Naïve Bayes

Prof. Dr. Christina Bauer

christina.bauer@th-deg.de

Faculty of Computer Science

Classification

- Email: Spam/not Spam
- Online transactions: Fraudulent (yes/no)
- Tumour: malignant/benign

$y \in \{0,1\}$ 0: "negative class" (e.g. benign tumour)

1: "positive class" (e.g. malignant tumour)

- Multiclass classification problem → $y \in \{0,1,2,3\}$

Classification

- The relationship between attribute set and the class variable is non-deterministic.
- Even if the attributes are same, the class label may differ in training set even and hence can not be predicted with certainty.
- Reason: noisy data, certain other attributes are not included in the data.

→ We want to model a probabilistic relationship between the attribute set and the class variable.

Bayesian Classification

- Problem statement
- Given features $X_1,X_2,...,X_n$
- Predict a label Y

Bayesian Classification

- A probabilistic framework for solving classification problems
- Conditional Probability:
- $P(C|A) = \frac{P(C,A)}{P(A)}$
- $P(A|C) = \frac{P(C,A)}{P(C)}$

- Bayes theorem:
- $P(C|A) = \frac{P(A|C) * P(C)}{P(A)}$

EXAMPLE OF BAYES THEOREM

- Given:
  - A doctor knows that Cold causes fever 50 % of the time
  - Prior probability of any patient having cold is 1/50,000
  - Prior probability of any patient having fever is 1/20
- If a patient has fever, what's the probability they have a Cold?

- $P(Cold|Fever) = \frac{P(Fever|Cold)*P(Cold)}{P(Fever)}$ = $\frac{0.5*\frac{1}{50000}}{\frac{1}{20}}$ = 0.0002

Bayesian Classification

- 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 the data?

Bayesian Classification

- Approach:
  - Compute the posterior probability P(C|A_1, A_2,...,A_n) for all values of C using the Bayes theorem
  - $P(C|\mathrm{A}_1, \mathrm{A}_2, \dots, \mathrm{A}_\mathrm{n}) = \frac{P(\mathrm{A}_1, \mathrm{A}_2, \dots, \mathrm{A}_\mathrm{n}|C) * P(C)}{P(\mathrm{A}_1, \mathrm{A}_2, \dots, \mathrm{A}_\mathrm{n})}$
  - Choose value of C that maximizes P(C|A_1, A_2,...,A_n)
- How to estimate $P(\mathrm{A}_1, \mathrm{A}_2, \dots, \mathrm{A}_\mathrm{n}|C)$?

NAÏVE BAYES CLASSIFIER

- Assume independence among attributes A_i when class is given:
- $P(A_1, A_2, \ldots, A_n|Cj) = P(A_1|Cj) * P(A_2|Cj) * \ldots * P(A_n|Cj)$

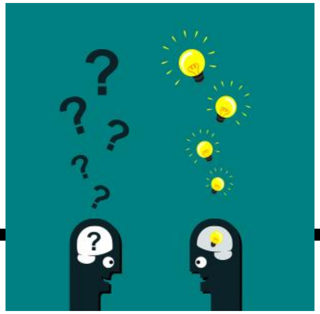
- Can estimate $P(A_i|C_j)$ for all A_i and C_j.
- New point is classified to C_j if $P(C_j) \prod P(A_i| C_j)$ is maximum

ESTIMATE PROBABILITIES FROM DATA

Refund	Marital Status	Taxable Income	Evade
Yes	Single	125K	No
No	Married	100K	No
No	Single	70K	No
Yes	Married	120K	No
No	Divorced	95K	Yes
No	Married	60K	No
Yes	Divorced	220K	No
No	Single	85K	Yes
No	Married	75K	No
No	Single	90K	Yes

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

QUESTION



Given the dataset, how many different probabilities do you have to calculate for P(Marital Status|Evade)?

A: 2

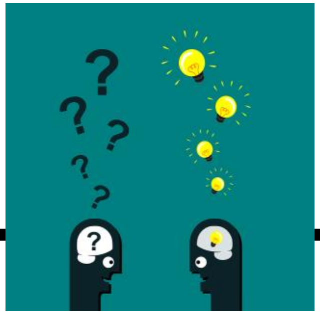
B: 6

C: 4

D: 1

Refund	Marital Status	Taxable Income	Evade
Yes	Single	125K	No
No	Married	100K	No
No	Single	70K	No
Yes	Married	120K	No
No	Divorced	95K	Yes
No	Married	60K	No
Yes	Divorced	220K	No
No	Single	85K	Yes
No	Married	75K	No
No	Single	90K	Yes

QUESTION



Given the dataset, how many different probabilities do you have to calculate for P(Marital Status|Evade)?

