MACHINE INTELLIGENCE

UNIT - 3

Bayesian Learning

feedback/corrections: vibha@pesu.pes.edu

VIBHA MASTI

Bayesian Learning

- Assumption: quantities of interest are governed by probabilistic distributions

1. FOUNDATIONS FOR BAYESIAN LEARNING

1.1 Basics

(a) Event

- set of outcomes from a random experiment
- Eg: experiment : tossing a fair coin

  events:

  E_1 = neither heads nor tails
  E_2 = H
  E_3 = T
  E_4 = H or T

  $P(E_1) = 0$
  $P(E_2) = P(E_3) = 1/2$
  $P(E_4) = 1$

(b) Random Variable

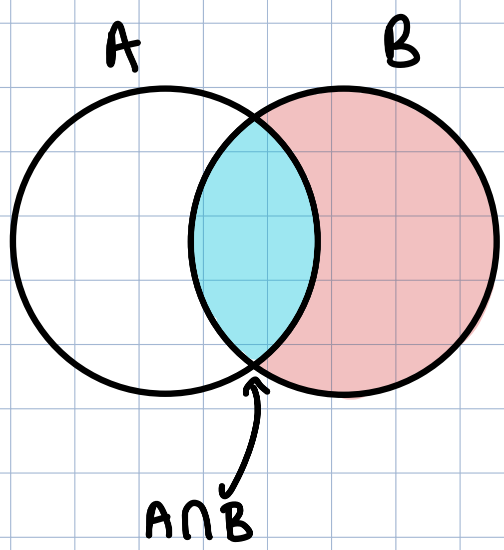
- Numerical value of each outcome in a sample space

- Convention: uppercase letters (X, Y, Z)

(c) Conditional Probability

- $P(A|B)$ = probability of A given that B has occurred

- For independent events A & B



$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

(d) Addition Theorem

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

(e) Independent Events

- If probability of occurrence of one event is unaffected by the occurrence of the other
- Eg: tossing a coin : $P(H)$ independent of previous coin toss

(f) Multiplication Theorem

- Calculate probability of both events A and B

(i) A & B independent

$$P(A \cap B) = P(A)\,P(B)$$

(ii) A & B dependent

$$P(A \cap B) = P(A)\,P(B \mid A)$$

$$P(A \cap B) = P(B)\,P(A \mid B)$$

- For 3 events (dependent)

$$P(A \cap B \cap C) = P(A)\,P(B \mid A)\,P(C \mid A \cap B)$$

- For n events

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)\,P(A_2 \mid A_1) \ldots P(A_n \mid A_{n-1} \cap \ldots \cap A_1)$$

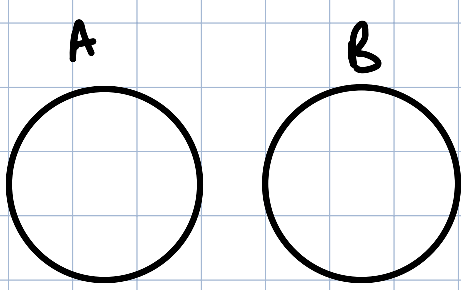
(g) Joint Probability of Independent Events

- Joint probability = $P(A \cap B) = P(A)\,P(B)$
- Mutually independent : if they are all pairwise independent

$$P(A \cap B) = P(A)\,P(B)$$
$$P(B \cap C) = P(B)\,P(C)$$
$$P(C \cap A) = P(C)\,P(A)$$

(h) Mutually Exclusive Events

- If A occurs, then that automatically means B has not occurred
- In other words, $P(A \cap B) = 0$



(i) Collectively Exhaustive Events

- If union of events is SS
- $P(A \cup B) = P(A) + P(B) = 1$ (mutually exclusive also)

1.2 Independent, Identically Distributed RV (IID)

- RV is IID if each RV has the same probability distribution as others and all are mutually independent

- Eg: outcomes of flipping a coin (unfair or fair) $X_1, X_2, \ldots, X_n$

  - independent: outcome of i^{th} flip unaffected by outcome of $i-1^{th}$ flip

  - identical distribution: probability of heads or tails does not change at every flip

