PROBABILISTIC LEARNING NAÏVE BAYES CLASSIFIERS

pasquale Rullo

rullo@mat.unical.it

Probabilistic classifiers

- Let an instance X and a set of classes $\{c_1, \ldots, c_n\}$ be given
- A probabilistic classifier determines a probability distribution function
  - $p(c_1|X)$
  - …
  - $p(c_i|X)$
  - …
  - $p(c_n|X)$
- where $p(c_i|X)$ is the conditional probability that X belongs to c_i
- Then, outputs class c_j with the highest probability

Conditional probability

A	B	C
a	b	c
a	b	d
e	b	c
d	h	c

p(X|Y): probability of X given Y

What is the probability of having B=b, given C=c?

Notation: p(B=b|C=c) or p(b|c)

Conditional probability

A	B	C
a	**b**	**c**
a	b	d
e	**b**	**c**
d	**h**	**c**

By definition of conditional probability:

$$p(b|c) = \frac{p(b,c)}{p(c)}$$

$$p(b|c) = \frac{p(b,c)}{p(c)} = 2/4*4/3= 2/3$$

Product rule - Joint probability

- From the definition of conditional probability, the joint probability is

$$p(X,Y) = p(X|Y)\,p(Y) = p(Y|X)\,p(X)$$

- X and Y are independent if $p(X|Y) = p(X)$

$$\Rightarrow p(X,Y) = p(X)\,p(Y)$$

- X and Y are incompatible (mutually exclusive) if $p(X|Y) = 0$

$$\Rightarrow p(X,Y) = 0$$

Sum rule

- $p(X \vee Y) = p(X) + p(Y) - p(X,Y) =$
  $p(X) + p(Y) - p(X|Y)\, p(Y) =$
  $p(X) + p(Y) - p(Y|X)\, p(X)$

- If X and Y are independent
  - $p(X \vee Y) = p(X) + p(Y) - p(X)\, p(Y)$

- If X and Y are incompatible
  - $p(X \vee Y) = p(X) + p(Y)$

Sum rule

$p(A \vee B) = p(A) + p(B) - p(A \wedge B)$

Events A,B

independent

$p(A \wedge B) = p(A)p(B)$

$p(A \vee B) = p(A) + p(B) - p(A)p(B)$

dependent

incompatible

$p(A \wedge B) = 0$

$p(A \vee B) = p(A) + p(B)$

compatible

$p(A \wedge B) = p(A|B)p(B)$

$p(A \vee B) = p(A) + p(B) - p(A|B)p(B)$

Sum rule - Example

- What is the probability of getting A={1,2} from the first die or B={2,3} from the second one in the throw of two dice?
- A and B are independent on each other
  - $p(A \vee B) = p(A) + p(B) – p(A)p(B)$
- A={1,2}: 1 and 2 are incompatible
  - $p(A) = p(1)+p(2) = 1/6+1/6 = 1/3$
- B={2,3}: 2 and 3 are incompatible
  - $p(B) = p(2)+p(3) = 1/6+1/6 = 1/3$

- $p(A \vee B) = p(A) + p(B) – p(A)p(B) = 1/3+1/3-1/9= 5/9$

Sum rule – Example (cont'd)

- There are 24 configurations favorable to event A $\vee$ B
  - A=1 with any B – 6 configurations: <1, 1>, …, <1, 6>
  - A=2 with any B – 6 configurations: <2, 1>, …, <2, 6>
  - B=2 with any A – 6 configurations: <1, 2>, …, <6, 2>
  - B=3 with any A – 6 configurations: <1, 3>, …, <6, 3>

Sum rule – Example (cont'ed)

- There are 24 configurations favorable to event A ∨ B
  - A=1 with any B – 6 configurations: <1, 1>, <1,2>, <1,3>…, <1, 6>
  - A=2 with any B – 6 configurations: <2, 1>,<2,2>, <2,3>, …, <2, 6>
  - B=2 with any A – 6 configurations: <1, 2>, <2,2>, …, <6, 2>
  - B=3 with any A – 6 configurations: <1, 3>, <2,3>, …, <6, 3>

