The University of Texas ARLINGTON

Bayesian Classifiers
(Put together from many sources)

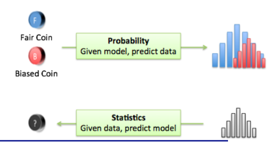
Sharma Chakravarthy

Department of Computer Science and Engineering
University of Texas at Arlington
Fall 2019

Probability and Statistics

- We need a little bit of probability and statistics before going into Bayes approach to classification
- So, what is the difference between probability and statistics?
  - **Probability** is starting with an animal, and figuring out what footprints it will make.
  - **Statistics** is seeing a footprint, and guessing the animal.

Probability & Statistics



Basics

- Consider rolling two dice (a blue and yellow one) and noting down the numbers on each roll (alternatively, drawing from a deck of cards)
- Let *X* and *Y* denote the number of dots we get on the blue and yellow dice, respectively, and consider the meaning of P(X+Y=6) in the large!
- Sample space is the list of Possible outcomes as shown
- We place a weight of $\frac{1}{36}$ on each Outcome reflecting they are equally likely (important!)
- Outcomes (1,5), (2,4),…, (5,1) have a total weight of $\frac{5}{36}$

1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5.6
6,1	6,2	6,3	6,4	6,5	6,6

Sample Space for the Dice Example

Intuitively

- Sample space may be difficult to comprehend for some problems
- If you are doing an experiment recording the outcome in a note book, it may look like →
- After many, many repetitions,
  Approximately 5/36 of the lines
  Will have a 'yes' value!
- If you did the experiment 1008 times, approximately X+Y =6 will occur 5/36*1008 =140 times

notebook line	outcome	blue+yellow = 6?
1	blue 2, yellow 6	No
2	blue 3, yellow 1	No
3	blue 1, yellow 1	No
4	blue 4, yellow 2	Yes
5	blue 1, yellow 1	No
6	blue 3, yellow 4	No
7	blue 5, yellow 1	Yes
8	blue 3, yellow 6	No
9	blue 2, yellow 5	No

- This is what probability really is: In what fraction of the lines does the event of interest happen? **It sounds simple, but if you think about this "as counting the outcome of an experiment", probability problems are a lot easier to understand**. And it is the fundamental basis of computer simulation.

Basics

- We assume experiments are repeatable!
- We assume that experiments are performed a large number of times (in the large is important)
- An *event* A has a Boolean outcome (yes/no) of the experiment. Examples:
  - X+Y = 6
  - X =1
  - Y = 3
  - X – Y = 4
- A *random variable (X or Y)* is a numerical outcome of the experiment: X+Y, 2*X*Y,

Definition of P(A ^ B) Boolean AND

- For any event of interest *A*, imagine a column on *A* in the notebook. The k^{th} line (or row) in the notebook, *k* = 1, 2, 3, …, will say Yes or No, depending on whether *A* occurred or not during the k^{th} repetition of the experiment. For instance, we can have such a column in our table above, for the event {blue + yellow = 6}.
- For any event of interest *A*, we define *P*(*A*) to be the long-run fraction of lines with Yes entries.
- For any events *A* ^ *B* (Boolean AND), imagine a new column in our notebook, labeled "A ^ B." In each line, this column will say Yes if and only if there are Yes entries in columns *A* and *B*.
- *P*(*A* ^ *B*) is then defined to be the long-run fraction of lines with Yes entries in the new column labeled "A ^ B."

Definition of P(A v B)

- For any events *A*, *B* imagine a new column in our notebook, labeled "A ˅ B." In each line, this column will say Yes if and only if at least one of the entries for *A* or *B* says Yes.
- *P*(*A* ˅ *B*) is then defined to be the long-run fraction of lines with Yes entries in the new column labeled "A ˅ B."

Definition

- For any events *A*, *B* imagine a new column in our notebook, labeled "A | B" and pronounced "A given B." In each line:
  - This new column will say "NA" ("not applicable") if the B entry is No.
  - If it is a line in which the B column says Yes, then this new column will say Yes or No, depending on whether the A column says Yes or No.
- Then *P*(*A*|*B*) means the long-run fraction of lines in the notebook in which the A | B column says Yes—**among the lines which do NOT say NA**.
  - Conditional probability

Common mistake

- **A common mistake is to confuse** *P*(*A* ^ *B*) and *P*(*A*|*B*). This is important to understand. Compare the values of *P*(*X* = 1 ^ *S* = 6) and *P*(*X* = 1|*S* = 6) , where S = X+Y
- *P*(*X* = 1 ^ *S* = 6) is 1/36 and comes from (1,5) out of possible 36 outcomes (35 no and 1 yes)
- *P*(*X* = 1 | *S* = 6) is 1/5 comes from (1,5) out of possible 5 relevant outcomes.

outcome	S=6	X=1?
(1,5)	yes	yes
(2, 4)	yes	no
(3, 3)	yes	no
(4,2)	yes	no
(5, 1)	yes	no
All other cases	NA	hence, not used!