A: 2

B: 6

C: 4

D: 1

- P(Status=Married|No)
- P(Status=Single|No)
- P(Status=Divorced|No)
- P(Status=Married|Yes)
- P(Status=Single|Yes)
- P(Status=Divorced|Yes)

Refund	Marital Status	Taxable Income	Evade
Yes	Single	125K	No
No	Married	100K	No
No	Single	70K	No
Yes	Married	120K	No
No	Divorced	95K	Yes
No	Married	60K	No
Yes	Divorced	220K	No
No	Single	85K	Yes
No	Married	75K	No
No	Single	90K	Yes

Estimate Probabilities From Data

- For continuous attributes:
  - Discretize the range into bins
    - one ordinal attribute per bin
    - violates independence assumption
- 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, you can use it to estimate the conditional probability $P(A_i|c)$

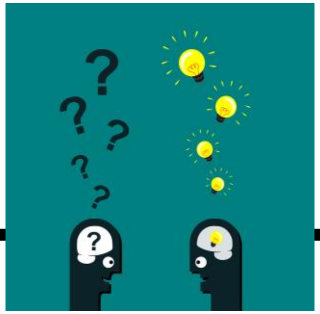
Estimate Probabilities From Data

Refund	Marital Status	Taxable Income	Evade
Yes	Single	125K	No
No	Married	100K	No
No	Single	70K	No
Yes	Married	120K	No
No	Divorced	95K	Yes
No	Married	60K	No
Yes	Divorced	220K	No
No	Single	85K	Yes
No	Married	75K	No
No	Single	90K	Yes

- Normal distribution:
  - $P(A_i|c_j) = \frac{1}{\sqrt{2\pi\sigma^2_{ij}}} * e^{-\frac{(A_i-\mu_{ij})^2}{2\sigma^2_{ij}}}$
  - one for each $(A_i|c_j)$ pair
- Example: (Income, Class = No):

- if Class = No
  - sample mean = 110
  - sample variance = 2975
  - $P(\text{Income} = 120|\text{No}) = \frac{1}{\sqrt{2\pi 2975}} * e^{-\frac{(120-110)^2}{2*2975}} =$ 0.0072

QUESTION



Given the dataset, how do you calculate P(Income =60|No)?

(Hint: mean = 110; variance = 2975)

A: $\frac{1}{\sqrt{2\pi 2975}} * e^{-\frac{(120-110)^2}{2*2975}}$

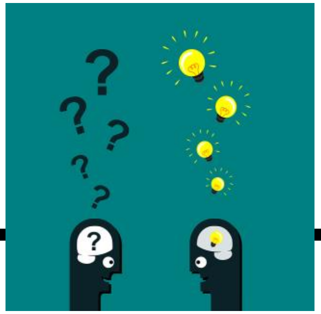
B: $\frac{1}{\sqrt{2\pi 2975}} * e^{-\frac{(60-110)^2}{2*2975}}$

C: $\frac{1}{\sqrt{2\pi}} * e^{-\frac{(60-110)^2}{2*2975}}$

D: Not possible with the given data.

Refund	Marital Status	Taxable Income	Evade
Yes	Single	125K	No
No	Married	100K	No
No	Single	70K	No
Yes	Married	120K	No
No	Divorced	95K	Yes
No	Married	60K	No
Yes	Divorced	220K	No
No	Single	85K	Yes
No	Married	75K	No
No	Single	90K	Yes

QUESTION



Given the dataset, how do you calculate P(Income =60|No)?

(Hint: mean = 110; variance = 2975)

A: $\frac{1}{\sqrt{2\pi 2975}} * e^{-\frac{(120-110)^2}{2*2975}}$

~~B~~: $\frac{1}{\sqrt{2\pi 2975}} * e^{-\frac{(60-110)^2}{2*2975}}$

C: $\frac{1}{\sqrt{2\pi}} * e^{-\frac{(60-110)^2}{2*2975}}$

D: Not possible with the given data.

Refund	Marital Status	Taxable Income	Evade
Yes	Single	125K	No
No	Married	100K	No
No	Single	70K	No
Yes	Married	120K	No
No	Divorced	95K	Yes
No	Married	60K	No
Yes	Divorced	220K	No
No	Single	85K	Yes
No	Married	75K	No
No	Single	90K	Yes

EXAMPLE

Given: X =(Refund = No, Married, Income = 120k)

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

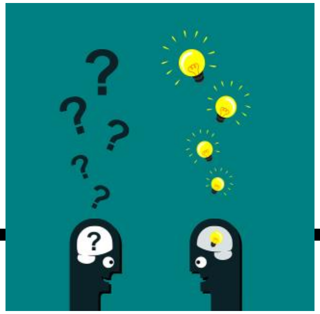
For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

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

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)
=> Class = No

QUESTION



Given:

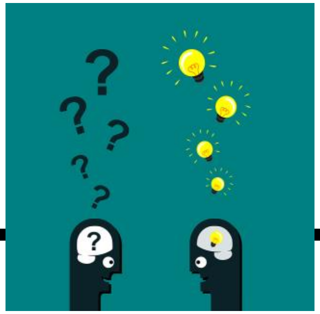
X =(Refund = yes, Married, Income = 120k)

What kind of decision do you make?