  4 of which are duplicated: <1,2>, <1,3>, <2,2>,<2,3>
- So the number of favorable configurations without repetitions is 20 (over 36)
  - $p(A \vee B) = 20/36 = 5/9$
- This explains the need of the joint probability for computing the total probability

$$p(A \vee B) = p(A) + p(B) - p(A \wedge B)$$

Sum rule - Examples

- What is the probability of getting A={1,2} or B={3,4} in the throw of one die?
  - A and B are incompatible, so $p(A,B)=0$
  - $p(A \vee B) = p(A) + p(B) = 1/3+1/3 = 2/3$
- What is the probability of getting A={1,2} or B={2,3} in the throw of one die?
  - A and B are compatible (when 2 occurs, both A and B occur)
  - $p(A \vee B) = p(A) + p(B) – p(A|B)p(B)$
  - $p(A) = p(B) = 1/3$
  - $p(A|B) = 1/2$
  - $p(A \vee B) = 1/3+1/3-1/2*1/3= 1/2$

Theorem of total probability

- If $\{X_1, \ldots, X_n\}$ are mutually exclusive events such that $p(X_1) + \ldots + p(X_n) = 1$, then

$$p(Y) = \Sigma_{i=1,n}\, p(Y|X_i)\, p(X_i)$$

- Example: In a school, 60% of students are female. The percentage of males who passed the final exam of Math is 0.5, while that of females is 0.66. What is the probability of event Y = "a student has passed the exam"?

- $p(Y) = p(Y|f)\, p(f) + p(Y|m)\, p(m) = 0.66*0.6+0.5*0.4 = 0.55$

Summary of basic probability formulas

- **Product rule**: $p(X,Y) = p(X|Y)\ p(Y) = p(Y|X)\ p(X)$
- **Sum rule**: $p(X \vee Y) = p(X)+p(Y)-p(X,Y)$
- **Total probability**: $p(Y) = \ \Sigma_{i=1,n}\ p(Y|X_i)\ p(X_i)$, if $\{X_1, \ldots, X_n\}$ are mutually exclusive events such that $p(X_1) + \ldots + p(X_n) = 1$

Bayes' Theorem

- Bayes' theorem may be derived from the definition of conditional probability:

  - $p(X|Y) = p(X,Y) / p(Y)$
  - $p(Y|X) = p(X,Y) / p(X)$
  - $\Rightarrow p(X,Y) = p(X|Y)\, p(Y) = p(Y|X)\, p(X) \Rightarrow$

$$p(X|Y) = \frac{p(Y|X)\, p(X)}{p(Y)}$$

Terminology:

- $p(X|Y)$: posterior probability
- $p(X)$: prior probability – initial degree of belief in *X*
- $p(X|Y)/p(Y)$: support *Y* provides for *X*

Bayes theorem: an example

- In a school, 60% of students are female. The percentage of males passing the final exam of Math is 0.5, while that of females is 0.66.

Passed Math	Sex
y	f
y	f
y	f
y	f
n	f
n	f
y	m
y	m
n	m
n	m

- Event Y = “a student has passed the Math exam”
- What is the probability that the student is a female?
- p(f |Y) ?

Bayes theorem: an example

- Question: p(f |Y)?
- Input data
  - p(m)=0.4, p(f)=0.6
  - p(Y|m)=0.5, p(Y|f)=0.66
- Bayes theorem
  - p(f|Y) = p(Y|f) * p(f)/p(Y)

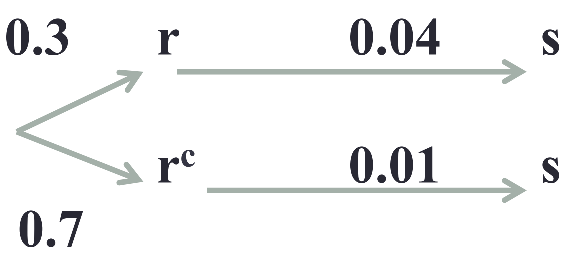
  where
  - p(Y) = p(Y|m)*p(m) + p(Y|f)*p(f) = 0.57 (total probability)

Answer: p(f|Y) = 0.66*0.6/0.57 = 0.69

Bayes theorem: an example

- In a Formula 1 Gran prix, the rain probability is 30%. The probability that Vettel wins when it's raining is 4%, and 1%, otherwise. Now, assuming that Vettel won the race, what is the probability that it has rained?

r=rain s=Vettel won



p(r|s)=0.3*0.04/(0.3*0.04+0.7*0.01) =0,94488

Bayes theorem: an example

- The entire output of a car factory is produced on three plants. The three plants account for 10%, 40%, and 50% of the output, respectively. The fraction of red cars produced by each plant is: 8% for the first plant; 5% for the second plant; 1% for the third plant. If a car is chosen at random from the total output and is found to be red, what is the probability that it was produced by the second plant?