1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5.6
6,1	6,2	6,3	6,4	6,5	6,6

Please understand this clearly!

HW: What is P(X=2 v S =6) where S = X _ Y? What is P(X=2 v Y=4)?

More Definitions

- *Suppose A, B are events such that it is impossible for them to occur at the same time. They are said to be* **disjoint** or "mutually exclusive" *events*. "being freshman" and "being sophomore" are disjoint events
- If A and B are disjoint events, then P(A ˅ B) = p(A) + P(B). Can be generalized to

  P(A_1 ˅ A_2, …, ˅ A_k) $= \sum_{i=1}^{k} P(Ai)$

  are x=2 and x+y = 2 Disjoint events?

  Yes 6/36 + 1/36

1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

- if A and B are not disjoint, then P(A ˅ B) = P(A) + P(B) – P(A ^ B)
- To compute P(X=3 or Y=4), first check whether they are disjoint! 1/6 + 1/6 – 1/36 (we have counted (3, 4) twice! Hence the subtraction)
- No compact generalization for ˅!

What about "female" or "freshman"?

More Definitions

- *Events A and B are independent (or unrelated) if*

  *P(A ^ B) = P(A) * P(B)* = *P(B^A)*

  P(A_1 and A_2, …, and A_k) = $\prod_{i=1}^{k} P(Ai)$

What about P(x=3 ^ X=7)? Events can be independent and disjoint if one of them is null or never Happens! "promoted" ^ "audited"?

- How do we know whether two events are disjoint, independent?
  - Must be inferred from the problem, domain, context!
  - 2 cards are drawn from a deck without replacement! Independent or dependent? (HW)
  - What about with replacement? (HW)
- For example, when two dices are rolled, outcome on each dice is independent of each other!

  P(X=4 ^ Y =5) = 6/36 * 6/36 = 1/36

  where as events "good grade" ^ "completing the project" are not independent events (related)

  Contrast with P(X=1^S=6)!

1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

More Definitions

- *If Events A and B are not independent, the previous expression generalizes to*

  *P(A ^ B) = P(A) * P(B|A) instead of* P(A) * P(B). Suppose we use this for

  P(X=4 ^ Y=5) = P(X=4)*P(Y=5|X=4)

  = 6/36 * 1/6 = 1/36 (same as earlier, why?)

  Check P(X=1^S=6)
- Suppose 20% of all UTA students are in the COE, and 15% of engineering majors are female. (there are non-COE females! At UTA)
- P(COE student ^ female) =

  P(COE student) * P(female | COE student)

  0.2 * 0.15 = 0.3 or 3% of UTA students are "female engineers"
- When A and B are actually independent

  P(B|A) is the same as P(B) and reverts to P(A ^ B) = P(A) * P(B)

- P(B|A) is also know as conditional probability!
- Probability of B conditionally occurring on A
- P(B|A) = P(B) if A and B are independent!

Bayes theorem

- We can express P(A ^ B) = P(A) * P(B|A) as

$$P(B|A) = \frac{P(A \wedge B)}{P(A)}$$

Similarly,

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

Bayes theorem provides a way of Going from P(X|Y) to P(Y|X) or from A sample labeled population to a unseen data outcome prediction!
This is what classification is

giving rise to Bayes theorem

Significance: if I know individual probabilities of A and B, and P(B|A), I can compute the Probability of P(A|B)!

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)}$$

Example of Bayes Theorem

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time P(S|M) is 0.5 conditional probability
  - Prior probability of any patient having meningitis is 1/50,000 P(M) sample or population probability
  - Prior probability of any patient having stiff neck is 1/20 P(S)
- If a patient has stiff neck, what's the probability he/she has meningitis? Posterior probability P(M|S)

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Bayesian Classifiers

- How can we apply this to classification using a training data set
- Consider each attribute and class label as random variables
- Given a record with attributes (A_1, A_2,…,A_n)
  - Goal is to predict class C
  - Specifically, we want to find the value of C that maximizes P(C| A_1, A_2,…,A_n)
- Can we estimate P(C| A_1, A_2,…,A_n) directly from data?

Bayesian Classifiers

- Approach:
  - Compute the posterior probability P(C | A_1, A_2, …, A_n) for all values of C using the Bayes theorem

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}$$

  - Choose value of C that maximizes P(C | A_1, A_2, …, A_n) posterior probability!
  - Equivalent to choosing value of C that maximizes numerator) P(A_1, A_2, …, A_n|C) P(C)
- How to estimate P(A_1, A_2, …, A_n | C)? How do we know P(C)?

Naïve Bayes Classifier

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}$$

- In the above expression, we do not know how to compute probability $P(A_1, A_2, \ldots, A_n \mid C)$ because of multiple A's!
- This is where the assumption of independence of A_i events comes into use!
  - Also why it is called Naïve!
- Assuming independence among attributes A_i when class is given:
  - $P(A_1, A_2, \ldots, A_n) = P(A_1)*p(A_2)* \ldots * P(A_n)$
  - $P(A_1, A_2, \ldots, A_n \mid C)$ can be written as $P(A_1 \mid C_j) * P(A_2 \mid C_j) * \ldots * P(A_n \mid C_j)$

  Can estimate $P(A_i \mid C_j)$ for all A_i and C_j.

