CS102: Big Data

Tools and Techniques, Discoveries and Pitfalls

Spring 2017

Ethan Chan, Lisa Wang

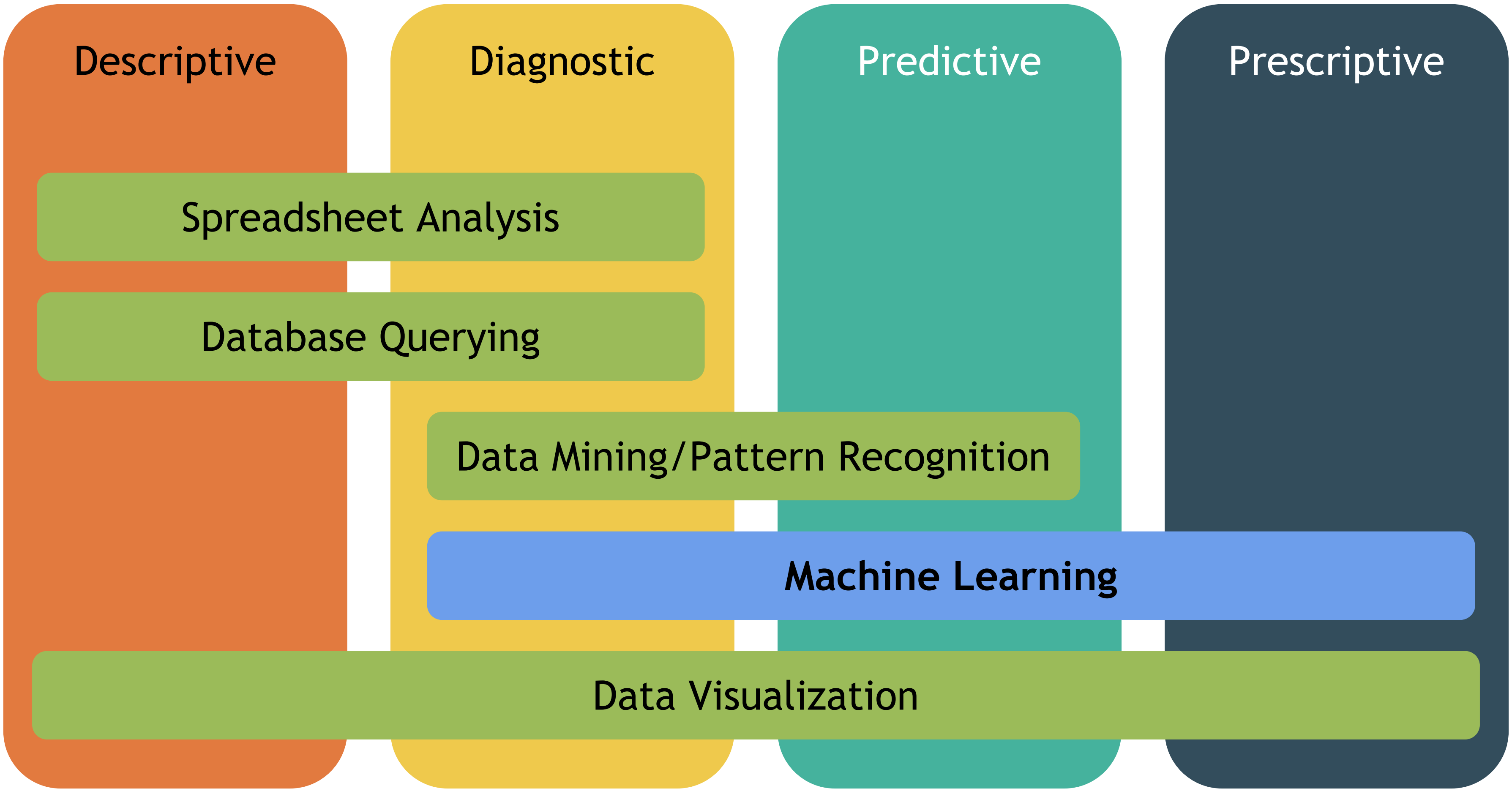
Lecture 12: Classification + Evaluation

Announcements

- We’re trying to get better, VPTL Small Group feedback session at 230pm today
- Assignment 4B have been released
  - “Predicting White or Red wine!”

- Assignment 4A and 4B <u>due May 23rd Tuesday</u>
- Final Project Help, come to Office Hours!