Summary of basic probability formulas

- **Product rule**: $p(X,Y) = p(X|Y)\ p(Y) = p(Y|X)\ p(X)$
- **Sum rule**: $p(X \vee Y) = p(X)+p(Y)-p(X,Y)$
- **Total probability**: $p(Y) = \Sigma_{i=1,n}\ p(Y|X_i)\ p(X_i)$, if $\{X_1, \ldots, X_n\}$ are mutually exclusive events such that $p(X_1)+\ldots+p(X_n) = 1$
- **Bayes Theorem**: $p(X|Y) = \dfrac{p(Y|X)\ p(X)}{p(Y)}$

Naïve Bayes (NB) classifier

- Question: Given a new instance $<x_1, \ldots, x_n>$, what is the probability that the class is c?

$$p(c|<x_1, \ldots, x_n>) \ ?$$

- By the Bayes theorem

- $$p(c|<x_1, \ldots, x_n>) = \frac{p(<x_1, \ldots, x_n>|c)\, p(c)}{p(<x_1, \ldots, x_n>)}$$

- $p(c|<x_1, \ldots, x_n>)$ is the posterior probability for c
- p(c) is the prior probability for c

NB classifier

- Given the set of classes $C=\{c_1, \ldots, c_m\}$

$$c_{NB} = \underset{c_j \in C}{\text{argmax}}\, p(c_j | \langle x_1, \ldots, x_n \rangle) =$$

- $$= \underset{c_j \in C}{\text{argmax}} \frac{p(\langle x_1, \ldots, x_n \rangle | c_j)\, p(c_j)}{p(\langle x_1, \ldots, x_n \rangle)}$$

- the denominator is equal for all classes ➔

$$c_{NB} = \underset{c_j \in C}{\text{argmax}}\, p(\langle x_1, \ldots, x_n \rangle | c_j)\, p(c_j)$$

Evaluating prior probabilities

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}}\ p(<x_1, \ldots, x_n>|c_j)\ \ p(c_j)$$

- where
  - $p(c_j)$ and $p(<x_1, \ldots, x_n>|c_j)$ are called prior probabilities
  - $p(c_j)$ is the fraction of examples with class label c_j
  - $p(<x_1, \ldots, x_n>|c_j)$ is the number of examples of type $<x_1, \ldots, x_n>$ over the total number of examples with label c_j
- Evaluating $p(<x_1, \ldots, x_n>|c_j)$ would require a very, very large set of training data

The Conditional Independence Assumption (CIA)

The NB classifier is based on the simplifying assumption that the attribute values are conditionally independent, i.e., the probability of observing the conjunction $<x_1, \ldots, x_n>$ is given by the product of the probabilities of the single attributes, i.e.,

$$p(<x_1, \ldots, x_n>|c_j) = p(x_1|c_j) \ldots p(x_n|c_j)$$

➔

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}}\, p(c_j)\, p(x_1|c_j) \ldots p(x_n|c_j)$$

The conditional independence assumption (CIA)

- The CIA states the following

$$p(X,Y|C) = p(X|C)\ p(Y|C)$$

- Indeed

$$p(X,Y|C) = p(X,Y,C)/p(C) =$$
$$p(X,Y,C)/p(Y,C) * p(Y,C)/p(C) =$$
$$p(X|Y,C) * p(Y|C)$$

- Since $p(X|Y,C) = p(X|C)$ (i.e., X is conditionally independent of Y), it turns out that

$$p(X,Y|C) = p(X|C)*p(Y|C)$$

NB classifier – independent attributes

- For instance, in the mammal data set, the attributes gives birth and #legs are independent
- On the contrary, if the examples represent persons, then the attributes Height and Shoe Size are NOT independent

