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 profile, etc.)
  - Determine if it is a fraud
- Machine learning / pattern recognition applications
  - Vision
  - 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

- $x = (x_1, x_2, \ldots, x_d)$
  - In the above example, x_1, x_2 are diameter and weight

- $y \in \{1,2, \ldots, K\} = [K]$

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

Bayes Classifier

- Suppose $Y \in \{1,2,\ldots,K\}$,
- $P(Y = k)$ 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? — Bayes error r.

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	don't play
rain	65	70	true	don't play
rain	75	80	false	play
rain	68	80	false	play
rain	70	96	false	play

Decision Tree

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

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 gain
  - 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

Detail Compact Column

5 of 5 columns

Outlook		Temperature		Humidity	Wind		Play Tennis
Sunny	36%	Mild	43%	2 unique values	Weak	57%	true 9 64%
Rain	36%	Hot	29%		Strong	43%	false 5 36%
Other (4)	29%	Other (4)	29%				
Sunny		Hot		High	Weak		No
Sunny		Hot		High	Strong		No
Overcast		Hot		High	Weak		Yes
Rain		Mild		High	Weak		Yes
Rain		Cool		Normal	Weak		Yes
Rain		Cool		Normal	Strong		No
Overcast		Cool		Normal	Strong		Yes
Sunny		Mild		High	Weak		No
Sunny		Cool		Normal	Weak		Yes
Rain		Mild		Normal	Weak		Yes
Sunny		Mild		Normal	Strong		Yes
Overcast		Mild		High	Strong		Yes
Overcast		Hot		Normal	Weak		Yes
Rain		Mild		High	Strong		No