Tools & Techniques



Learning Goals

- Classification (continued)
  - Naive Bayes
  - Support Vector Machines
- Summary of Supervised Learning
- Guide to building Machine Learning models

Naive Bayes

Naive Bayes

- Define probability
- Define conditional probability
- Define Bayes Rule
- Define Conditional Independence
- Define Naive Bayes

Probability

<u>**Definition**</u>

Let event "Y" be if a person has (<u>cancer</u> or <u>no cancer</u>)

Let event "X" be the outcome of a test (<u>positive</u> or <u>negative</u>)

<u>**Basic Probability**</u>

P(Y = cancer) = 0.01 means a person has 1% chance of having cancer

P(Y = no cancer) = 0.99 , 99% chance of no cancer

<u>**Meaning**</u>

If you pick anyone on the street, there's a 1% chance that person has cancer.

Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
B	Cancer	Negative
C	No Cancer	Negative
D	Cancer	Positive
E	No Cancer	Negative
F	No Cancer	Negative
G	No Cancer	Positive

P(Y = cancer) = 3/7, P(Y = no cancer) = 4/7

P(X = positive) = 3/7 , P(X = negative) = 4/7

Conditional Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
B	Cancer	Negative
C	No Cancer	Negative
D	Cancer	Positive
E	No Cancer	Negative
F	No Cancer	Negative
G	No Cancer	Positive

$$P(Y = \text{cancer} \mid X = \text{positive}) = \frac{\#\text{cancer and positives}}{\#\text{positives}}$$

Probability of having cancer **given** that we know the test is positive.

Conditional Probability

Person	**Y (cancer or not)**	**X (test positive or negative)**
A	Cancer	Positive
[redacted]		
D	Cancer	Positive
[redacted]		
G	No Cancer	Positive

P(Y = cancer |X = positive)

$$= \frac{\text{\#cancer and positives}}{\text{\#positives}}$$

$$= \frac{\text{\#cancer and positives}}{3}$$

We only want to look at rows where X = positive first

We can see that there are 3 rows that have positive tests

Conditional Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
[redacted]		
D	Cancer	Positive
[redacted]		
G	No Cancer	Positive

P(Y = cancer |X = positive)

$$= \frac{\text{\#cancer and positives}}{\text{\#positives}}$$

$$= \frac{\mathbf{2}}{\mathbf{3}}$$

We can see that there are 2 rows where Y = cancer and X = Positive.

We know know that P(Y = cancer|X = positive) = ⅔ = 0.66

Conditional Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
B	Cancer	Negative
C	No Cancer	Negative
D	Cancer	Positive
E	No Cancer	Negative
F	No Cancer	Negative
G	No Cancer	Positive

What is P(Y = Cancer | X = negative)?

Conditional Probability

Person	Y (cancer or not)	X (test positive or negative)
[redacted]		
B	Cancer	Negative
C	No Cancer	Negative
[redacted]		
E	No Cancer	Negative
F	No Cancer	Negative
[redacted]		

What is $P(Y = \text{Cancer} \mid X = \text{negative})$? ¼.

Conditional Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
B	Cancer	Negative
C	No Cancer	Negative
D	Cancer	Positive
E	No Cancer	Negative
F	No Cancer	Negative
G	No Cancer	Positive

P(Y = cancer) = 3/7, P(Y = no cancer) = 4/7

P(X = positive) = 3/7 , P(X = negative) = 4/7

P(Y = cancer | X = positive) = ⅔

P(Y = cancer | X = negative) = ¼

Conditional Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
B	Cancer	Negative
C	No Cancer	Negative
D	Cancer	Positive
E	No Cancer	Negative
F	No Cancer	Negative
G	No Cancer	Positive

P(Y = cancer) = 3/7, P(Y = no cancer) = 4/7

P(X = positive) = 3/7 , P(X = negative) = 4/7

P(Y = cancer | X = positive) = ⅔

P(Y = cancer | X = negative) = ¼

How about P(X = positive | Y = cancer)?

Your Turn!

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
B	Cancer	Negative
C	No Cancer	Negative
D	Cancer	Positive
E	No Cancer	Negative
F	No Cancer	Negative
G	No Cancer	Positive

P(Y = cancer) = 3/7, P(Y = no cancer) = 4/7

P(X = positive) = 3/7 , P(X = negative) = 4/7

P(Y = cancer | X = positive) = ⅔

P(Y = cancer | X = negative) = ¼

How about P(X = positive | Y = no cancer)?

Your Turn!

