Probabilistic classification models

Outline:

1. Naïve Bayes: main idea
2. Naïve Bayes classifier
3. Smoothing
4. Continuous inputs

1. Naïve Bayes: main idea

- Naïve Bayes (NB) classifier belongs to the family of probabilistic classification models.
- Given the information about the values of explanatory variables **x**, what is the probability that the sample belongs to class y?
- We need to calculate the *posterior* probability P(y|**x**) by means of Bayes' theorem.
- This could be done if we have the values of the *prior* probability P(y) and the *class conditional* probability P(**x**|y) .
- Suppose there are *H* distinct values for the target variable y denoted as $H = \{v_1, v_2, \ldots, v_H\}$.
- The posterior probability P(y|**x**) according to Bayes' theorem:

$$P(y=v_h|\mathbf{x}) = P(\mathbf{x}|\ y=v_h)P(y=v_h)/P(\mathbf{x}) = P(\mathbf{x}|\ y=v_h)P(y=v_h)/\sum_{i=1}^{H} P(\mathbf{x}|y=v_i)P(y=v_i)$$

1. Naïve Bayes: main idea

- Bayes' formula: $P(y=v_h|\mathbf{x}) = P(\mathbf{x}|\ y=v_h)P(y=v_h)/P(\mathbf{x})$

  The simple formula for conditional probability $P(A|B) = P(A \cap B)/P(B)$ is sufficient!

- We get: $P(A,B) = P(A \cap B) = P(A|B) \times P(B)$

  $P(B|A) = P(A \cap B)/P(A)$

  $= P(A|B) \times P(B)/P(A) \longrightarrow P(y=v_h\ |\mathbf{x}) = P(\mathbf{x}|\ y=v_h)\ P(y=v_h)\ /P(\mathbf{x})$

  Also

  $P(\mathbf{x}) = P(\mathbf{x} \cap y = v_1) + P(\mathbf{x} \cap y = v_2) + \ \dots . + P(\mathbf{x} \cap y = v_H) \ = \sum_{i=1}^{H} P(\mathbf{x}|y=v_i)P(y=v_i)$

- Suppose we have a binary classification problem with a target feature y that can have values 0 or 1.
- What is the probability that y = 1 given the value of the descriptive feature **x**? $P(y = 1|\mathbf{x})$
- What is the probability that y = 0 given the value of the descriptive feature **x**? $P(y = 0|\mathbf{x})$

1. Naïve Bayes: main idea

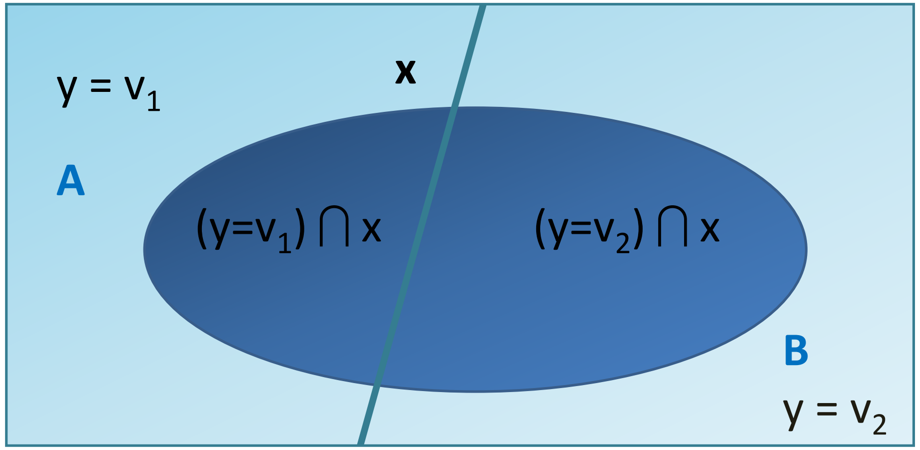
- Bayes' formula: $P(y{=}v_h|\mathbf{x}) = P(\mathbf{x}|\ y{=}v_h)P(y{=}v_h)/P(\mathbf{x}) = P(\mathbf{x}|\ y{=}v_h)P(y{=}v_h)/\sum_{i=1}^{H} P(\mathbf{x}|y{=}v_i)P(y{=}v_i)$

