UCIrvine | The Paul Merage School of Business

2022.11.24

BANA 273 Session 4

Ex ayes

Prof. Vibs Abhishek

The Paul Merage School of Business

University of California, Irvine

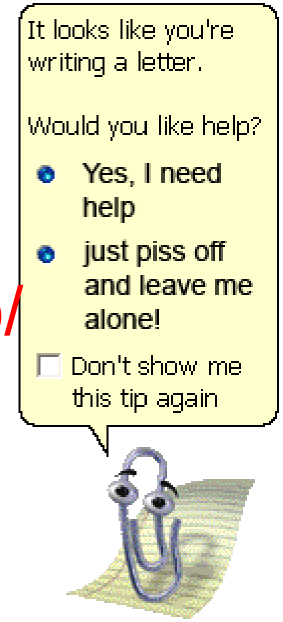
Agenda

- Classification using Exact Bayes & Naïve Bayes
- Reminders
  - Assignment 1 due on Canvas
  - Assignmen
  - Project proposal (1 para) d ck Canvas for all due dates)
  - Project guidelines posted to Canvas (Announcements page)

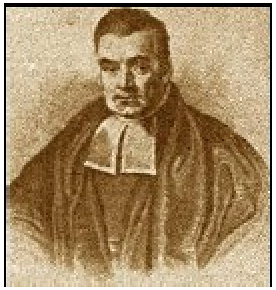
Big Picture View of Course Progress

- Databases, Data Warehousing, SQL
- RFM & Pivot Tables
- Classificatio
  - Bayesian (
  - Decision Tree (ID3)
- Association Rules
  - Apriori
- Clustering
  - K Means

A classic: Microsoft's Paperclip



Exact Bayes



Thomas Bayes

For each record to be classified:

1. Find all other records just like it (i.e. where all the predict
2. Determine elong to and which class is more prev
3. Assign that class to the new record

Predict class attribute "Play" using Exact Bayes

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Nor	[illegible]	[illegible]
Overcast	Cool	Nor	[illegible]	[illegible]
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

[illegible]	p.	Humidity	Windy	Play
[illegible]	t	High	False	?

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Notes

- Bayesian classifier works best with categorical attributes
  - Unlikely to merical variables
- Numerical a d and converted to categorical at
- When the number of attributes is large (say 20), it becomes hard to find exact matches

Exact Bayes – Cutoff Probability Method

- Establish a cutoff probability for the class of interest above which we consider that a record belongs to that class
- Find all the t e the new record
- Determine the probability records belong to the class of interest
- If that probability is above the cutoff probability, assign the new record to the class of interest

Example – Exact Bayes

	Sunny	Overcast	Rainy	Total
Play=Yes	2	3	2	7
Play=No	3	9	4	16
Total	5			23

P(Play=Yes | outlook=sunny) =

P(Play=Yes | outlook=overcast) = 25%

P(Play=Yes | outlook=rainy) = 33%

Conclusion: No matter what the outlook, predict Play = No

Cutoff probability method: Specify cutoff probability p

If Probability(Play=Yes | outlook = ?) > p then predict Play = Yes

Suppose p = 37%

Under what outlook would we forecast play = Yes?

UCIrvine | The Paul Merage School of Business

2022.11.24

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Prof. Vibs

The Paul Merage School of Business

University of California, Irvine

Conditional Probability

- Rules of probability: $P(A_1, \ldots, A_p \mid B=1) = P(A_1|B=1) * P(A_2|B=1) * \ldots * P(A_p|B=1)$

This is correct only if the events A_1 __________

- Let's start by assuming that they are, then:

P(Outlook=Sunny, Temp=High| Play=Yes) =

P(Outlook=Sunny| Play=Yes) * P(Temp=High| Play=Yes)

Apply Bayes’ Rule

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

B = the event “Play=Yes”

