1.6	2	0	3	maybe
124	64	f	4	130	303	0	122	n	2	2	2	3	maybe
146	62	f	4	138	294	0	106	n	1.9	2	3	3	maybe
208	121	f	4	300	679	0	317	y	0.6	1	0	3	maybe
191	121	f	4	324	493	0	297	y	1.4	2	0	3	maybe
62	51	f	4	130	305	0	142	y	1.2	2	0	7	necessary
101	63	f	4	108	269	0	169	y	1.8	2	2	3	maybe
130	66	f	4	178	228	0	165	y	1	2	2	7	necessary
344	103	f	4	248	488	2	319	n	0	1	0	3	not necessary
15	57	f	4	128	303	2	159	n	0	1	1	3	not necessary
693	215	f	4	472	1,171	2	544	n	3.2	2	0	3	maybe
60	128	f	4	300	632	2	268	n	5	2	3	7	necessary
214	62	f	4	140	268	2	160	n	3.6	3	2	3	maybe
206	62	f	4	160	164	2	145	n	6.2	3	3	7	maybe
166	91	f	4	276	479	2	304	y	0.2	2	0	3	maybe
399	104	f	4	277	648	2	282	y	4	2	0	7	necessary

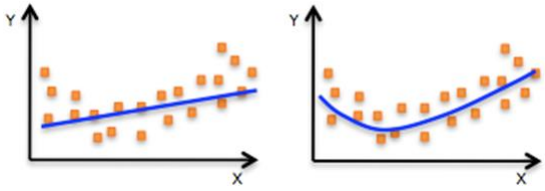
Classification Example

Venue	Type of Wicket	Type of match	Batted first	Winning Team
Pakistan	Slow	ODI	Pakistan	Pakistan
India	Fast	Test	Pakistan	Pakistan
India	Slow	ODI	India	India
Pakistan	Slow	ODI	Pakistan	India
Third country	Fast	ODI	India	Pakistan
India	Fast	ODI	India	India
Pakistan	Fast	Test	India	Pakistan
Third country	Fast	Test	Pakistan	India
Third country	Slow	Test	India	Pakistan
Third country	Slow	ODI	Pakistan	Pakistan
Pakistan	Fast	ODI	Pakistan	India
Third country	Slow	Test	Pakistan	Pakistan
Pakistan	Fast	ODI	India	Pakistan
Third country	Fast	Test	Pakistan	India
India	**Slow**	**ODI**	**Pakistan**	**???**

Decision Boundaries

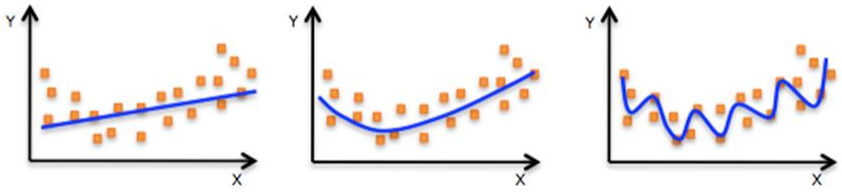


Decision Boundary

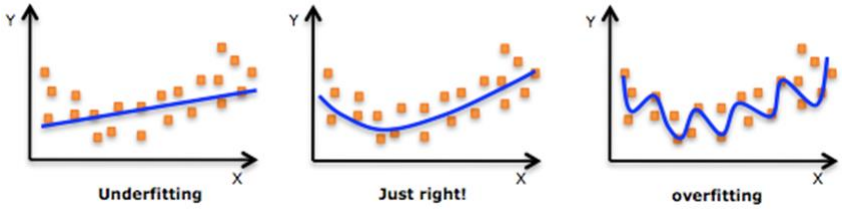


https://web.stanford.edu/class/cs46n

Decision Boundary



Decision Boundary



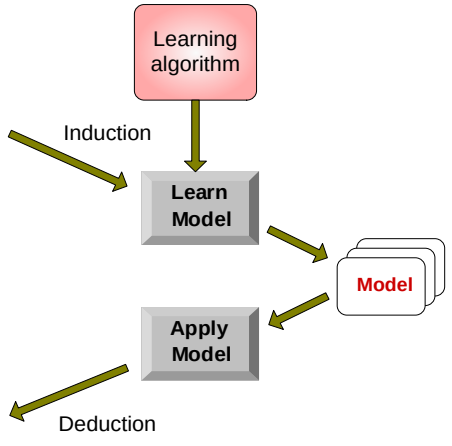
Illustrating Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Tan, Steinbach, Kumar Introduction to Data Mining