  We know: $P(A|\mathbf{x}) = P(A \cap \mathbf{x})/P(\mathbf{x})$

  $P(A \cap \mathbf{x}) = P(A|\mathbf{x}) \times P(\mathbf{x})$

  $P(\mathbf{x}|A) = P(A \cap \mathbf{x})/P(A)$

  $= P(A|\mathbf{x}) \times P(\mathbf{x})/P(A)$



- Suppose $H = 2$, $v_1 = 0$, $v_2 = 1$
- What is the probability that $y = v_1$ given $\mathbf{x}$? Answer: $P(y{=}v_1|\mathbf{x})$
- For simplicity, let A be the event that $y = v_1$ and B be the event that $y = v_2$, then
  - probability that $y = v_1$ is $P(A)$ and probability that $y = v_2$ is $P(B)$ ← prior probability
  - and $P(A|\mathbf{x}) = P(\mathbf{x}|A)\ P(A)/P(\mathbf{x})$ $P(B|\mathbf{x}) = P(\mathbf{x}|B)\ P(B)/P(\mathbf{x})$ ← posterior probability
  - also $P(\mathbf{x}) = P(\mathbf{x} \cap A) + P(\mathbf{x} \cap B)$

    $= P(\mathbf{x}|A)P(A) + P(\mathbf{x}|B)P(B)$ ← marginal probability

1. Naïve Bayes: main idea

- Posterior probilities:

$$P(A|\mathbf{x}) = P(\mathbf{x}|A)\, P(A)/P(\mathbf{x})$$

$$P(B|\mathbf{x}) = P(\mathbf{x}|B)\, P(B)/P(\mathbf{x})$$

- If $P(A|x)$ is greater than $P(B|x)$, classify the sample as class A ($y = v_1$), otherwise classify it as class B ($y = v_2$).

Maximum a posteriori hypothesis (MAP):

- Since the denominator $P(\mathbf{x})$ is the same for both posterior probabilities, we need only to compare

$$P(\mathbf{x}|A)\, P(A) \quad \text{and} \quad P(\mathbf{x}|B)\, P(B)$$

- Conclude $y = v_1$ if $P(\mathbf{x}|A)\, P(A)$ is the larger of the two, otherwise conclude $y = v_2$

The classifier assumes that given the target class, the explanatory variables are <u>conditionally independent</u>:

$$P(\mathbf{x}|y) = P(x_1|y) \times P(x_2|y) \times P(x_3|y) \times \ldots\ldots \times P(x_n|y)$$

- The descriptive feature/explanatory variable **x** normally consists of many components:

it is an n dimensional input, $\mathbf{x} = \{x_1, x_2, \ldots., x_n\}$.

2. Naïve Bayes classifier

- The classifier assumes that given the target class, the explanatory variables are **conditionally independent**:

$$P(\mathbf{x}|y) = P(x_1|y) \times P(x_2|y) \times P(x_3|y) \times \ldots.. \times P(x_n|y)$$

- For example, without the conditional independence assumption:

$$P(C,B,A|D) = P(C|D) \times P(B|C \cap D) \times P(A|\ B \cap C \cap D)$$

With the independence assumption:

$$P(C,B,A|D) = P(C|D) \times P(B|D) \times P(A|D)$$

- Two events are said to be **independent** of each other if knowledge of one event has no effect on the probability of the other event, and vice versa:

$$P(X|Y) = P(X)$$
$$P(X,Y) = P(X \cap Y) = P(X) \times P(Y)$$

$$\boxed{\begin{aligned} P(X|Y) &= P(X \cap Y)/P(Y) \\ &= P(X) \times P(Y)/P(Y) = P(X) \end{aligned}}$$

- **Conditional independence**: two or more events may be independent if a third event has happened

$$P(X|Y,Z) = P(X|Z)$$
$$P(X,Y|Z) = P(X|Z) \times P(Y|Z)$$

$$\boxed{\begin{aligned} P(X,Y|Z) &= P(X \cap Y \cap Z)/P(Z) \\ &= P(X|Y \cap Z) \times P(Y \cap Z)/P(Z) \\ &= P(X|Z) \times P(Y|Z) \times P(Z)/P(Z) \\ &= P(X|Z) \times P(Y|Z) \end{aligned}}$$

2. Naïve Bayes classifier

- For categorical or discrete numerical attribute $\mathbf{a_j}$:

Target attribute Meningitis is True or False

$$P(x_j | y) = P(x_j = r_{jk} | y = v_h) = s_{jhk}/m_h$$

where s_{jhk} is the number of class v_h for which the attribute takes value r_{jk} and m_h is the total number of samples of class v_h in data set $\mathbb{D}$. For example: Do you have Meningitis given Headache, no Fever and Vomiting?