1.3 Total Probability Theorem

$$P(B) = \sum_{i=1}^{n} P(A_i) P(B | A_i)$$

where A_i's are mutually exclusive and exhaustive

Proof:

$$\sum_{i=1}^{n} P(A_i) P(B | A_i)$$

$$= \sum_{i=1}^{n} P(A_i) \frac{P(B \cap A_i)}{P(A_i)}$$

$$= \sum_{i=1}^{n} P(B \cap A_i)$$

$$= P(B \cap A_1) + \dots + P(B \cap A_n)$$

$$= P(B \cap (A_1 \cup A_2 \cup \dots \cup A_n))$$

$$= P(B \cap S)$$

sample space S

$$= P(B)$$

1.4 Bayes Theorem

- Hypothesis h
- Training data D

- Terms
  - $P(h)$: prior probability of hypothesis h
  - $P(D)$: prior probability of training data D
  - $P(h|D)$: posterior probability of h given D
  - $P(D|h)$: likelihood of D given h

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

2. MAP and ML Hypothesis

2.1 Maximum A Posteriori (MAP) hypothesis

- Most probable hypothesis h given observed data D
- Such hypothesis is called h_{MAP}
- Terms
  - h: a specific hypothesis
  - H: hypothesis space (set of hypotheses)
- Maximise probability of h given D

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} \; P(h|D)$$

$$= \underset{h \in H}{\operatorname{argmax}} \; \frac{P(D|h)\,P(h)}{P(D)}$$

P(D) is a constant independent of h

$$= \underset{h \in H}{\operatorname{argmax}} \; P(D|h)\,P(h)$$

2.2 Maximum Likelihood Hypothesis

- If every hypothesis in H is equally probable a priori, h_{MAP} is reduced to h_{ML}
- $P(h_i) = P(h_j) \ \forall \ h_i, h_j \in H$

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \ P(D|h)$$

Q: Cancer test

If patient + : test + 98%
If patient − : test − 97%
0.008 of pop has cancer

Patient test +.

What should we diagnose?

Hypothesis space: $\underset{h_1}{\text{patient }+}$, $\underset{h_2}{\text{patient }-}$

Outcome (data) : test + , test −

$P(h_1) = 0.008$
$P(h_2) = 0.992$

$P(+|h_1) = 0.98$ $\quad$ $P(+|h_2) = 0.03$

$P(-|h_1) = 0.02$ $\quad$ $P(-|h_2) = 0.97$

$$P(h_1|+) = \frac{P(+|h_1)\,P(h_1)}{P(+|h_1)P(h_1) + P(+|h_2)\,P(h_2)}$$

$$= \frac{0.98 \times 0.008}{0.98 \times 0.008 + 0.03 \times 0.992}$$

$$= \frac{7.84 \times 10^{-3}}{0.0376} = 0.2085$$

$$P(h_2|+) = \frac{P(+|h_2)\,P(h_2)}{P(+|h_1)P(h_1) + P(+|h_2)\,P(h_2)}$$

$$= \frac{0.03 \times 0.992}{0.98 \times 0.008 + 0.03 \times 0.992}$$

$$= \frac{0.02976}{0.0376} = 0.7915$$

$\therefore \quad h_{MAP} = \text{argmax } P(h|D) = h_2 =$ not cancer

3. Apply Bayesian Learning to Concept Learning

Assumptions

- conjunctive hyp space
- Noise free training data
- Target hyp in hyp space
- All hyp equally probable

Brute Force MAP Learning

Calculate posterior probability of each poss hypothesis $h \in H$ given training data D

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

- $P(h_i) = P(h_j) \;\forall\; h_i, h_j \in H$
- Target concept $\in H$, $\sum_{i=1}^{N} P(h_i) = 1$
- $\therefore\; h_i = \frac{1}{|H|} \;\forall\; h_i \in H$
- $P(D|h)$ is either 0 or 1

$$P(D|h) = \begin{cases} 1, & d_i = h(x_i) \;\forall\; d_i \in D \\ 0 & \text{otherwise} \end{cases}$$