A = the event “ p = High”

$$\mathrm{P}(Play = \mathrm{Yes} \mid Outlook = \mathrm{sunny}, Temp = \mathrm{High}) =$$

$$= \frac{\mathrm{P}(Outlook = \mathrm{sunny}, Temp = \mathrm{High} \mid Play = \mathrm{Yes}) \cdot \mathrm{P}(Play = \mathrm{Yes})}{\mathrm{P}(Outlook = \mathrm{sunny}, Temp = \mathrm{High})}$$

$$= \frac{\mathrm{P}(Outlook = \mathrm{sunny} \mid Play = \mathrm{Yes}) \cdot \mathrm{P}(Temp = \mathrm{High} \mid \mathrm{Play} = \mathrm{Yes}) \cdot \mathrm{P}(Play = \mathrm{Yes})}{\mathrm{P}(Outlook = \mathrm{sunny}, Temp = \mathrm{High})}$$

Meaning of conditional independence

- *P(outlook=sunny,Temp=High | Yes)* with

  *P(outlook=sunny|Yes) * P(Temp=High | Yes)*

- This means [illegible] onditional independence between *ou*[illegible] *emp*[illegible]

- If the conditional dependence is not extreme, it will work reasonably well

Probabilities for weather data

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool										

		midity	Windy	Play
		h	False	No
Sunny			True	No
Overcast			False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Terminology

- Frequency Chart also called contingency table (on previous slide)
- Probability
- Create the c xcel – Pivot Table
- How to open ARFF file in Excel?
  - Launch Excel, Open File, Delimited, comma delimited
- Can also use SQL to compute entries in table.

Probabilities for weather data

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool										

- A new day:

Outlook	Temp.		ndy	Play
Sunny	Cool		rue	?

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Tem
Sunny	Coo

← *Evidence E*

Pr [yes|E]=Pr [Outlook=Sunny|yes]×Pr [Te ool|yes]
×Pr [Humidity=High|yes]×Pr [Windy=True|yes]×Pr [yes]/Pr [E]

$$= \frac{\frac{2}{9}\cdot\frac{3}{9}\cdot\frac{3}{9}\cdot\frac{3}{9}\cdot\frac{9}{14}}{\Pr[E]} = \frac{0.0053}{\Pr[E]}$$

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/ 14	5/ 14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Tem
Sunny	Coo

← *Evidence E*

Pr [no|E]=Pr [Outlook=Sunny|no]×Pr [Tem | no]
×Pr [Humidity=High| no]×Pr [Windy=True| no]×Pr [no]/Pr [E]

$$= \frac{\frac{3}{5}\cdot\frac{1}{5}\cdot\frac{4}{5}\cdot\frac{3}{5}\cdot\frac{5}{14}}{\Pr[E]} = \frac{0.0206}{\Pr[E]}$$

Normalize…

- Pr[Yes | E] + Pr [No | E] = 1
  - Play can be either “Yes” or “No”

$$\frac{0.0053}{\Pr[E]} + \frac{0.0206}{\Pr[E]} = 1$$

$$\Pr[Yes \mid E] = \frac{0.00}{\Pr[E]} \Big/ \left(\frac{0}{\Pr[E]} + \frac{}{\Pr[E]} \right)$$

$$\Pr[Yes \mid E] = \frac{0.0053}{0.0053 + 0.0206} = 0.205$$

$$\Pr[No \mid E] = \frac{0.0206}{0.0053 + 0.0206} = 0.795$$

Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	ye		
turtle	no	no	so		
penguin	no	no	so		
porcupine	yes	no	no	yes	mammal
eel	no	no	yes	no	non-ma
salamander	no	no	sometimes	yes	non-ma
gila monster	no	no	no	yes	non-ma
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

Degenerate Probabilities (Pr[Outlook=Overcast|No)=0

- Could be a “true” representation of the real-world
  - Of course, one does not have to worry in that case
  - Rare

- The training ugh

  - Is it EVER =rainy” when “Play=no”?

  - If the answer is yes, a large uld have captured that fact
    - What does one do when data set is not big enough?
- We treat degeneracy seriously and try to remove it
  - Laplace approach

The “zero-frequency problem”

- Why does degeneracy matter?
(e.g. “Humidity = high” for class “yes”)

$$Pr\,[Humidity{=}High \mid yes]{=}0$$

$$Pr\,[yes \mid E]{=}0$$

  - Probability

  - (No matter how likely the o are!)

