Applications

- Text classification
 - Classify emails into spam / non-spam
 - NLP Problems
 - · Tagging: Classify words into verbs, nouns, etc.
- Risk management, Fraud detection, Computer intrusion detection
 - Given the properties of a transaction (items purchased, amount, location, customer · Determine if it is a fraud
- Machine learning / pattern recognition applications

 - · Speech recognition etc.
- All of science & knowledge is about predicting future in terms of past
 - So classification is a very fundamental problem with ultra-wide scope of applications

We collect different measurements/facts/about certain features

$$\bullet \ x = (x_1, x_2, \dots, x_d)$$

• In the above example, x_1, x_2 are diameter and weight

•
$$y \in \{1, 2, ..., K\} = [K]$$

• If K=2 binary classification, else, multi-class classification

Bayes Classifier

- Suppose $Y \in \{1, 2, ..., K\}$,
- P(Y = k) Prob that one observes a data point from class k (Prior)
- P(Y = k | X = x) Conditional Probability (the label of the data point point x) = k (Posterior)
- P(X = x | Y = k) Probability distribution on X given that the data point is in class k (Likelihood)

· Which class we assign?

•
$$k \in argmax_{j \in [k]} \Pr(Y = j | X = x)$$

- One can argue it is optimal classifier
- · What is optimal? Boyes error Y.

Naïve Bayes Classifier

What can we say using Bayes theorem for Posterior and Prior?

• Posterior =
$$\frac{Prior * Liklihood}{P(X=x)}$$

Outlook	Temp (°F)	Humidity (%)	Windy?	Class
sunny	75	70	true	play
sunny	80	90	true	don't play
sunny	85	85	false	don't play
sunny	72	95	false	don't play
sunny	69	70	false	play
overcast	72	90	true	play
overcast	83	78	false	play
overcast	64	65	true	play
overcast	81	75	false	play
rain	71	80	true	THE REAL PROPERTY.
rain	65	70	true	don't play
rain	75	80	false	don't play
rain	68	80	false	play
rain	70	00	folas	play play

Decision Tree

- Recall cricket-ball vs tennis-ball example
- $x = (x_1, x_2) \in \mathbb{R}^2$
- $y \in \{0,1\}$ 1 is 'cricket' 0 is 'tennis'
- If weight of a ball is > 100 gm, ball is cricket ball else tennis ball
- · Mathematically,
- $f(x) = 1 \text{ if } x_1 > 100$ = 0 ow

ID3

- Iterative Dichotomizer Ross Quinlan in 1980s
- · Which feature should be used to make decision?
 - · The one which will separate all classes well
 - Information theoretically, the one from which we have highest information
 - · How to measure Information Gain?

Entropy

- I need to send you one of the two observations
 - It is raining today
 - It is not raining today
- How many bits I need to send you?
- · One bit
 - ='0' means It is raining today
 - ='1' means It is not raining today
- Suppose there is four messages: Two bits

- It is raining today and is sunny
- · It is raining today and is cloudy
- It is not raining today and is sunny
- · It is not raining today and is cloudy
- Suppose it is never going to be rainy and sunny

- Can we do better than 'two bits' in the three message or do we need one bit '0' to say no rain everyday?
- Every system of discrete signals has some 'information' and we need those many bits to represent the system