Person	Y (cancer or not)	X (test positive or negative)
[redacted]		
C	No Cancer	Negative
[redacted]		
E	No Cancer	Negative
F	No Cancer	Negative
G	No Cancer	Positive

P(Y = cancer) = 3/7, P(Y = no cancer) = 4/7

P(X = positive) = 3/7 , P(X = negative) = 4/7

P(Y = cancer | X = positive) = ⅔

P(Y = cancer | X = negative) = ¼

How about P(X = positive | Y = no cancer)?
¼

Conditional Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
B	Cancer	Negative
[redacted]		
D	Cancer	Positive
[redacted]		

$P(Y = \text{cancer}) = 3/7$, $P(Y = \text{no cancer}) = 4/7$

$P(X = \text{positive}) = 3/7$, $P(X = \text{negative}) = 4/7$

$P(Y = \text{cancer} \mid X = \text{positive}) = \frac{2}{3}$

$P(Y = \text{cancer} \mid X = \text{negative}) = \frac{1}{4}$

How about $P(X = \text{positive} \mid Y = \text{cancer})$?
2/3.

Conditional Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
B	Cancer	Negative
[redacted]		
D	Cancer	Positive
[redacted]		

P(Y = cancer) = 3/7, P(Y = no cancer) = 4/7

P(X = positive) = 3/7 , P(X = negative) = 4/7

P(Y = cancer | X = positive) = ⅔

P(Y = cancer | X = negative) = ¼

How about P(X = positive | Y = cancer)? 2/3.

We can also get this answer ⅔ without counting by using our previous results obtained through <u>Bayes Rule.</u>

Bayes Rule

$$P(A|B) = P(B|A) * P(A) \;/\; P(B)$$

Not going to prove it in class, its a few lines for those who are interested.

Remember we wanted to find **P(X = positive | Y = Cancer).**

P(X=positive | Y=Cancer)

= P(Y=Cancer | X=positive) * P(X = positive) / P(Y = Cancer)

= (2/3) * (3/7) / (3/7)

= ⅔ (same as direct counting!)

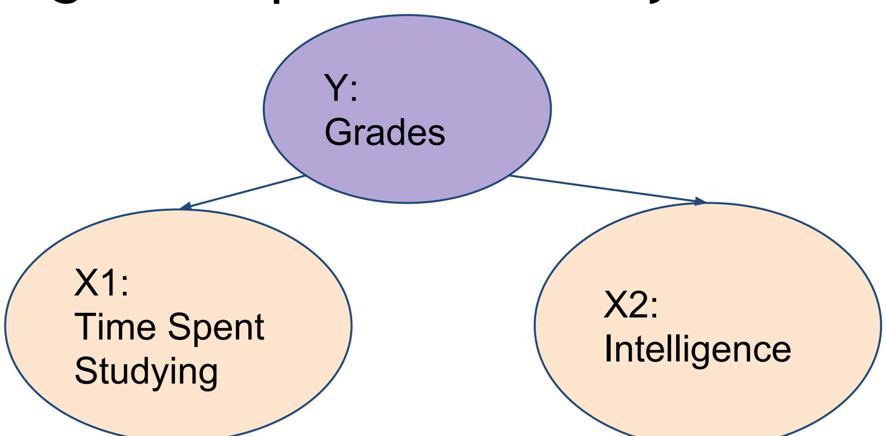
Conditional Independence

Definition: Given we know the Label, the probability of feature X1 occuring is independent of feature X2.

In Math: P(X1,X2 | Y) = P(X1 | Y) * P(X2 | Y)

<u>Naive Bayes</u> assumes all variables are conditionally independent, hence it is called "*naive*".

Example: The grade you get from a class through the time you spent studying is independent from your intelligence



Naive Bayes (Classification)

Features

- X1: Age [young / old]
- X2: Tumor Size [none / small / large]

Labels

- Y: [Cancer / No Cancer]

Person	X1 Age [young / old]	X2 Tumor Size [none / S / L]	Y (cancer or not)
A	old	none	Cancer
B	old	small	Cancer
C	young	none	No Cancer
D	young	large	Cancer
E	old	none	No Cancer
F	old	none	No Cancer
G	young	small	No Cancer

Naive Bayes (Classification)

Suppose this person who is (X1) old and has a (X2) small tumor comes to you..

Determine if P(Y = Cancer | X1 = old, X2 = small)

or P(Y = No Cancer | X1 = old, X2 = small) is greater

Naive Bayes (Classification)

Suppose this person who is (X1) old and has a (X2) small tumor comes to you..