- $VS(H,D)$ = consistent hypotheses

$$P(D) = \sum_{h_i \in H} P(D|h_i)P(h_i)$$

$$P(D) = \sum_{h_i \in VS} 1 \times \frac{1}{|H|} = \frac{|VS|}{|H|}$$

$$P(h|D) = 0 \text{ for } h \notin VS$$

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

$$= \frac{1 \times 1/|H|}{|VS|/|H|} = \frac{1}{|VS|}$$

- Summary

$$P(h|D) = \begin{cases} \frac{1}{|VS|} & \text{if } h \in VS \\ 0 & \text{otherwise} \end{cases}$$

Training

Choose hyp with max posterior

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} \; P(h|D)$$

$$= \underset{h \in H}{\operatorname{argmax}} \; P(D|h)P(h)$$

Testing

Compute $h_{MAP}(x)$ where x is new data point

Drawback

Need to compute all $P(D|h)$ and $P(h)$

BAYES OPTIMAL CLASSIFIER

- Suppose we have 3 hypotheses h_1, h_2 and h_3 in the hypothesis space H
- Suppose we have training data D
- Let $P(h_1|D) = 0.4$, $P(h_2|D) = 0.3$, $P(h_3|D) = 0.3$
- $h_{MAP} = \underset{h_i}{\operatorname{arg\,max}} (P(h_i|D)) = P(h_1|D) = 0.4$

- For a new instance x, suppose we have the predictions $h_1(x)$, $h_2(x)$ and $h_3(x)$

- Most probable classification of x?
  - Not the same as prediction of h_{MAP}

Bayes Optimal Classification

- Most probable output label produced from all possible hypotheses

$$\arg\max_{y} \sum_{h_i \in H} P(y|h_i)\, P(h_i|D)$$

Q: Find most probable classification of x

Hypothesis	Probability of hypothesis P(hi|D)	new instance (x) classified by hi as	Probability of (sign| hi)
h1	0.4	+	P(+|h1)=1 P(-|h1)=0
h2	0.3	-	P(+|h2)=0 P(-|h2)=1
h3	0.3	-	P(+|h3)=0 P(-|h3)=1

$h_{MAP}(x) = + \longrightarrow$ not correct

$$\sum_{h_i \in H} P(+|h_j)\, P(h_i|D) = 1\times0.4 + 0\times0.3 + 0\times0.3 = 0.4$$

$$\sum_{h_i \in H} P(-|h_j)\, P(h_i|D) = 0 \times 0.4 + 1 \times 0.3 + 1 \times 0.3 = 0.6$$

$$\underset{v_j \in V}{\arg\max} = -ve$$

Drawbacks

- Costly — must apply all possible hypotheses on instance
- Size of H is huge

Gibbs Algorithm

- At most ≤ 2 error of bayes optimal
- Uses Gibbs Sampling

Algorithm

1. Choose random $h \in H$ according to $P(h|D)$ posterior probability
2. Use one hypothesis from the distribution to predict classification of new instance

4. Naive Bayes Classifier

1. Gaussian

   - features follow ND
   - continuous data
   - classification

2. Multinomial

   - discrete counts
   - text classifier

3. Binomial

   - feature vectors binary
   - 'bag of words' text classifier (occurs or not)