Evaluating prior probabilities

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}}\ p(c_j)\, p(x_1|c_j) \ldots p(x_n|c_j)$$

- where
  - $p(c_j)$ and $p(x_1|c_j) \ldots p(x_n|c_j)$ are called prior probabilities
  - $p(c_j)$ is the probability that class c_j is the label of some instance of the training set
  - $p(x_i|c_j)$ is the probability that the value x_i appears in some instance of c_j

- They can be estimated over the training data

  - $p(c_j) = Nc_j / N$
    - Nc_j = number of instances labeled c_j
    - N = total number of instances
  - $p(x_i|c_j)$ = fraction of instances with label c_j where x_i appears

- Evaluating prior probabilities is all a NB classifier has to do during the training phase

Classifying by NB - An Example

- p(Yes) = 9/14=0.64
- p(No) = 5/14 = 0.36
- p(Outlook=sunny | Yes) = 2/9 = 0.22
- p(Temp=cool | Yes) = 3/9 = 0.33
- p(Hum=high | Yes) = 3/9 = 0.33
- p(Wind=strong | Yes) = 3/9 = 0.33
- …
- p(Wind=strong | No) = 3/5 = 0.60

Day	Outlook	Temperature	Humidity	Wind	playTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Classifying by NB - An Example (cont'ed)

Classify the following instance:

X = <Outlook=sunny, Temp=cool, Hum=high, Wind= strong>

$$c_{NB} = \underset{c \in \{Yes, No\}}{\operatorname{argmax}} \; p(c|X) = p(c)\, p(x_1|c) \ldots p(x_n|c)$$

- p(Yes|X) = p(Yes) p(sunny|yes) p(cool|yes) p(high|yes) p(strong|yes)
- p(no|X) = p(no) p(sunny|no) p(cool|no) p(high|no) p(strong|no)

Classifying by NB - An Example (cont'ed)

- p(Yes) = 0.64
- p(No) = 0.36
- p(sunny | Yes) = 2/9 = 0.22
- p(cool | Yes) = 3/9 = 0.33
- p(high | Yes) = 3/9 = 0.33
- p(strong | Yes) = 3/9 = 0.33
- …
- p(strong | No) = 3/5 = 0.60

- p(Yes|X) = p(Yes) p(sunny|yes) p(cool|yes) p(high|yes) p(strong|yes) = 0.0053

- p(No|X) = p(No) p(sunny|No) p(cool|No) p(high|No) p(strong|No) = 0.026

- ➔ c_{NB} = No

On the conditional independence assumption (CIA) - Example

- On the training set, the following holds:
  - p(X=0|No) = 0.4, p(X=1|No) = 0.6
  - p(X=0|Yes) = 0.6, p(X=1|Yes) = 0.4
  - p(Y=0|No) = 0.4, p(Y=1|No) = 0.6
  - p(Y=0|Yes) = 0.6, p(Y=1|Yes) = 0.4
  - p(No) = p(Yes) = 0.5

X	Y	C
0	0	No
0	0	No
1	1	No
1	1	No
1	1	No
0	1	Yes
0	0	Yes
0	1	Yes
1	0	Yes
1	0	Yes

On the conditional independence assumption (CIA) – Example (cont'ed)

- Classify E = <X=0, Y=0>

- By using NB (with the CIA)

  $P(No|E) = p(E|No)\, p(No) = p(<X=0,Y=0>|No)p(No) =$

  $p(X=0|No)\, p(Y=0|No)\, p(No) = 0.08$

  $P(Yes|E) = p(E|Yes)\, p(Yes) = p(<X=0,Y=0>|Yes)\, p(Yes) =$

  $p(X=0|Yes)\, p(Y=0|Yes)\, p(Yes) = 0.18$

- P(Yes|E) > P(No|E) ➔ E is assigned to class Yes

On the conditional independence assumption (CIA) - Example

- On the training set, the following holds:
  - p(X=0|No) = 0.4, p(X=1|No) = 0.6
  - p(X=0|Yes) = 0.6, p(X=1|Yes) = 0.4
  - p(Y=0|No) = 0.4, p(Y=1|No) = 0.6
  - p(Y=0|Yes) = 0.6, p(Y=1|Yes) = 0.4
  - p(No) = p(Yes) = 0.5

