Machine Learning-
Theory and Practice

1. LECTURE 1: BAYESIAN LEARNING - BASIC CONCEPTS
2. LECTURE 2: NAÏVE BAYES CLASSIFIER
3. LECTURE 3: BAYESIAN NETWORKS

Lecture 1:

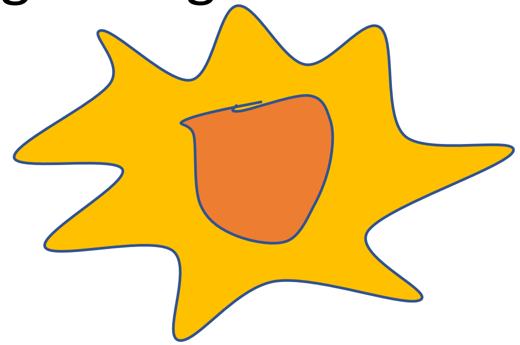
Bayesian Learning - Basic Concepts

Different Styles of Learning

- Discriminative
- Generative

Lion or Tiger? Discriminative style

Distinguishing feature: A male lion has a mane while a tiger does not have a mane



P(Male lion|Mane) >>P(Male tiger|Mane)

Distinguishing feature: A tiger has striped skin while a lion does not

Discriminative Style for learning

- Discriminative Style: Identify on the basis of distinguishing features

  - *If you see an animal with mane it is mostly lion, not tiger.*

    P(Lion|Mane) is high, P(Tiger|Mane) is low

  - *If you see an animal with stripes, it is most probably tiger*

    *P(*Tiger|Stripes) is high, *P(*Lion|Stripes) is low

- In general, form of P(y|X) is known, and Model fitting is carried out on historical data. Example:
  1. Logistic Regression assumes the form: $P(y|x) = 1/(1+e^{-w.x})$
  2. Estimates w-parameters by maximizing Likelihood.
  3. *For new data the trained classifier calculates **y|X** using this relationship*

Generative Style of learning – Given a Class, get general idea about its features

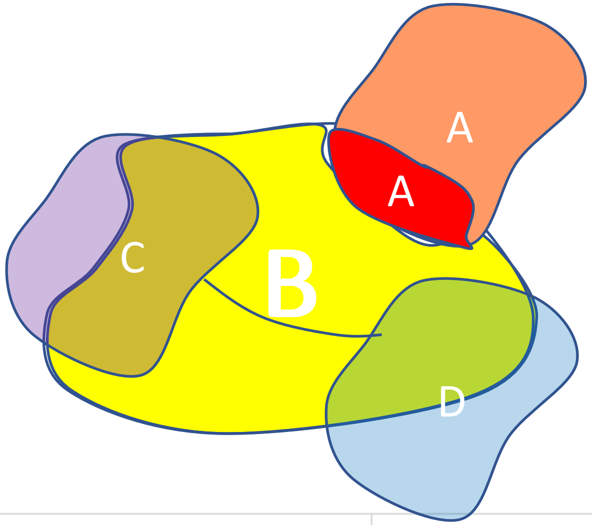
Generative Style of Learning

- *Generally, tigers have stripes, golden yellow skin colour, are longer*
- *Generally, male lions have a mane, light brown skin colour, are shorted*
- Probability of a set of features X, given a Class y:
  - P(*stripes, golden yellow skin, long*|tiger)
  - P(*mane, dull brown skin, short*|male lion)
- Learn P(X|y) from historical data, then find P(**y|X**) for new data by using Bayes Rule

Revision of basic Laws of Probability

1. Rule of Addition/Union	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
2. Rule of Product/Intersection:	
Dependent Events	$P(A \cap B) = P(A\|B) \cdot P(B) = P(B\|A).P(A)$
Independent Events	$P(A \cap B) = P(A) \cdot P(B)$

Revision of basic Laws of Probability



3. Conditional Probability:	$P(A \mid B) = P(A \cap B) / P(B)$ $P(B\|A) = P(A \cap B)/ P(A)$

4. Total Probability

$P(B)=\sum_i P(B \cap A_i)=\sum_i P(B|A_i)P(A_i)$.