  - New point is classified to C_j if $P(C_j) \Pi P(A_i \mid C_j)$ is maximal.

Simple one attribute example

A_1 C

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

P(Ai|no) P(Ai|yes)

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				P(A)
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64	P(C)	

Compute P(Yes|Sunny) = P(Sunny|yes) * P(Yes)/P(Sunny)
= 3/9*0.64/0.36 = 0.6

P(No|Sunny) = 0.4

Compute P(yes|overcast)
P(overcast|yes) * P(yes)/P(overcast)
= 4/9*0.64/0.29 = 0.98

Compute P(No|overcast)

Step 1: convert data into frequency table (can be done for each attribute)

Step 2: create probabilities (likelihood) table for the sample data

Step 3: use Naïve Bayes equation to compute the posterior probability for each class

- Can be easily extended to multi-attribute and multi-class!
- For numerical attributes, normal distribution is assumed!

Simple one attribute example

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Compute P(Yes|Sunny) = P(Sunny|yes) * P(Yes)/P(Sunny)
= 3/9*0.64/0.36 = 0.6

Compute P(yes|overcast)
P(overcast|yes) * P(yes)/P(overcas
= 4/9*0.64/0.29 = 0.98

P(No|Sunny) = 0.4

- To compute P(No|overcast), we need to compute P(overcast|No) what is the value!
- In this sample, there were no no-play days when it was overcast!
- Due to this, the probability of the computation P(No|overcast) becomes 0.
- This is corrected by adding a small value (usually 1) as correction called 'Laplace Correction' (may be a parameter)

Naïve Bayes Classifier

- If one of the conditional probability is zero, then the entire expression becomes zero
- Most algorithms use one of the following or take a parameter for this correction
- Probability estimation:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m - estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes

p: prior probability

m: parameter

HW: What is P(No|overcast) With Laplace Correction!

- M is a constant; p is a weight relative to observed data

Multiple attribute example

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Class: $P(C) = N_c/N$
  - e.g., P(No) = 7/10, P(Yes) = 3/10
- For discrete attributes:
  $P(A_i \mid C_k) = |A_{ik}| / N_c$
  - where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k
  - Examples:
    P(Status=Married|No) = 4/7
    P(Refund=Yes|Yes)=0
    P(Refund=Yes|No)= 3/7
    P(Refund=No|No) = 4/7

What is Gaussian Naïve Bayes?

- So far we have seen computations when the A's are categorical. But how to compute the probabilities when A is a continuous variable?
- If we assume that A follows a particular distribution, then you can plug in the probability density function of that distribution to compute the probability of likelihoods.
- If you assume the A's follow a Normal (aka Gaussian) Distribution, which is fairly common, we substitute the corresponding probability density of a Normal distribution and call it the Gaussian Naive Bayes. You need the mean and variance of the A to compute this formula.

How to Estimate Probabilities from Data?

- For continuous attributes:
  - Discretize the range into bins and thus transform the attribute into an ordinal attribute.
  - Probability density estimation:
    - Assume attribute follows a normal distribution
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, can use it to estimate the conditional probability $P(A_i|c)$

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Normal distribution:

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - One for each (A_i, c_j) pair
- For (Income, Class=No):
  - If Class=No
    - sample mean = 770K/7 = 110K
    - sample variance = 2975
    - Sample std =sqrt(2975) = 54.54

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Example of Naïve Bayes Classifier

- Given a Test Record: X = (Refund=No, Married, Income = 120K)

- P(X|Class=No) = P(Refund=No|Class=No)
  × P(Married| Class=No)
  × P(Income=120K| Class=No)
  = 4/7 × 4/7 × 0.0072 = 0.0024
- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  × P(Married| Class=Yes)
  × P(Income=120K| Class=Yes)
  = 1 × 0 × 1.2 × 10^{-9} = 0
- Since P(X|No)P(No) > P(X|Yes)P(Yes)
- Therefore P(No|X) > P(Yes|X)
  => Class = No

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/3
P(Marital Status=Divorced|Yes)=1/3
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

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

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

A: attributes
M: mammals
N: non-mammals

$$P(A|M)=\frac{6}{7}\times\frac{6}{7}\times\frac{2}{7}\times\frac{2}{7}=0.06$$

$$P(A|N)=\frac{1}{13}\times\frac{10}{13}\times\frac{3}{13}\times\frac{4}{13}=0.0042$$

$$P(A|M)P(M)=0.06\times\frac{7}{20}=0.021$$

$$P(A|N)P(N)=0.004\times\frac{13}{20}=0.0027$$

P(A|M)P(M) > P(A|N)P(N)
=> Mammals

Naïve Bayes (Summary)

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques such as Bayesian Belief Networks (BBN)

Thank You !!!

For more information visit: http://itlab.uta.edu