• Entropy =
$$\sum_{i} p_i * \log(\frac{1}{p_i})$$

Information Gain

• Entropy
$$H(S) = \sum_{i \in [n]} \frac{S_i}{S} * \log \frac{S}{S_i}$$

Information Gain due to A

$$IG(S,A) = H(S) - H(S|A)$$

$$= H(S) - \sum_{t} \frac{S_t}{S} H(S_t)$$

where, $t \in T$ is the different values present in S

Sunny 36% Mild 45% 26% Weak 57% Frue 9 64% Chef (4) 29% Other (4) 28% Unique values Strong 43% false 5 36% Sunny Hot High Strong No Overcast Hot High Beak Yes Rain Cool Normal Reak Yes Rain Cool Normal Strong No Overcast Cool Normal Strong No Overcast Cool Normal Strong No Overcast Cool Normal Strong No Reak Yes Rain Cool Normal Strong Yes Sunny Cool Normal Strong Yes Rain Rain Rain Raid Normal Strong Yes Rain Rain Raid Roreal Reak Yes Rain Rain Raid Roreal Reak Yes Rain Rain Raid Roreal Rain Raid Roreal Rain Yes Rain Raid Rain Rain Raid Roreal Rain Rain Raid Rain Raid Rain Rain Rain Rain Raid Rain Rain Rain Rain Rain Rain Rain Raid Rain Rain Rain Rain Rain Rain Rain Rain	Application of the second		5.01.5.00	Humns
Hain 36% Hot 26% unique values Strong 43% 9 64% faise	utlook	ature = 4 Homidity = 4 Wind	F Y Play Tennis F	
Rain 36% Hot 28% 2 Unique values Strong 43% false 5 36% Sumy Hot High Weak No Sumy Hot High Strong No Overcast Not High Weak Yes Rain Cool Normal Weak Yes Rain Cool Normal Strong No Overcast Cool Normal Strong No Overcast Cool Normal Strong No Normal Normal Strong No Normal Norm	ıy.			
Cther(4) 28% Other(4) 28% faise Summy Hot High Weak No Summy Hot High Strong No Overcast Not High Weak Yes Bain Wild High Weak Yes Rain Cool Normal Weak Yes Rain Cool Normal Strong No Overcast Cool Normal Strong No Summy Mild High Weak Yes Summy Mild Normal Strong Yes Overcast Mild Normal Strong Yes Overcast Mild High Strong Yes Overcast Mild High Strong Yes			WAN D	
Surny Hot High Weak No Surny Hot High Strong No Overcast Hot High Weak Yes Rain Nild High #eek Yes Rain Cool Normal Weak Yes Rain Cool Normal Strong No Overcast Cool Normal Strong Yes Sunny Kild High Weak Yes Sunny Hild Normal Weak Yes Rain Nild Normal Strong Yes Sunny Hild Normal Weak Yes Rain Nild Normal Strong Yes Sunny Hild Normal Strong Yes	r (4)	28%	faise	
Overcast Hot High Wesk Yes Bain Nild High Wesk Yes Rain Cool Normal Wesk Yes Rain Cool Normal Strong No Overcast Cool Normal Strong Yes Dunny Maid High Renk No Sunny Cool Normal Wesk Yes Sunny Cool Normal Wesk Yes Sunny Cool Normal Wesk Yes Sunny Hild Normal Wesk Yes Dunny Hild Normal Wesk Yes Dunny Hild Normal Strong Yes Dunny Hild Normal Strong Yes Decent Hild High Strong Yes	y	High Weak		
Rain Nild High Strong West Rain Cool Normal West Yes Rain Cool Normal Strong No Overcast Cool Normal Strong West Sunny Kild High Best No Rein Nild Horsel West Yes Sunny Cool Normal West Yes Sunny Hild Normal Strong Yes Descript Hild Normal Strong Yes Sunny Hild Horsel Best Yes Sunny Hild Normal Strong Yes Descript Hild High Strong Yes	y	High Strong	No	
Rain Cool Normal Heak Yes Rain Cool Normal Strong No Descript Cool Normal Strong Yes Survey Mild High Reak No Sunny Cool Normal Weak Yes Hain Nild Normal Weak Yes Sunny Rild Normal Strong Yes Mormal Weak Yes Sunny Rild Normal Strong Yes Descript Rild Normal Strong Yes Descript Rild High Strong Yes		High West	Yes	
Rain Cool Normal Neak Yes Rain Cool Normal Strong No Overcast Cool Normal Strong Yes Sunny Mild High Neak No Hein Nild Normal Normal Yes Sunny Rile Normal Strong Yes Overcast Rile Normal Strong Yes Overcast Rile High Strong Yes		High Book		
Overcast Cool Normal Strong No Overcast Cool Mormal Strong Yes Sunny Mild High Meek No Sunny Cool Mormal Week Yes Nein Mild Hormal Week Yes Sunny Mile Normal Strong Yes Overcast Mild High Strong Yes		Normal Weak		
Overcast Cool Normal Strong Yes Sunny Kild High Reak Ho Sunny Cool Hormal Reak Yes Hein Nild Hormal Reak Yes Sunny Rile Normal Strong Yes Overcast Rild High Strong Yes		Normal Strong		
Sunny Mild High Mesk No Sunny Cool Mormel Week Yes Mein Mild Normel Week Yes Sunny Hile Normel Strong Yes Overcast Mild High Strong Yes		Mormal Strong		
Mein Mild Normal Weak Yes Sunny Mild Normal Strong Yes Overcost Mild High Strong Yes Overcast Mild High Strong Yes		tics		
Nain Mild Normal Beak Yes Summy Mild Normal Strong Yes Overcest Rild High Strong Yes		Hormal Weak		
Overcast Rild High Strong Yes Description of the Strong Yes		Moreot		
Drercast Rid High Strong Yes		Norma's		
Process 105		High		
	resut	Hormal	Yes	
Rain Hile Teak Yes	6 37	NOAK (Yes	
High Strong No.		High Strong	No.	