SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Bayes & Uncertainty III

PROF LIM KWAN HUI

50.021 Artificial Intelligence

Outline & Objectives

- Recap on statistical concepts such as product rule, chain rule, conditional independence, Bayes rules
- Able to represent a problem in terms of a Bayesian network and its corresponding conditional probability table
- Learn about how Bayes net can be used in various scenarios
- Learn about the Naïve Bayes Classifier and its application to text

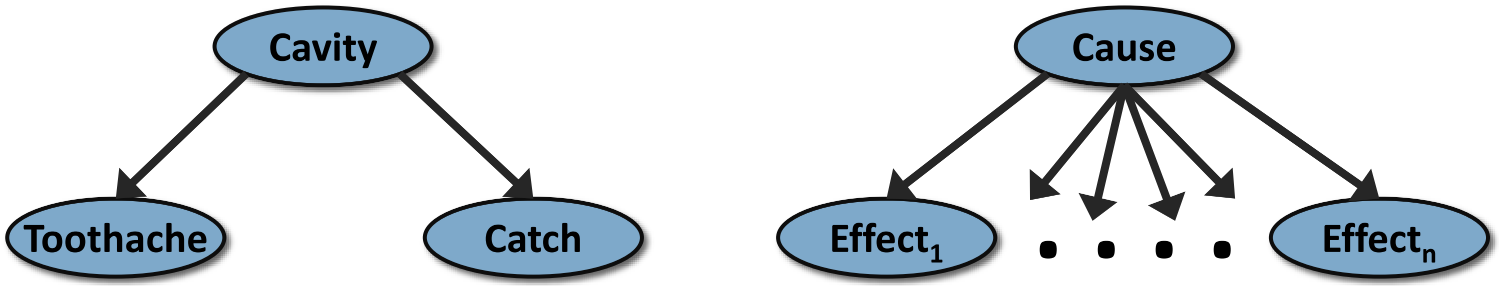
Bayes Rule and conditional independence

- $P(\text{Cavity} \mid \text{toothache, catch})$

  $= \alpha\, P(\text{toothache, catch} \mid \text{Cavity})\, P(\text{Cavity})$

  $= \alpha\, P(\text{toothache} \mid \text{Cavity})\, P(\text{catch} \mid \text{Cavity})\, P(\text{Cavity})$

- This is an example of a naive Bayes model:
  - $P(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) = P(\text{Cause}) \prod_i P(\text{Effect}_i \mid \text{Cause})$



- Total number of parameters is linear in n

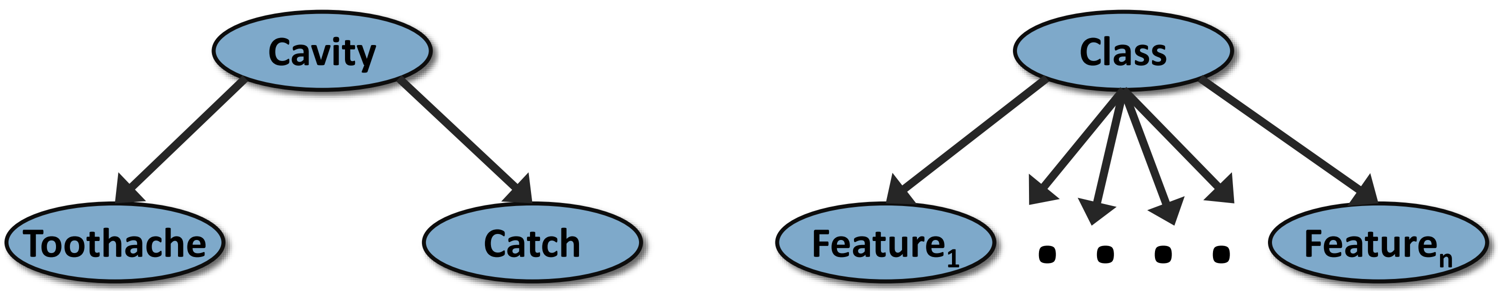
Naïve Bayes Classifier

- P(Cavity | toothache, catch)

  $= \alpha$ P(toothache, catch | Cavity) P(Cavity)

  $= \alpha$ P(toothache|Cavity) P(catch|Cavity) P(Cavity)

- Similar to finding the most likely class given a set of features, e.g.,
  - P(cavity=true | toothache, catch) or P(cavity=false | toothache, catch)



Bayes for Classification

- Given the following:
  - An observation o represented by a feature set $X_o = \{x_1, x_2, \ldots, x_m\}$
  - A fixed set of classes $Y = \{y_1, y_2, \ldots, y_n\}$

