Introduction to Artificial Intelligence

COMP307
Reasoning Under Uncertainty 2: Naïve Bayes Classifier

Yi Mei

yi.mei@ecs.vuw.ac.nz

Outline

- Rules from last lecture
- Bayes Rule
- Naive Bayes Classifier
  - Assumption
  - Deal with zero count
- Summary

Important Rules

- The product rule:
  - P(A, B) = P(B) * P(A | B) = P(A) * P(B | A)

- The sum rule
  - $P(X = x) = \sum_{y \in \Omega} P(X = x, Y = y)$

- The normalisation rule
  - $\sum_x P(X = x) = 1$
  - $\sum_x P(X = x \mid Y = y) = 1$

- Independence
  - P(A | B) = P(A)
  - P(B | A) = P(B)
  - P(A, B) = P(A) * P(B)

Bayes Rules

- The product rule:
  - P(A, B) = $\boxed{P(B) * P(A \mid B) = P(A) * P(B \mid A)}$
- Transform to Bayes Rule
  - $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$
- More variables
  - $P(Y \mid X_1, \ldots, X_n) = \frac{P(X_1,\ldots,X_n \mid Y)P(Y)}{P(X_1,\ldots,X_n)}$

Thomas Bayes (/ˈbeɪz/; c. 1701 – 7 April 1761)

Interpretation of Bayes Rules

- Proposition A and evidence B
  - P(A l B): the posterior degree of belief in A, given evidence B
  - P(B l A): if A is true, the degree of belief that the evidence B is shown
  - P(A): the prior degree of belief in A, without any evidence
  - P(B): the degree of belief that evidence B is shown

- $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$

- For calculating P(A l B), need to estimate P(B l A), P(A) and P(B)

Example: Medical Test

- You are worried about having a rare cancer.
- The cancer is very rare, occurring in only one of every 10,000 people.
- You go with the test, which has 99% accuracy (if you have the disease, it shows that you do with 99% probability, and if you don't have the disease, it shows that you do not with 99% probability).
- If your test results come back positive, what are your chances that you actually have the disease?

- (a) 99% (b) 90% (c) 10% (d) 1%

Example Training Dataset

Applicant	Job	Deposit	Family	Class
1	true	low	single	Approve
2	true	low	couple	Approve
3	true	high	single	Approve
4	true	high	single	Approve
5	false	high	couple	Approve
6	true	low	couple	Decline
7	false	low	couple	Decline
8	true	low	children	Decline
9	false	low	single	Decline
10	false	high	children	Decline

Example Classification Task

- Determine whether to approve a mortgage application, given data/features about the client:
  – Whether they have a job (true or false)
  – The level of their deposit (low or high)
  – Their family status (single, couple[but no kids], children)

- Classification: either Approve or Decline

- Given a set of data about past clients and the classification by the Bank's experts

- Construct a classifier that will output the right answer (class) when given a new (unseen) client (instance)

Bayes Rules for Classification

- Very simple probability-based technique
- Computes P(class l instance data) for each class, and choose the class with the highest probability.

- **Problem: Hard to measure** P(class l data)
- e.g. P(Decline l Job=true, Dep=high, Fam=children)
- Needs lots of examples of (Job=true & Dep=high & Fam=children)
- Then count the fraction that are Decline.
- Usually do NOT have enough data

- **Use Bayes Rules**

$$P(Decline|Job = true, Dep = high, Fam = children)$$

$$= \frac{P(Deline) * P(Job = true, Dep = high, Fam = children|Deline)}{P(Job = true, Dep = high, Fam = children)}$$

Naïve Bayes

- Why this is better?
  - No better if just like this
  - We still need a lot of data to have a comprehensive estimation of the multivariate distribution (Job, Dep, Fam) and (Job, Dep, Fam I Decline)
  - But what if the features are independent?

$$P(Decline|Job = true, Dep = high, Fam = children)$$
$$= \frac{P(Deline) * P(Job = true, Dep = high, Fam = children|Deline)}{P(Job = true, Dep = high, Fam = children)}$$

- A naïve Bayes approach assumes that the features are conditionally independent
  - If A and B are conditional independent on C, then P(A, B I C) = P(A I C) * P(B I C)
  - More variables $P(X_1, \ldots, X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$
- Example:

$$P(Job = true, Dep = high, Fam = children|Decline)$$
$$= P(Job = true|Decline) * P(Dep = high|Decline) * P(Fam = children|Decline)$$

- There is usually enough data for the univariate distributions

Computing Probabilities: Example

Class	Approve	Decline
Total	5	5
Job = true	4	2
Job = false	1	3
Dep = low	2	4
Dep = high	3	1
Fam = single	3	1
Fam = couple	2	2
Fam = children	0	2

	Approve	Decline
P(Class)	5/10	5/10
P(Job = true | Class)	4/5	2/5
P(Job = false | Class)	1/5	3/5
P(Dep = low | Class)	2/5	4/5
P(Dep = high | Class)	3/5	1/5
P(Fam = single | Class)	3/5	1/5
P(Fam = couple | Class)	2/5	2/5
P(Fam = children | Class)	0/5	2/5