Application of Classification

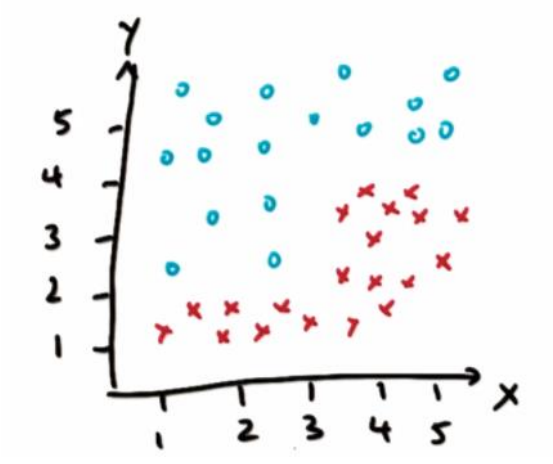
- E-mail Classification (Spam vs. Inbox)
- Object Recognition
- Intrusion Detection
- Loan Defaulter
- Fraud Detection
- Biometric Identification
  - Fingerprinting
  - Handwriting
  - Speech Recognition
- Search Engines

Popular Machine Learning Techniques

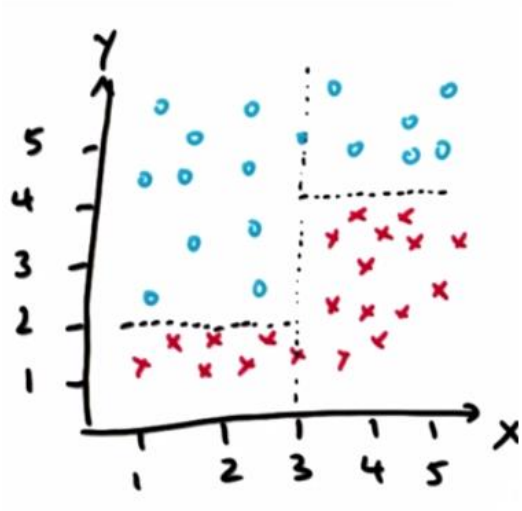
- Classification
  - Decision Trees
  - Naïve Bayes
  - Artificial Neural Networks
  - Logistic Regression
  - Support Vector Machine
  - Random Forest
  - K-Nearest Neighbor