ID	Headache	Fever	Vomiting	Meningitis
1	True	True	False	False
2	False	True	False	False
3	True	False	True	False
4	True	False	True	False
5	False	True	False	True
6	True	False	True	False
7	True	False	True	False
8	True	False	True	True
9	False	True	False	False
10	True	False	True	True

We first compute from the table,

P(Headache, no Fever, Vomiting|Meningitis):

Number of samples with Meningitis = true: 3

Among those with Meningitis, the number that have Headache, no Fever, Vomiting = 2

Hence,

P(Headache,no Fever, Vomiting|Meningitis) = 2/3

- For numerical attributes, $P(x_j | y)$ is estimated by making some assumption regarding its distribution.

2. Naïve Bayes classifier

- Example 1.

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

There are 4 discrete attributes in the data: Outlook, Temperature, Humidity and Wind.

The decision is to PlayTennis or not.

For tomorrow, the weather forecast is:

- Outlook: sunny
- Temperature: cool
- Humidity: high
- Wind: strong

Do we play tennis?

Prior probabilities:

P(Playtennis = Yes) = 9/14

P(Playtennis = No) = 5/14

2. Naïve Bayes classifier

- Example 1.

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

Estimate conditional probabilities, for example:

P(Wind = strong|Playtennis = Yes) = 3/9

P(Wind=strong|Playtennis = No) = 3/5

Compute conditional probabilities for the other attributes.

Then compute:

- P(Yes)P(Sunny|Yes)P(Cool|Yes)P(High|Yes)P(Strong|Yes) = (9/14)(2/9)(3/9)(3/9)(3/9) = 0.00529
- P(No)P(Sunny|No)P(Cool|No)P(High|No)P(Strong|No) = (5/14)(3/5)(1/5)(4/5)(3/5) = **0.02057**

Maximum a posteriori hypothesis:
Compare: P(**x**|A) P(A) and P(**x**|B) P(B)

Decision: **No**, we do not play tennis.

2. Naïve Bayes classifier

- Example 2.

ID	Credit History	Guarantor/ CoApplicant	Accomodation	Fraud
1	current	none	own	true
2	paid	none	own	false
3	paid	none	own	false
4	paid	guarantor	rent	true
5	arrears	none	own	false
6	arrears	none	own	true
7	current	none	own	false
8	arrears	none	own	false
9	current	none	rent	false
10	none	none	own	true
11	current	coapplicant	own	false
12	current	none	own	true
13	current	none	rent	true
14	paid	none	own	false
15	arrears	none	own	false
16	current	none	own	false
17	arrears	coapplicant	rent	false
18	arrears	none	free	false
19	arrears	none	own	false
20	paid	none	own	false

Loan application fraud detection

Three descriptive attributes:

Credit history: current, paid, arrears, none

Guarantor/CoApplicant: none, guarantor, coapplicant

Accomodation: own, rent, free

Binary target:

Fraud: true, false

P(true) = 6/20 = 0.3, P(false) = 14/20 = 0.7

Query:

- Credit history = paid
- Guarantor/CoApplicant = none
- Accommodation = rent
- Fraud = ?

2. Naïve Bayes classifier

- Example 2.