Using Naïve Bayes Classifier

- Classify a new case: (Job=true, Dep=high, Fam=children)
- Calculate P(Decline I Job=true, Dep=high, Fam=children)
- Calculate P(Approve I Job=true, Dep=high, Fam=children)
- See which probability is higher

$$P(\boxed{Decline}|Job = true, Dep = high, Fam = children)$$
$$= \frac{P(Deline) * P(Job = true, Dep = high, Fam = children|Decline)}{P(Job = true, Dep = high, Fam = children)}$$
$$= \frac{P(Deline) * P(Job = true|Decline) * P(Dep = high|Decline) * P(Fam = children|Decline)}{P(Job = true, Dep = high, Fam = children)}$$
$$= \frac{0.4 \times 0.2 \times 0.4 \times 0.5}{P(Job = true, Dep = high, Fam = children)}$$
$$= \frac{0.016}{P(Job = true, Dep = high, Fam = children)}$$

Using Naïve Bayes Classifier

- Classify a new case: (Job=true, Dep=high, Fam=children)
- Calculate P(Decline I Job=true, Dep=high, Fam=children)
- Calculate P(Approve I Job=true, Dep=high, Fam=children)
- See which probability is higher

$$P(Approve|Job = true, Dep = high, Fam = children)$$
$$= \frac{P(Approve) * P(Job = true, Dep = high, Fam = children|Approve)}{P(Job = true, Dep = high, Fam = children)}$$
$$= \frac{P(Approve) * P(Job = true|Approve) * P(Dep = high|Approve) * P(Fam = children|Approve)}{P(Job = true, Dep = high, Fam = children)}$$
$$= \frac{0.8 \times 0.6 \times 0 \times 0.5}{P(Job = true, Dep = high, Fam = children)}$$
$$= \frac{0}{P(Job = true, Dep = high, Fam = children)}$$

- Denominator does not need to calculate (the same for all the classes)
- Probability of Approve = 0? Just because (Fam = children) has never occurred for Approve. Need to deal with **zero occurrence**

Computing Probabilities: Example

Class	Approve	Decline
Total	5	5
Job = true	4	2
Job = false	1	3
Dep = low	2	4
Dep = high	3	1
Fam = single	3	1
Fam = couple	2	2
Fam = children	0	2

	Approve	Decline
P(Class)	5/10	5/10
P(Job = true | Class)	4/5	2/5
P(Job = false | Class)	1/5	3/5
P(Dep = low | Class)	2/5	4/5
P(Dep = high | Class)	3/5	1/5
P(Fam = single | Class)	3/5	1/5
P(Fam = couple | Class)	2/5	2/5
P(Fam = children | Class)	0/5	2/5

Dealing with Zero Occurrence

- Initialise the table to contain small constant, e.g. 1
- This is not quite sound, but reasonable in practice

Class	Approve	Decline
Total	6	6
Job = true	5	3
Job = false	2	4
Dep = low	3	5
Dep = high	4	2
Fam = single	4	2
Fam = couple	3	3
Fam = children	1	3

	Approve	Decline
P(Class)	6/12	6/12
P(Job = true | Class)	5/7	3/7
P(Job = false | Class)	2/7	4/7
P(Dep = low | Class)	3/7	5/7
P(Dep = high | Class)	4/7	2/7
P(Fam = single | Class)	4/8	2/8
P(Fam = couple | Class)	3/8	3/8
P(Fam = children | Class)	1/8	3/8

- Denominator of Job and Dep is 7 (e.g. (Job = true) = 5, (Job = false) = 2, 5+2=7
- Denominator of Fam is 8, 4+3+1=8

Using Naïve Bayes Classifier

$$P(Decline|Job = true, Dep = high, Fam = children)$$
$$= \frac{P(Decline) * P(Job = true, Dep = high, Fam = children|Decline)}{P(Job = true, Dep = high, Fam = children)}$$
$$= \frac{P(Decline) * P(Job = true|Decline) * P(Dep = high|Decline) * P(Fam = children|Decline)}{P(Job = true, Dep = high, Fam = children)}$$
$$= \frac{3/7 \times 2/7 \times 3/8 \times 1/2}{P(Job = true, Dep = high, Fam = children)}$$
$$= \frac{0.0230}{P(Job = true, Dep = high, Fam = children)}$$

$$P(Approve|Job = true, Dep = high, Fam = children)$$
$$= \frac{P(Approve) * P(Job = true, Dep = high, Fam = children|Approve)}{P(Job = true, Dep = high, Fam = children)}$$
$$= \frac{P(Approve) * P(Job = true|Approve) * P(Dep = high|Approve) * P(Fam = children|Approve)}{P(Job = true, Dep = high, Fam = children)}$$
$$= \frac{5/7 \times 4/7 \times 1/8 \times 1/2}{P(Job = true, Dep = high, Fam = children)}$$
$$= \frac{0.0255}{P(Job = true, Dep = high, Fam = children)}$$

Summary

- Bayes rule:
  - $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$
  - $P(Y \mid X_1, \dots, X_n) = \frac{P(X_1,\dots,X_n \mid Y)P(Y)}{P(X_1,\dots,X_n)}$
- In classification, Y is the class label, X1, …, Xn are features. The probability of an instance belonging to a class is

$$P(Y \mid X_1, \dots, X_n) = \frac{P(X_1,\dots,X_n \mid Y)P(Y)}{P(X_1,\dots,X_n)}$$

- Calculate $P(Y \mid X_1, \dots, X_n)$ for each class, and predict as the class with the highest conditional probability
  - The denominator $P(X_1, \dots, X_n)$ can be ignored, as it is the same for all the classes
  - $P(X_1, \dots, X_n \mid Y)$ is still hard to estimate (high-dimensional multivariate distribution)
- Assume conditional independence (Naïve Bayes)
  - $P(X_1, \dots, X_n \mid Y) = P(X_1|Y) \times P(X_2|Y) \times \cdots \times P(X_n|Y)$
  - Easy to estimate the univariate distribution