Decision Tree Classifier

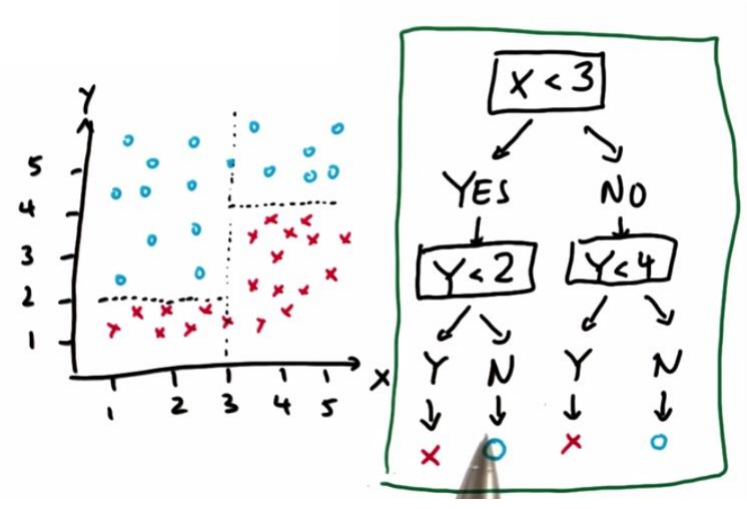
Identifying Decision Boundary



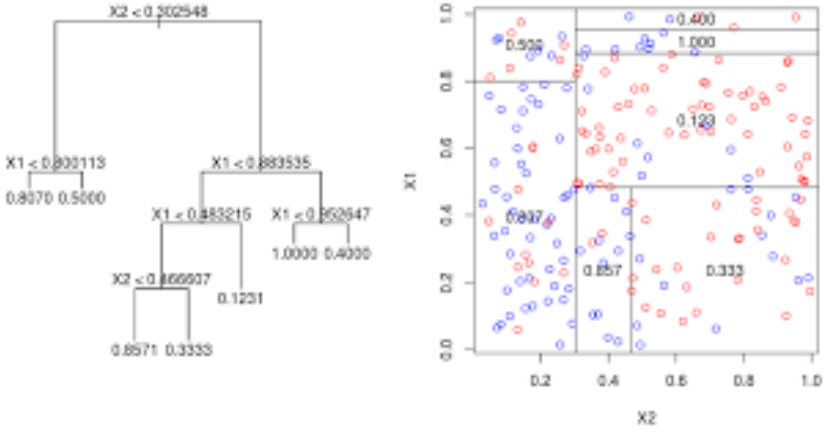
Identifying Decision Boundary



Identifying Decision Boundary



Identifying Decision Boundary



Classification Example

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Classification Tree

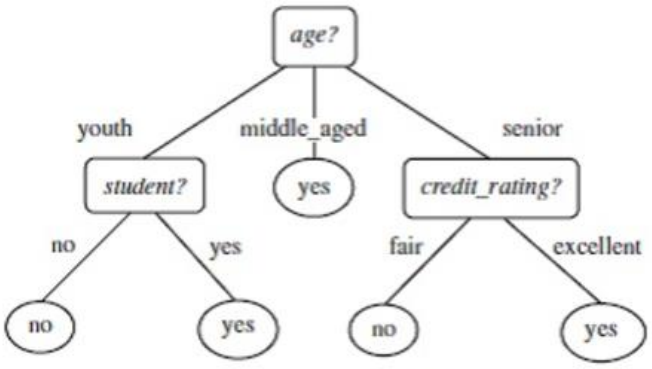
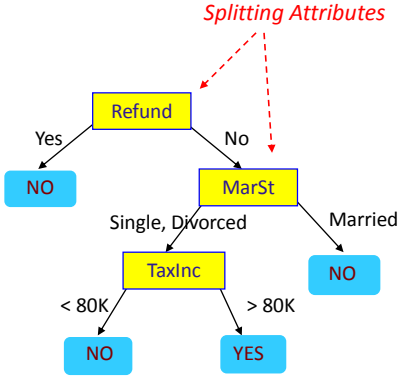


Fig.1: 'Buys Computer?' Decision Tree (Han, Kamber & Pei).

Example of a Decision Tree

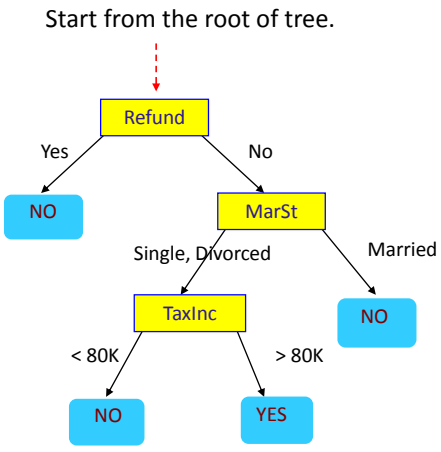
	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



Model: Decision Tree

Apply Model to Test Data

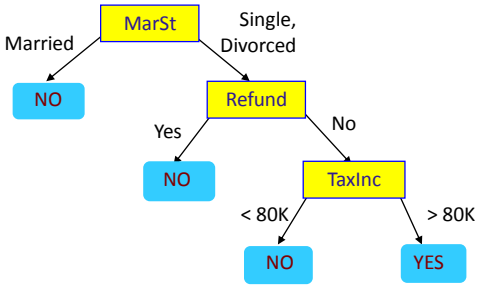


Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

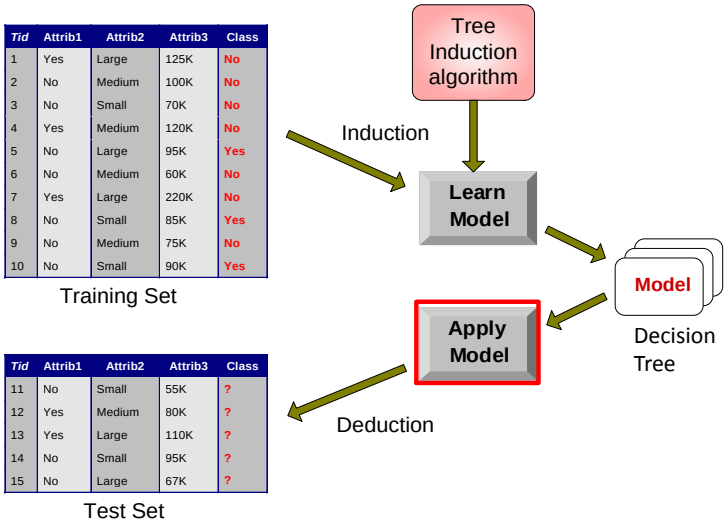
Another Example of Decision Tree

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



There could be more than one tree that fits the same data!