X	Y	C
0	0	No
0	0	No
1	1	No
1	1	No
1	1	No
0	1	Yes
0	0	Yes
0	1	Yes
1	0	Yes
1	0	Yes

Note: X and Y are perfectly correlated when C=No, so

- p(X=v,Y=v|No) = p(X=v|No) = p(Y=v|No)

On the conditional independence assumption (CIA) – Example (cont'ed)

- Since X and Y are perfectly correlated when C=No

$$p(\langle X=0,Y=0\rangle|No) = p(X=0|No)=p(Y=0|No) = 0.4$$

- Thus

$$p(No|E) = p(E|No)\ p(No)= p(\langle X=0,Y=0\rangle|No)\ p(No) =$$
$$p(X=0|No)\ p(No) = 0.4*0.5 = 0.2$$

- Since $p(No|E) > p(Yes|E)=0.18$, E should correctly be assigned to No (instead of Yes, to which is assigned based on the CIA)

Attribute Values with zero probability

- Classify the following instance:

  E = <Outlook=sunny, Temp=cool, Hum=high, Wind= strong>

- Assume that there is no example with Hum=high in the training set, so that

$$p(Hum{=}high|Yes) = p(Hum{=}high|No) = 0$$

and, thus,

$$p(Yes|E) = p(No|E) = 0$$

An Example (cont'ed)

Day	Outlook	Temperature	Humidity	Wind	playTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Attribute Values with zero probability

- REMEDY: recall that

$$p(x|c) = n_x/N_c$$

- i.e., p(x|c) is the fraction of instances under c where attribute A has value x
- Now, we set

$$p(x|c) = \frac{n_x + kq}{N_c + k}$$

where

- k is a constant between 0 and 1 (usually 1)
- q = 1/n, where n is the number of possible values for attribute A

Attribute Values with zero probability

- Thus, to classify the instance

  E = <Outlook=sunny, Temp=cool, Hum=high, Wind= strong>

- we evaluate

$$p(\text{Hum=high}|\text{Yes}) = \frac{n_{high} + kq}{N_{yes} + k}$$

where

  - n_{high} = 3 is the number of Yes examples with Hum=high
  - q=1/3, since Hum takes on 3 possible values
  - N_{yes} = 9 is the number of Yes examples
- By setting k=1 ($0 \leq k \leq 1$)

$$p(\text{Hum=high}|\text{YES}) = 3.33/(9+1) = 0.33$$

NB - Exercise

Id	Home owner	Marital status	Annual income	Defaulted borrower
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	Married	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Classify E = <No, Married, 120K>

NB – Exercise (cont'ed)

- Classify E = <No, Married, 120K>

p(Yes|E) = p(Yes) * p(HomeOw=no|Yes) * p(status=married|Yes) * p(Income=120|Yes)

p(No|E) = p(No) * p(HomeOw=no|No) * p(status=married|No) * p(Income=120|No)

- where
  - p(yes) = 0.3, p(No) = 0.7

NB – Exercise (cont'ed)

Classify E = <No, Married, 120K>

- p(HomeOw=no|No) = 4/7
- p(HomeOw=no|Yes) = 1

- p(status=Married|No) = 4/7
- p(status=Married|Yes) = 1/3

Annual income:

Class=No

- Mean=110
- Standard deviation= 54,54
- p(120|No) = 0.0072

Class=Yes

- Mean=90
- Standard deviation = 5
- p(120|yes) = 0

- p(Yes|E) = 0
- p(No|E) = 0.7* 4/7 * 4/7 * 0.0072 > p(Yes|E)

Conclusions

- The classification function of an instance $X=<x_1, \ldots, x_n>$ is

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} \; p(c_j)\, p(x_1|c_j) \ldots p(x_n|c_j)$$

- The instance X is classified under the class c which maximizes the conditional probability p(c|X)
- There is no explicit search of the hypothesis space.
- Correlated attributes may degrade performance because of the CIA
- Very efficient