- We are interested to classify the class Y that observation o belongs to given its feature set X_o

$$y_{MAP} = \text{argmax}\, P(Y \mid X_o)$$

Bayes for Classification

- We are interested to classify the class *Y* that observation *o* belongs to given its feature set X_o.
- Applying Bayes Theorem, we have:

$$y_{MAP} = \text{argmax}\, P(Y \mid X_o)$$

$$y_{MAP} = \text{argmax} \frac{P(X_o \mid Y)\, P(Y)}{P(X_o)}$$

$$y_{MAP} = \text{argmax}\, P(X_o \mid Y)\, P(Y)$$

$$y_{MAP} = \text{argmax}\, P(x_1, x_2, \dots, x_m \mid Y)\, P(Y)$$

Example: Tax Avoidance

- Given an observation *o* with feature set X_o = *{Industry=IT, Status=Married, Income=low}*, how to estimate class *Y*?

$$y_{MAP} = \text{argmax}\, P(Y \mid X_o)$$

- Find probabilities by counting:
  - $P(Y) = N_c/N$
    - E.g., P(AvoidTax=Yes) = 3/10
  - $P(X_i \mid Y_k) = |X_{ik}| / N_c$
    - E.g., P(Industry=Sales | AvoidTax=Yes) = 2/3

ID	Industry	Marital Status	Income Level	Avoid Tax
1	IT	Single	High	***No***
2	Sales	Married	Medium	***No***
3	Sales	Single	Low	***No***
4	IT	Married	High	***No***
5	Sales	Divorced	Low	***Yes***
6	Sales	Married	Low	***No***
7	IT	Divorced	High	***No***
8	IT	Married	Medium	***Yes***
9	Sales	Married	Low	***No***
10	Sales	Single	Medium	***Yes***

Example: Tax Avoidance

- Given an observation *o* with feature set X_o = *{Industry=IT, Status=Married, Income=low}*, how to estimate class *Y*?

  - P(Y=Yes | X) = P(X | Y=Yes) P (Y=Yes)
    = P(Industry=IT | Yes) x P(Status=Married | Yes) x P(Income=Low | Yes) x P(Yes)
    = 1/3 x 1/3 x 1/3 x 3/10

  - P(Y=No | X) = P(X | Y=No) P (Y=No)
    = P(Industry=IT | No) x P(Status=Married | No) x P(Income=Low | No) x P(No)
    = 3/7 x 4/7 x 3/7 x 7/10

ID	Industry	Marital Status	Income Level	Avoid Tax
1	IT	Single	High	***No***
2	Sales	Married	Medium	***No***
3	Sales	Single	Low	***No***
4	IT	Married	High	***No***
5	Sales	Divorced	Low	***Yes***
6	Sales	Married	Low	***No***
7	IT	Divorced	High	***No***
8	IT	Married	Medium	***Yes***
9	Sales	Married	Low	***No***
10	Sales	Single	Medium	***Yes***

Naïve Bayes on Text

- Given a document *d*, we want to find the most likely class *c*:

$$c_{NB} = \underset{c_i \in C}{\operatorname{argmax}} P(c_i \mid d)$$

Naïve Bayes on Text

- Given a document *d*, we want to find the most likely class *c*:

$$c_{NB} = \underset{c_i \in C}{\operatorname{argmax}} P(c_i | d)$$

$$c_{NB} = \operatorname{argmax} \frac{P(d | c_i) \, P(c_i)}{P(d)}$$

Bayes Rule

Naïve Bayes on Text

- Given a document *d*, we want to find the most likely class *c*:

$$c_{NB} = \underset{c_i \in C}{\operatorname{argmax}} P(c_i | d)$$

$$c_{NB} = \operatorname{argmax} \frac{P(d | c_i) P(c_i)}{P(d)}$$

$$c_{NB} = \operatorname{argmax} P(d|c_i) P(c_i)$$

Normalization

Naïve Bayes on Text

- Given a document *d*, we want to find the most likely class *c*:

$$c_{NB} = \underset{c_i \in C}{\operatorname{argmax}} P(c_i | d)$$

$$c_{NB} = \operatorname{argmax} \frac{P(d | c_i)\, P(c_i)}{P(d)}$$

$$c_{NB} = \operatorname{argmax} P(d|c_i)\, P(c_i)$$

$$c_{NB} = \operatorname{argmax} P(w_1, w_2, \ldots, w_m | c_i)\, P(c_i)$$

Document as words

Naïve Bayes on Text

- Given a document *d*, we want to find the most likely class *c*:

$$c_{NB} = \underset{c_i \in C}{\operatorname{argmax}} P(c_i | d)$$

$$c_{NB} = \operatorname{argmax} \frac{P(d | c_i) P(c_i)}{P(d)}$$

$$c_{NB} = \operatorname{argmax} P(d | c_i) P(c_i)$$

$$c_{NB} = \operatorname{argmax} P(w_1, w_2, \ldots, w_m | c_i) P(c_i)$$

Conditional Independence

$$c_{NB} = \operatorname{argmax} P(w_1 | c_i) P(w_2 | c_i) \ldots P(w_m | c_i) P(c_i)$$