Decision Tree Classification Task



Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Tree Induction

- Greedy strategy.
  - Split the records based on an attribute test that optimizes certain criterion.
- Issues
  - Determine how to split the records
    - How to specify the attribute test condition?
    - How to determine the best split?
  - Determine when to stop splitting

Measuring Impurity

Disease	Symptom 1	Symptom 2	Symptom 3
True	9	5	3
False	1	5	7

Measuring Impurity

- Greedy approach:
  - Nodes with homogeneous class distribution are preferred
- Need a measure of node impurity:

C0: 5
C1: 5

Non-homogeneous,

High degree of impurity

C0: 9
C1: 1

Homogeneous,

Low degree of impurity

Measures of Node Impurity

- Gini Index

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

- Entropy

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

Examples for computing GINI

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Gini = 1 – P(C1)² – P(C2)² = 1 – 0 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Gini = 1 – (1/6)² – (5/6)² = 0.278

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Gini = 1 – (2/6)² – (4/6)² = 0.444

Measure of Impurity: Entropy

- Entropy at a given node t:

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

(NOTE: $p(j \mid t)$ is the relative frequency of class j at node t).

- Measures homogeneity of a node.
  - Maximum ($\log n_c$) when records are equally distributed among all classes implying least information
  - Minimum (0.0) when all records belong to one class, implying most information
- Entropy based computations are similar to the GINI index computations

Examples for computing Entropy

$$Entropy(t) = -\sum_j p(j \mid t) \log_2 p(j \mid t)$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (5/6) = 0.65

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Entropy = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

Inducing a decision tree

- The *key* to building a decision tree - which attribute to choose in order to branch.
- The *heuristic* is to choose the attribute with the minimum GINI/Entropy.

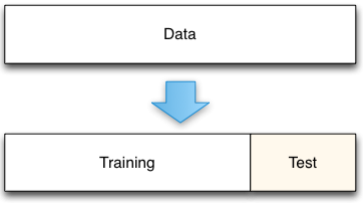
Algorithm for Decision Tree Induction

- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a top-down recursive manner
  - At start, all the training examples are at the root
  - Attributes are categorical
  - Examples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., GINI/Entropy)
- Conditions for stopping partitioning
  - All examples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
  - There are no examples left

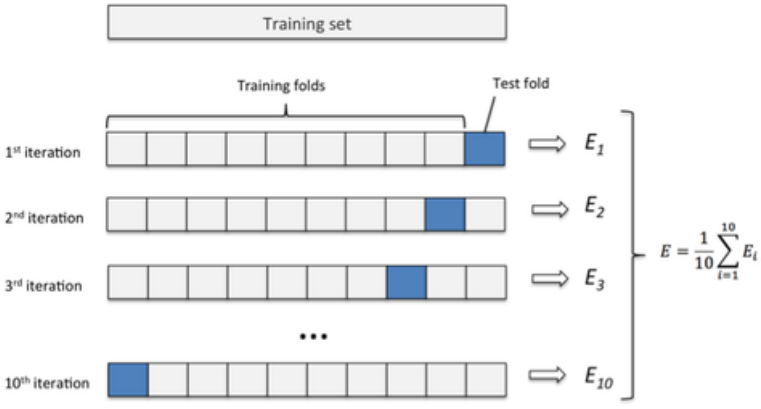
Evaluating a Classifier

Evaluating a Classifier

Partitioning Data into Train & Test



Cross Validation



https://sebastianraschka.com/faq/docs/evaluate-a-model.html

Measuring Classification Accuracy

- Confusion Matrix:

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)
b: FN (false negative)
c: FP (false positive)
d: TN (true negative)

Metrics for Performance Evaluation…

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

- d metric:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10
- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

Classification Accuracy Metrics

Measure	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Misclassification rate (1 – Accuracy)	$\frac{FP + FN}{TP + TN + FP + FN}$
Sensitivity (or Recall)	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
Precision (or Positive Predictive Value)	$\frac{TP}{TP + FP}$

Code Walkthrough

Hand-written digit recognition in Python using scikit-learn
http://bit.ly/2ud6pPZ

Exercise -
Classification in Python using scikit-learn

http://bit.ly/2ud6pPZ