A: $\text{Class} = \text{No}$

B: Class = Yes

QUESTION



Given:

X =(Refund = yes, Married, Income = 120k)

What kind of decision do you make?

~~A~~: Class = No

→P(X|No) = P(Refund = yes|No) * p(Married|No) * p(Income = 120k|No) * p(No) =

→(3/7*4/7*0,00072) * 7/10

B: Class = Yes

→ P(X|Yes) = P(Refund = yes|Yes) * p(Married|Yes) * p(Income = 120k|Yes) * p(Yes) =

→(0*0*1.2*10^{-9})*3/10

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

PROBABILITY OF A SINGLE DECISION

Binary Problem:

P(X|Class = No) / [P(X|Class = No) + P(X|Class = Yes)]

Example:

0.0024/ 0.0024 + 0=

100%

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

- 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

NAÏVE BAYES CLASSIFIER

- 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
- Really easy to implement and often works well
- Often a good first thing to try
- Really fast (no iterations; not optimization needed)
- Commonly used as a "punching bag" for "smarter" algorithms

Wrap-Up

- We assume that the relationship between the attribute set and the class variable is non-deterministic. Therefore, we set up a probabilistic framework for solving classification problems using the Bayes theorem:
  - $P(C|A) = \frac{P(A|C)*P(C)}{P(A)}$
- Our approach is to compute the posterior probability $P(C|A_1, A_2, \ldots, A_n)$ for all values of the class C using the Bayes theorem (with $A_1, A_2, \ldots, A_n$ being our attribute, e. g. features):
  - $P(C|A_1, A_2, \ldots, A_n) = \frac{P(A_1, A_2, \ldots, A_n|C)*P(C)}{P(A_1, A_2, \ldots, A_n)}$
- After that we choose the value of C that maximizes $P(C|A_1, A_2, \ldots, A_n)$.

WRAP-UP

- We use an Naïve Bayes Classifier and assume independence among the attributes A_i when the class is given:
  - $P(A_1, A_2, \ldots, A_n|Cj)$= $P(A_1|Cj)$* $P(A_2|Cj)$* …* $P(A_n|Cj)$

- Therefore we can estimate $P(A_i|C_j)$ for all A_i and C_j.
- Every new point is classified to C_j if $P(C_j) \prod P(A_i| C_j)$ is maximum.

WRAP-UP

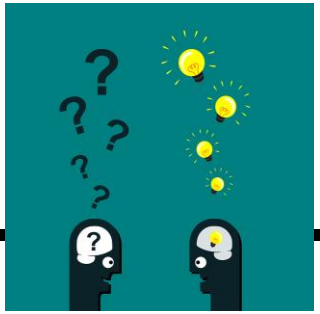
- The procedure is to calculate a look-up-table that includes:
  - The probability of the class: $P(C) = N_c/N$
  - The probability of an attribute given a class
    - For discrete attributes:
      - $P(A_i|\ C_k) = |A_{ik}|/\ N_c$ where $|A_{ik}|$ is the number of instances having attribute A_i and belongs to class C_k
    - For continuous attributes:
      - We assume the attribute follows a normal distribution
      - $P(A_i|c_j) = \frac{1}{\sqrt{2\pi\sigma^2{}_{ij}}} * \ e^{-\frac{(A_i-\mu_{ij})^2}{2\sigma^2{}_{ij}}}$
- If we want to calculate the probability of a single decision for a binary Problem we use:
  - P(X|Class = {No or Yes}) / [P(X|Class = No) + P(X|Class = Yes)]

WRAP-UP

The main characteristics of the Naïve Bayes Classifier are:

- It is robust to isolated noise points.
- It can handle missing values by ignoring the instance during probability estimate calculations.
- It is robust to irrelevant attributes.
- Nevertheless, the independence assumption may not hold for some attributes, e.g. if they are lineary dependent.
- It is really easy to implement and often works well.
- It is often a good first thing to try.
- It is eally fast (no iterations; not optimization needed).
- Therefore, it is commonly used as a “punching bag” for “smarter” algorithms.

QUESTION



See “Naive Bayes Exercise” in ilearn