Naïve Bayes on Text

- Given a document *d*, we want to find the most likely class *c*:

$$c_{NB} = \underset{c_i \in C}{\text{argmax}}\, P(c_i|\, d)$$

$$c_{NB} = \text{argmax} \frac{P(d\,|c_i)\, P(c_i)}{P(d)}$$

$$c_{NB} = \text{argmax}\, P(d|c_i)\, P(c_i)$$

$$c_{NB} = \text{argmax}\, P(w_1, w_2, \dots, w_m\,|c_i)\, P(c_i)$$

$$c_{NB} = \text{argmax}\, P(w_1\,|c_i)\, P(w_2\,|c_i) \dots P(w_m\,|c_i)\, P(c_i)$$

$$c_{NB} = \underset{c_i \in C}{\text{argmax}}\, P(c_i) \prod_{w_j \in W} P(w_j \mid c_i)$$

Naïve Bayes on Text

- How do we estimate the different probabilities?

$$c_{NB} = \underset{c_i \in C}{\operatorname{argmax}} P(c_i) \prod_{w_j \in W} P\left(w_j \mid c_i\right)$$

- For a set of documents *D,* vocabulary of words *W*, and classes C,

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

$$P(w_1 \mid c_i) = \frac{count(w_1, c_i)}{\sum_{w_j \in W} count(w_j, c_i)}$$

Naïve Bayes on Text

- How do we estimate the different probabilities?

$$c_{NB} = \underset{c_i \in C}{\operatorname{argmax}} P(c_i) \prod_{w_j \in W} P(w_j \mid c_i)$$

- For a set of documents *D,* vocabulary of words *W*, and classes C,

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

Proportion of class c_i in the entire dataset

$$P(w_1 \mid c_i) = \frac{count(w_1, c_i)}{\sum_{w_j \in W} count(w_j, c_i)}$$

 # times word w_1 is used in all documents of class c_i

Total words in all documents of class c_i

Naïve Bayes on Text

- How do we use this to test if an email is a scam or real?

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

$$\boldsymbol{P(w_1 \mid c_i) = \frac{count(w_1, c_i)}{\sum_{w_j \in W} count(w_j, c_i)}}$$

ID	Email Text	Class
1	iphone, free, password	Scam
2	job, easy, password	Scam
3	lunch, restaurant, discount, iphone	Real
4	restaurant, reservation, discount	Real
5	iphone, discount, reservation	?

$$\begin{aligned} &P(Real \mid iphone, discount, reservation) \\ &\quad = P(iphone, discount, reservation \mid Real)\ P(Real) \\ &\quad = P(iphone|Real)\ P(discount|Real)\ P(reservation|Real)\ P(Real) \\ &\quad = (1/7) \times (2/7) \times (1/7) \times (2/4) = 0.0029 \end{aligned}$$

Exercise: Naïve Bayes on Text

- How do we use this to test if an email is a scam or real?

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

$$P(w_1 \mid c_i) = \frac{count(w_1, c_i)}{\sum_{w_j \in W} count(w_j, c_i)}$$

ID	Email Text	Class
1	iphone, free, password	Scam
2	job, easy, password	Scam
3	lunch, restaurant, discount, iphone	Real
4	restaurant, reservation, discount	Real
5	iphone, discount, reservation	?

$$P(Scam \mid iphone, discount, reservation) = ?$$

Which is the most likely class?

Exercise: Naïve Bayes on Text

- How do we use this to test if an email is a scam or real?

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

$$\boldsymbol{P(w_1 \mid c_i) = \frac{count(w_1, c_i)}{\sum_{w_j \in W} count(w_j, c_i)}}$$

ID	Email Text	Class
1	iphone, free, password	Scam
2	job, easy, password	Scam
3	lunch, restaurant, discount, iphone	Real
4	restaurant, reservation, discount	Real
5	iphone, discount, reservation	?

Which is the most likely class?

$$P(Real \mid i, d, r) = 0.0029$$
$$P(Scam \mid i, d, r) = 0$$

$$\begin{aligned} &P(Scam \mid iphone, discount, reservation) \\ &\quad = P(iphone, discount, reservation \mid Scam)\; P(Scam) \\ &\quad = P(iphone \mid Scam)\; P(discount \mid Scam)\; P(reservation \mid Scam)\; P(Scam) \\ &\quad = (1/6) \times (0/6) \times (0/6) \times (2/4) = 0 \end{aligned}$$

What is the issue here?