Revision of basic Laws of Probability

Bayes Formula	

Consider :
A to be hypothesis or possible outcome h
B to be attributes or evidences gathered from Data D
So:

$$P(h|D) = \frac{P(D|h)}{P(D)} P(h)$$

Bayesian Training: Learn Priors and Likelihood

- PRIOR PROBABILITIES P(h): These are overall probabilities of possible outcomes / hypothesis in nature:
  - P(Cancer) = .008   P(Not Cancer) = 0.992

$$P(h|D) = \frac{P(D|h)}{P(D)} \boldsymbol{P(h)}$$

- LIKELIHOODS P(D|h): These are conditional probabilities of evidences (Data) given an outcome / hypothesis exists:
  - P(Lab test is +|Cancer)= 0.98
  - P(Lab test is +|Not Cancer) = 0.03
  - P(Lab test is –ve|Cancer)= 0.02
  - P(Lab test is –ve |Not Cancer)= 0.97

$$P(h|D) = \frac{\boldsymbol{P(D|h)}}{P(D)} P(h)$$

- Priors and Likelihoods are calculated from historical data

Evaluate Aposterior Probabilities

- APOSTERIORI P(h|D): Probability of hypothesis given evidence
- P(D): Total probability – A normalization factor: $\mathrm{P(D)} = \sum_{h \in H} \mathrm{P(h|D)}$
- BAYESIAN LEARNING predicts the possibility of each hypothesis, given some evidence D:

$$\forall h \in H: \boldsymbol{P(h|D)} = \frac{P(D|h)}{P(D)} P(h)$$

- Posterior probability = $\frac{Likelihood}{Total\ Probability} \times Prior$

Maximum A-Posteriori MAP

- BAYESIAN LEARNING predicts the possibility of each hypothesis, given some evidence D:

$$\forall h \in H: \quad P(h|D) = \frac{P(D|h)}{P(D)} P(h)$$

- Prediction: Find Maximum APosteriori hypothesis:

$$P(h|D) = Argmax_{h \in H}\{P(h|D)$$

- Note: Denominator remains same for each P(h|D)
- MAP prediction is:

$$Argmax_{h \in H}\{P(D|h) \times P(h)$$

Maximum Likelihood

- If all hypotheses are equiprobable, then Priors are equal:

$$P(h_1) = P(h_2) \ldots \ldots .$$

- Then, we predict Maximum Likelihood ML hypothesis:

$$P(h|D) = Argmax_{h \in H}\{P(D|h\}$$

Cancer Case:

-
  - P(Cancer) = .008 P(Not Cancer) = 0.992
  - P(Lab test is +|Cancer)= 0.98
  - P(Lab test is +|Not Cancer) = 0.03
  - P(Lab test is –ve|Cancer)= 0.02
  - P(Lab test is –ve |Not Cancer)= 0.97

- The two hypotheses are not equiprobable. Calculate MAP.
- Given a new patient with +ve test result:
- Prediction: **Argmax**{P(+ve|Cancer}xP(cancer), P(+ve|Not Cancer}xP(Not cancer)}

$=Argmax\{0.98 \times 0.008, 0.03 \times 0.992\} = Argmax\{0.00784, 0.02976\}$

Not Cancer

Exact Probabilities

- { P(Cancer) = 0.98 x 0.008, P(no cancer) 0.03x0.992}

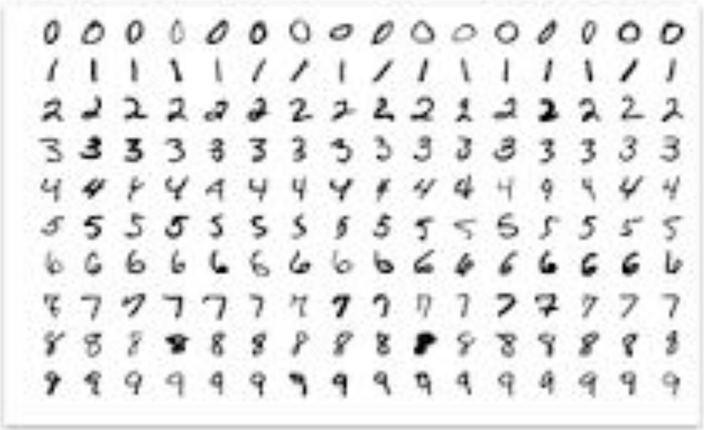
= {.0078, .0298}

$$\text{Thus } P(\text{Cancer}|+ve) = \frac{.0078}{.0078+.0298}$$

$$\text{Thus } P(\text{No Cancer}|+ve) = \frac{.0298}{.0078+.0298}$$

Handling Multiple Features

- So are, we considered a single evidence
- What if we have multiple features?
- Examples: Distinguish between handwritten images 1 and 7



Training data

- Classify text documents into Science and Arts based on words

Handling Multiple Features

- Digit Recognition: A set of 128x128 images. Each of 4096 pixels is an individual feature.
- Document Classification: A vocabulary of 10000 words spanning all documents!
- Let $D = \{x_1, x_2, x_3, x_4 \dots \dots\}$

Now , classifier needs to evaluate the Joint probabilities for all combination of features:

$$\forall h \in H: \quad P(h|x_1, x_2, x_3, x_4 \dots \dots) = \frac{P(\{x_1, x_2, x_3, x_4 \dots \dots |h)}{P(D)} P(h)$$

- ***How many joint probabilities are needed to train?***