Determine if P(Y = Cancer | X1 = old, X2 = small)

or P(Y = No Cancer | X1 = old, X2 = small) is greater

We can reform that equation using Bayes Rule:

P(Y = Cancer | X1 , X2)

= P(X1, X2 | Y = Cancer) * P(Y = Cancer) / P(X1,X2)

P(Y = No Cancer | X1 , X2)

= P(X1, X2 | Y = No Cancer) * P(Y = No Cancer) / P(X1,X2)

Naive Bayes (Classification)

Suppose this person who is (X1) old and has a (X2) small tumor comes to you..

Determine if P(Y = Cancer | X1 = old, X2 = small)

or P(Y = No Cancer | X1 = old, X2 = small) is greater

We can reform that equation using Bayes Rule:

P(Y = Cancer | X1 , X2)

= P(X1, X2 | Y = Cancer) * P(Y = Cancer) / P(X1,X2)

P(Y = No Cancer | X1 , X2)

= P(X1, X2 | Y = No Cancer) * P(Y = No Cancer) / P(X1,X2)

Since we only care about which one is bigger, we can drop the P(X1,X2) term.

Determine if P(X1, X2 | Y = Cancer) * P(Y = Cancer)

or P(X1, X2 | Y = No Cancer) * P(Y = No Cancer) is greater.

Naive Bayes (Classification)

Suppose this person who is (X1) old and has a (X2) small tumor comes to you..

Determine if P(Y = Cancer | X1 = old, X2 = small)

or P(Y = No Cancer | X1 = old, X2 = small) is greater

We can reform that equation using Bayes Rule:

P(Y = Cancer | X1 , X2)

= P(X1, X2 | Y = Cancer) * P(Y = Cancer) / P(X1,X2)

P(Y = No Cancer | X1 , X2)

= P(X1, X2 | Y = No Cancer) * P(Y = No Cancer) / P(X1,X2)

Since we only care about which one is bigger, we can drop the P(X1,X2) term.

Determine if P(X1, X2 | Y = Cancer) * P(Y = Cancer)

or P(X1, X2 | Y = No Cancer) * P(Y = No Cancer) is greater.

Use the Conditional Independence Assumption

P(X1 | Y = Cancer) * P(X2 | Y = Cancer)* P(Y = Cancer)

P(X1 | Y = No Cancer) * P(X2 | Y = No Cancer)* P(Y = No Cancer)

Naive Bayes (Classification)

Suppose this person who is (X1) old and has a (X2) small tumor comes to you..

Determine if P(Y = Cancer | X1 = old, X2 = small)

or P(Y = No Cancer | X1 = old, X2 = small) is greater

We can reform that equation using Bayes Rule:

P(Y = Cancer | X1 , X2)

= P(X1, X2 | Y = Cancer) * P(Y = Cancer) / P(X1,X2)

P(Y = No Cancer | X1 , X2)

= P(X1, X2 | Y = No Cancer) * P(Y = No Cancer) / P(X1,X2)

Since we only care about which one is bigger, we can drop the P(X1,X2) term.

Determine if P(X1, X2 | Y = Cancer) * P(Y = Cancer)

or P(X1, X2 | Y = No Cancer) * P(Y = No Cancer) is greater.

Use the Conditional Independence Assumption

P(X1 | Y = Cancer) * P(X2 | Y = Cancer)* P(Y = Cancer)

P(X1 | Y = No Cancer) * P(X2 | Y = No Cancer)* P(Y = No Cancer)

Depending on which term is larger, we predict if a person has cancer or not

In English

"We are finding the probability of features {length of nose, weight, height} occuring given that it is an elephant"

Score =

Probability of having nose length 1 meter given that is is an elephant *

Probability of having weight 500 pounds given that is is an elephant *

Probability of having height 3 meters given that is is an elephant

And then comparing this score for other possible animals

In English

“We are finding the probability of features {length of nose, weight, height} occuring given that it is an elephant”

Score =

Probability of having nose length 1 meter given that is is an elephant *

Probability of having weight 500 pounds given that is is an elephant *

Probability of having height 3 meters given that is is an elephant

And then comparing this score for other possible animals

Probability is just counting of number of rows in the data!

