UCIrvine | The Paul Merage School of Business

2022.11.24

BANA 273 Session 4

Classification

Exact Bayes & Naïve Bayes

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
  - Assignment 2 posted
  - Project proposal (1 para) due soon (check 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
- Classification
  - Bayesian (Naïve Bayes)
  - Decision Tree (ID3)
- Association Rules
  - Apriori
- Clustering
  - K Means

A classic: Microsoft’s Paperclip

The New York Times
Wednesday, April 2, 2008

Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

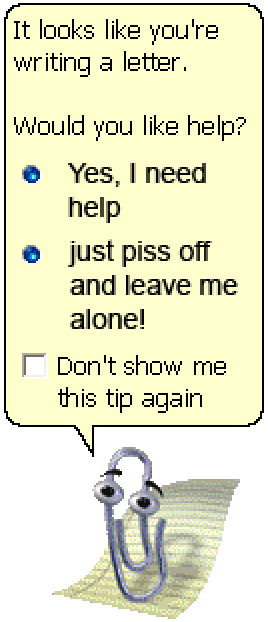
CIRCUITS CAMCORDERS CAMERAS CELL PHONES COMPUTERS HANDHELDS HOME VIDEO MUSIC PERIP

TECHNOLOGY; Microsoft Sees Software 'Agent' as Way to Avoid Distractions

By JOHN MARKOFF
Published: July 17, 2000

E-MAIL
PRINT
SINGLE-PAGE

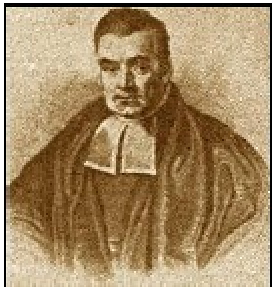
"The software considers what the value of the information is and the cost of the disruption," Mr. Horvitz said.



The Bayesian techniques have been widely adopted in Microsoft's products -- including the Paper Clip help wizard that pops up frequently to offer advice in the company's Office program. Many users, however, have criticized Paper Clip as an irritant, popping up too often with unwanted help.

Mr. Horvitz, who speaks apologetically about the Paper Clip program, said its shortcomings were the result of Microsoft's failure to implement all of his team's ideas.

Exact Bayes



Thomas Bayes

For each record to be classified:

1. Find all other records just like it (i.e. where all the predictor values are the same)
2. Determine what classes they all belong to and which class is more prevalent
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	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

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	?

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Notes

- Bayesian classifier works best with categorical attributes
  - Unlikely to find exact matches for numerical variables
- Numerical attributes must be binned and converted to categorical attributes
- 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 training records just like the new record
- Determine the probability that those 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	12	6	23

P(Play=Yes | outlook=sunny) = 40%
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?

2022.11.24

Classification

Using Naïve Bayes

Prof. Vibs Abhishek

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 $A_1, \ldots, A_p$ are ___________

- 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 "Outlook = Sunny and Temp = 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 that we are assuming conditional independence between *outlook* and *Temp*

- 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	3/9	1/5								

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	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 chart
- Create the chart using Microsoft Excel – 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	3/9	1/5								

- A new day:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

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	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← ***Evidence E***

Pr [yes|E]=Pr [Outlook=Sunny|yes]×Pr [Temperature=Cool|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	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← ***Evidence E***

Pr [no|E]=Pr [Outlook=Sunny|no]×Pr [Temperature=Cool| 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"

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

$$\Pr[Yes \mid E] = \frac{0.0053}{\Pr[E]} \Big/ \left(\frac{0.0053}{\Pr[E]} + \frac{0.0206}{\Pr[E]} \right)$$

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

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

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	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
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 data set is not big enough

  - Is it EVER possible to have “Outlook=rainy” when “Play=no”?

  - If the answer is yes, a larger data set would 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 will be zero!

  - (No matter how likely the other values are!)

- Remedy: add 1 to the count for every 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 play=no

  - 1 row with Outlook = Sunny, 2nd with Outlook = Overcast and 3rd with Outlook = Rainy. See resulting change in conditional probabilities below. 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 add μ rows.
- Example: attribute outlook for class 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
    - $H(A,B) = H(A) + H(B)$
    - $D(A,B) = 0$; this is the minimum
  - If A and B are linearly related (perfectly 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 $P(Y=1 \mid X_1,\ldots,X_p)$
- But we don't have enough examples of each possible profile $X_1\ldots, X_p$ in the training set

- If we had instead $P(X_1,\ldots,X_p \mid Y=1)$, we could separate it to

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

  - True if we can assume (conditional) independence between X_1, $\ldots,X_p$ 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 performance accuracy
  - Pretty robust to independence assumption
- The bad
  - Need to categorize continuous predictors
  - Predictors with “rare” categories -> zero probability (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	19	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 compute the distance between different values for categorical data

C: It needs the predictor values to match 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/preprocess.html

- File conversion and opening text files 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