- Remedy: add 1 to the cou attribute value-class combination (*Laplace estimator*)
- Result: probabilities will never be zero!
(also: stabilizes probability estimates)

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

- Pretend that we add 3 rows of data containing only columns Outlook and Play:
  - All 3 rows have pla
  - 1 row with Outlook rcast and 3rd with Outlook = Rainy. See resulting change in conditiona elow. This eliminates the degenerate probability:

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	4	Hot	2	2	High	3	4	False	6	2	9	8
Overcast	4	1	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	3	Cool	3	1								
Sunny	2/9	4/8	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/17	8/17
Overcast	4/9	1/8	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	3/8	Cool	3/9	1/5								

Modified probability estimates

- In some cases, the number of rows to be added may need to be different from 3. In a more general setting we a

- Example: att s Play=No

$=(3+\mu/3)/\mu$ $=(0+\mu/3)/\mu$ $=(2+\mu/3)/\mu$

Sunny ***Overcast*** ***Rainy***

Testing for Independence OPTIONAL (Information Theoretic Testing)

- Let A and B be two random variables
- Let D(A,B) = (H(A) + H(B) – H(A,B))/H(A,B)
  - If A and B are independent
    - H(A,B) =
    - D(A,B) =
  - If A and B are linearly rela y correlated)
    - H(A,B) = H(A) = H(B)
    - D(A,B) = 1; this is the maximum
- If D() value is close to zero, assume independence
  - No need for looking up of statistical tables
  - Easy to implement

Piecing it all together

- We want to estimate $P(Y=1 \mid X_1,\dots,X_p)$
- But we don't have enough examples of each possible profile $X_1,\dots,$
- If we had inste could separate it to

$$P(X_1|Y=1) \cdot P(X_2|Y=1) \quad P(X_p|Y=1)$$

  - True if we can assume (conditional) independence between X_1, …,X_p within each class

Piecing it all together

$$P(Y=1 \mid X_1,\ldots,X_p) = \frac{P(X_1,\ldots,X_p \mid Y=1)P(Y=1)}{P(X_1,\ldots,X_p)}$$

$$\approx \frac{P(X_1 \mid Y=1)\cdot P(X_2 \mid Y=1)\cdots P(X_p \mid Y=1)\cdot P(Y=1)}{P(X_1,\ldots,X_p)}$$

Proportion of Play=Yes in training set

Proportion of rows with that predictor combination in the training set

Use the cutoff to determine classification of this observation.

Default: cutoff = 0.5 (classify to group that is most likely)

Advantages and Disadvantages

- The good
  - Simple
  - Can handle large amount of predictors
  - High perform
  - Pretty robust
- The bad
  - Need to categorize continuous pr
  - Predictors with “rare” categories ility (Use Laplace fix)
  - No insight about importance/role of each predictor

What is the probability of Play=Yes | Humidity=Normal and what would you predict for Play?

	Humidity High	Humidity Normal	Total
Play=Yes	5	7	12
Play=No	7	12	19
Total	12		31

A: 5/12, Predict Play = Yes

B: 7/19, Predict Play = Yes

C: 5/12, Predict Play = No

D: 7/19, Predict Play = No

E: None of the above

Naive Bayes works better with categorical data because

A: It takes less time to compute probabilities for categorical data

B: It cannot co ween different values for ca

C: It needs the predictor val n to some rows to compute accurate conditional probabilities

D: Numeric data slows down the computation too much

E: None of the above

Data Preprocessing using Weka

- Follow steps on the following page:
- http://facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/p

- File conversion and openi s in different applications
  - Excel, WordPad/TextEdit, Weka
  - CSV (text), XLSX (binary), ARFF (text)

Weka

- Run Naïve Bayes Classifier on cleaned and binned version of 4bank-data.csv

Next Session

- Testing and Validation