Going back to example data

Person	X1 Age [young / old]	X2 Tumor Size [none / S / L]	Y (cancer or not)
A	old	none	Cancer
B	old	small	Cancer
C	young	none	No Cancer
D	young	large	Cancer
E	old	none	No Cancer
F	old	none	No Cancer
G	young	small	No Cancer

P(X1 = old | Y = Cancer) * P(X2 = small| Y = Cancer)* P(Y = Cancer)

=

P(X1 = old | Y = No Cancer) * P(X2 = small| Y = No Cancer)* P(Y = No Cancer)

=

Going back to example data

Person	X1 Age [young / old]	X2 Tumor Size [none / S / L]	Y (cancer or not)
A	old	none	Cancer
B	old	small	Cancer
C	young	none	No Cancer
D	young	large	Cancer
E	old	none	No Cancer
F	old	none	No Cancer
G	young	small	No Cancer

P(X1 = old | Y = Cancer) * P(X2 = small| Y = Cancer)* P(Y = Cancer)

$= \frac{2}{3} * \frac{1}{3} * 3/7 = 0.0952$

P(X1 = old | Y = No Cancer) * P(X2 = small| Y = No Cancer)* P(Y = No Cancer)

$= 2/4 * \frac{1}{4} * 4/7 = 0.0714$

Going back to example data

Person	X1 Age [young / old]	X2 Tumor Size [none / S / L]	Y (cancer or not)
A	old	none	Cancer
B	old	small	Cancer
C	young	none	No Cancer
D	young	large	Cancer
E	old	none	No Cancer
F	old	none	No Cancer
G	young	small	No Cancer

P(X1 = old | Y = Cancer) * P(X2 = small| Y = Cancer)* P(Y = Cancer)

= $\frac{2}{3}$ * $\frac{1}{3}$ * 3/7 = 0.0952

P(X1 = old | Y = No Cancer) * P(X2 = small| Y = No Cancer)* P(Y = No Cancer)

= 2/4 * $\frac{1}{4}$ * 4/7 = 0.0714

Model Predicted this person has cancer!!! :o
0.0952 > 0.0714

Naive Bayes Summary

This is all you need to know for purposes of this class

Given training data that follows this format..

Feature X1	Feature X2	Feature X3	...	Feature X999	Y (Label) [A or B]
..	..	..	..	..	..
..	..	..	..	..	..

And you are given new data without labels that you want to classify

Feature X1	Feature X2	Feature X3	...	Feature X999	Y (Label) [A or B]
..	..	..	..	..	???
..	..	..	..	..	???

Determine if

P(X1 | Y = A) * P(X2 | Y = A) * P(X3 | Y = A) * .. * P(X999 | Y = A) *** P(Y = A)**

OR

P(X1 | Y = B) * P(X2 | Y = B) * P(X3 | Y = B) * .. * P(X999 | Y = B) *** P(Y = B)**

Is greater

Now, your turn!

Person	**X1 Age [young / old]**	**X2 Tumor Size [none / S / L]**	**Y (cancer or not)**
A	old	none	Cancer
B	old	small	Cancer
C	young	none	No Cancer
D	young	large	Cancer
E	old	none	No Cancer
F	old	none	No Cancer
G	young	small	No Cancer

Given a new person who is X1 = young and X2 = small, what will the model predict?

Remember! Determine if
P(X1 | Y = A) * P(X2 | Y = A) * P(X3 | Y = A) * .. * P(X999 | Y = A) * **P(Y = A)**
OR
P(X1 | Y = B) * P(X2 | Y = B) * P(X3 | Y = B) * .. * P(X999 | Y = B) * **P(Y = B)**
Is greater

Now, your turn!

Person	X1 Age [young / old]	X2 Tumor Size [none / S / L]	Y (cancer or not)
A	old	none	Cancer
B	old	small	Cancer
C	young	none	No Cancer
D	young	large	Cancer
E	old	none	No Cancer
F	old	none	No Cancer
G	young	small	No Cancer

Given a new person who is X1 = young and X2 = small, what will the model predict?

P(X1 = young | Y = Cancer) * P(X2 = small| Y = Cancer)* P(Y = Cancer)

= $\frac{1}{3} * \frac{1}{3} * 3/7 = 0.0476$

P(X1 = young | Y = No Cancer) * P(X2 = small | Y = No Cancer)* P(Y = No Cancer)

= $2/4 * \frac{1}{4} * 4/7 = 0.0714$

0.0476 < 0.0714, NO CANCER!! :)

Code in Python

(refer to notebook for full code)

```
features = ['minutes', 'shots', 'passes', 'tackles', 'saves']
nb = GaussianNB()
nb.fit(playersTrain[features],playersTrain['position'])
predictions = nb.predict(playersTest[features])
```

Why Naive Bayes?

1. Simple and easy to implement (just counting)
2. Computationally fast
3. Works well on small datasets

Real World Examples

1. Classify an email as spam, or not spam
2. Classify a news article to its category

Support Vector Machines

Finds a line that best seperates 2 classes of points.