Laplace Smoothing

- Problem with new words, or unseen words for a specific class
  - E.g., Consider a new word w_1, which gives $P(w_1|c) = \frac{\mathrm{count(w_1,c_i)}}{\sum \mathrm{count(w,c_i)}} = 0$
  - The probability $P(w_1 \mid c)$ = 0 affects the entire equation

$$c_{NB} = \operatorname{argmax} P(c \mid d)$$

$$\vdots$$

$$c_{NB} = \operatorname{argmax} P(w_1 \mid c)\, P(w_2 \mid c)\, P(w_3 \mid c)\, P(w_4 \mid c)\, P(c)$$

Usually α = 1

- Laplace Smoothing: Add a constant α to our counts

$$P(w_1 \mid c_i) = \frac{count(w_1, c_i) + \alpha}{\sum_{w_j \in W} \left(count(w_j, c_i) + \alpha\right)} = \frac{count(w_1, c_i) + \alpha}{\sum_{w_j \in W} \left(count(w_j, c_i)\right) + \alpha|W|}$$

Exercise: Naïve Bayes on Text

- How do we use this to test if an email is a scam or real?

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

Laplace Smoothing

$$\boldsymbol{P(w_1 \mid c_i) = \frac{count(w_1, c_i) + 1}{\sum_{w_j \in W} count(w_j, c_i) + |W|}}$$

ID	Email Text	Class
1	iphone, free, password	Scam
2	job, easy, password	Scam
3	lunch, restaurant, discount, iphone	Real
4	restaurant, reservation, discount	Real
5	iphone, discount, reservation	?

$$P(Real \mid iphone, discount, reservation) = ?$$

Exercise: Naïve Bayes on Text

- How do we use this to test if an email is a scam or real?

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

Laplace Smoothing

$$\boldsymbol{P(w_1 \mid c_i) = \frac{count(w_1, c_i) + 1}{\sum_{w_j \in W} count(w_j, c_i) + |W|}}$$

ID	Email Text	Class
1	iphone, free, password	Scam
2	job, easy, password	Scam
3	lunch, restaurant, discount, iphone	Real
4	restaurant, reservation, discount	Real
5	iphone, discount, reservation	?

$$P(Real \mid iphone, discount, reservation)$$
$$= P(iphone, discount, reservation \mid Real)\ P(Real)$$
$$= P(iphone|Real)\ P(discount|Real)\ P(reservation|Real)\ P(Real)$$
$$= (1+1)/(7+9) \times (2+1)/(7+9) \times\ (1+1)/(7+9) \times\ (2/4) \approx 0.0015$$

Exercise: Naïve Bayes on Text

- How do we use this to test if an email is a scam or real?

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

Laplace Smoothing

$$P(w_1 \mid c_i) = \frac{count(w_1, c_i) + 1}{\sum_{w_j \in W} count(w_j, c_i) + |W|}$$

ID	Email Text	Class
1	iphone, free, password	Scam
2	job, easy, password	Scam
3	lunch, restaurant, discount, iphone	Real
4	restaurant, reservation, discount	Real
5	iphone, discount, reservation	?

$$P(Scam \mid iphone, discount, reservation) \\ = ?$$

Exercise: Naïve Bayes on Text

- How do we use this to test if an email is a scam or real?

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

Laplace Smoothing

$$P(w_1 \mid c_i) = \frac{count(w_1, c_i) + 1}{\sum_{w_j \in W} count(w_j, c_i) + |W|}$$

ID	Email Text	Class
1	iphone, free, password	Scam
2	job, easy, password	Scam
3	lunch, restaurant, discount, iphone	Real
4	restaurant, reservation, discount	Real
5	iphone, discount, reservation	?

$$\begin{aligned} &P(Scam \mid iphone, discount, reservation) \\ &\quad = P(iphone, discount, reservation \mid Scam)\, P(Scam) \\ &\quad = P(iphone|Scam)\, P(discount|Scam)\, P(reservation|Scam)\, P(Scam) \\ &\quad = (1+1)/(6+9) \times (0+1)/(6+9) \times (0+1)/(6+9) \times (2/4) \approx 0.0003 \end{aligned}$$

Naïve Bayes on Text

- Naïve Bayes is a decent baseline classifier but has its own limitations

- Assumes there is conditional independence (given a class)
  - $P(w_1, w_2, w_3, w_4 | \, c) = P(w_1 \mid c)\, P(w_2 \mid c)\, P(w_3 \mid c)\, P(w_4 \mid c)$
  - Is this always true?

- Assumes that order does not matter
  - “I like burger but dislike fruits” and “I like fruits but dislike burgers”
  - Does the NB classifier treat the above two sentences differently?

Summary

- Recap on statistical concepts such as product rule, chain rule, conditional independence, Bayes rules
- Problem representation in terms of a Bayesian network and its corresponding conditional probability table
- Learn about how Bayes net can be used in various scenarios
- Learn about the Naïve Bayes Classifier and its application to text