Known

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)}$$

$$P(Y|X) \propto P(X|Y)\,P(Y)$$

- X is input feature - can be vector of n features

$$P(Y|X) \propto P(X_1, X_2, \ldots, X_n | Y)\ P(Y)$$

- Joint probability of $X_1, X_2, \ldots, X_n$ is diff to calculate

$$P(Y|X) \propto P(X_1|Y)P(X_2|X_1,Y)\ldots P(X_n|X_{n-1}\ldots X_1,Y)P(Y)$$

Assumptions

- Individual features are independent given an observation

$$P(Y|X) \propto P(X_1|Y)P(X_2|Y)\ldots P(X_n|Y)\ P(Y)$$

$$Y^{new} = \arg\max_{y_k} P(Y=y_k) \prod_i P(X_i^{new} | Y=y_k)$$

$$Y^{new} = \arg\max_{y_k} P(Y = y_k) \prod_i P(x_i^{new}|Y = y_k)$$

$$Y^{new} = V_{MAP} = V_{NB}$$

Q: PlayTennis

Outlook	Temp	Humidity	Windy	Play tennis
Sunny	High	High	Weak	No
Sunny	High	High	Strong	No
Overcast	High	High	Weak	Yes
Rainy	Medium	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Medium	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Medium	Normal	Weak	Yes
Sunny	Medium	Normal	Strong	Yes
Overcast	Medium	High	Strong	Yes
Overcast	High	Normal	Weak	Yes
Rainy	Medium	High	Strong	No

classify:

(sunny, cool, high, strong)

(Recall: same eg in ID3 decision trees)

Naive Bayes:

$$V_{MAP} = \underset{v_k \in V}{\operatorname{argmax}}\ P(v_k) \prod_i P(a_i \mid v_k)$$

$P(yes) = \frac{9}{14}$ $P(no) = \frac{5}{14}$

$P(sunny|yes) = \frac{2}{9}$ $P(sunny|no) = \frac{3}{5}$

$P(cool|yes) = \frac{3}{9}$ $P(cool|no) = \frac{1}{5}$

$P(high|yes) = \frac{3}{9}$ $P(high|no) = \frac{4}{5}$

$P(strong|yes) = \frac{3}{9}$ $P(strong|no) = \frac{3}{5}$

$$P(yes|o/p) \propto \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = \frac{1}{189}$$

$$= 5.291 \times 10^{-3}$$

$$P(no|o/p) \propto \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = \frac{18}{875}$$

$$= 0.0206$$

$\therefore \; V_{MAP} = \text{PlayTennis} = \text{NO}$

Smoothing

- If no training instance with target v_i have att a_i

$$P(a_i | v_i) = 0$$

- Solution: add virtual rows

- Add m total rows and assume a_i has a proportion of p in those m rows

$$\hat{P}(a_i | v_j) = \frac{n_c + mp}{n + m}$$

- p = prior estimate for $\hat{P}(a_i | v_j)$
  n = training eg for which $v = v_j$
  n_c = training eg for which $v = v_j$ and $a = a_i$
  m = no of virtual examples

Add-One Smoothing

- Assume $mp = 1$

- Laplace Smoothing

Text Classification

D1 : Inspiring Address	Campaigning	
D2 : Aggressive Speech	Campaigning	
D3 : Cowardly act	Law & Order	
D4 : Spoke coward	Campaigning	
D5 : threatening speech	Law & Order	
D6 : Act aggression	Law & Order	
D7 : Aggression threatening address	?	

Speech = spoke

aggression = aggressive

Laplace smoothing

Word	Campaigning	Law and Order
Inspiring	1 +1	1
Address	1 +1	1
Aggressive	1 +1	1 +1
Speech	2 +1	1 +1
Cowardly	1	1 +1
Act	1	2 +1
Coward	1 +1	1
Threatening	1	1 +1
	6 +8	6 +8

$$P(\text{aggression} \mid c) \times P(\text{threat} \mid c) \times P(\text{add} \mid c) \times P(c)$$

$$= \frac{2}{14} \times \frac{1}{14} \times \frac{2}{14} \times \frac{1}{2} = \frac{1}{1372} = 7.29 \times 10^{-4}$$

$P(\text{aggression} \mid \ell) \times P(\text{threat} \mid \ell) \times P(\text{add} \mid \ell) \times P(\ell)$

$$= \frac{2}{14} \times \frac{2}{14} \times \frac{1}{14} \times \frac{1}{2} = \frac{1}{1372} = 7.29 \times 10^{-4}$$

$\therefore$ both have same probability
randomly choose 1

Advancements

- Removing stop words
- Stemming
  - election = elected
- Using n-grams