Compute all the conditional probabilities required for NB classification:

$$
\begin{array}{rclrcl}
P(fr) & = & 0.3 & P(\neg fr) & = & 0.7 \\
P(CH = \text{'none'} \mid fr) & = & 0.1666 & P(CH = \text{'none'} \mid \neg fr) & = & 0 \\
P(CH = \text{'paid'} \mid fr) & = & 0.1666 & P(CH = \text{'paid'} \mid \neg fr) & = & 0.2857 \\
P(CH = \text{'current'} \mid fr) & = & 0.5 & P(CH = \text{'current'} \mid \neg fr) & = & 0.2857 \\
P(CH = \text{'arrears'} \mid fr) & = & 0.1666 & P(CH = \text{'arrears'} \mid \neg fr) & = & 0.4286 \\
P(GC = \text{'none'} \mid fr) & = & 0.8334 & P(GC = \text{'none'} \mid \neg fr) & = & 0.8571 \\
P(GC = \text{'guarantor'} \mid fr) & = & 0.1666 & P(GC = \text{'guarantor'} \mid \neg fr) & = & 0 \\
P(GC = \text{'coapplicant'} \mid fr) & = & 0 & P(GC = \text{'coapplicant'} \mid \neg fr) & = & 0.1429 \\
P(ACC = \text{'own'} \mid fr) & = & 0.6666 & P(ACC = \text{'own'} \mid \neg fr) & = & 0.7857 \\
P(ACC = \text{'rent'} \mid fr) & = & 0.3333 & P(ACC = \text{'rent'} \mid \neg fr) & = & 0.1429 \\
P(ACC = \text{'free'} \mid fr) & = & 0 & P(ACC = \text{'free'} \mid \neg fr) & = & 0.0714
\end{array}
$$

2. Naïve Bayes classifier

- Example 2.

Compute all the conditional probabilities required for NB classification:

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = \text{'none'} \mid fr)$	$=$	0.1666	$P(CH = \text{'none'} \mid \neg fr)$	$=$	0
$P(CH = \text{'paid'} \mid fr)$	$=$	0.1666	$P(CH = \text{'paid'} \mid \neg fr)$	$=$	0.2857
$P(CH = \text{'current'} \mid fr)$	$=$	0.5	$P(CH = \text{'current'} \mid \neg fr)$	$=$	0.2857
$P(CH = \text{'arrears'} \mid fr)$	$=$	0.1666	$P(CH = \text{'arrears'} \mid \neg fr)$	$=$	0.4286
$P(GC = \text{'none'} \mid fr)$	$=$	0.8334	$P(GC = \text{'none'} \mid \neg fr)$	$=$	0.8571
$P(GC = \text{'guarantor'} \mid fr)$	$=$	0.1666	$P(GC = \text{'guarantor'} \mid \neg fr)$	$=$	0
$P(GC = \text{'coapplicant'} \mid fr)$	$=$	0	$P(GC = \text{'coapplicant'} \mid \neg fr)$	$=$	0.1429
$P(ACC = \text{'own'} \mid fr)$	$=$	0.6666	$P(ACC = \text{'own'} \mid \neg fr)$	$=$	0.7857
$P(ACC = \text{'rent'} \mid fr)$	$=$	0.3333	$P(ACC = \text{'rent'} \mid \neg fr)$	$=$	0.1429
$P(ACC = \text{'free'} \mid fr)$	$=$	0	$P(ACC = \text{'free'} \mid \neg fr)$	$=$	0.0714

Query:

- Credit history = paid
- Guarantor/CoApplicant = none
- Accommodation = rent
- Fraud = ?

P(fr) = P(Fraud = true) = 6/20	**P(¬fr) = P(Fraud = false) = 14/20**
P(CH = paid|fr) = 1/6	P(CH=paid|¬fr) = 4/14
P(GC = none|fr) = 5/6	P(GC=none|¬fr) = 12/14
P(ACC=rent|fr) = 2/6	P(ACC=rent|¬fr) = 2/14

2. Naïve Bayes classifier

- Example 2.

Compute all the conditional probabilities required for NB classification:

P(fr) = P(Fraud = true) = 6/20	**P(¬fr) = P(Fraud = false) = 14/20**
P(CH = paid|fr) = 1/6	P(CH=paid|¬fr) = 4/14
P(GC = none|fr) = 5/6	P(GC=none|¬fr) = 12/14
P(ACC=rent|fr) = 2/6	P(ACC=rent|¬fr) = 2/14

- For Fraud = true, compute:

P(CH=paid|fr) × P(GC=none|fr) × P(ACC=rent|fr) × P(fr)

= (1/6) × (5/6) × (2/6) × (6/20) = 0.01388

- For Fraud = false, compute:

P(CH=paid|¬fr) × P(GC=none|¬fr) × P(ACC=rent|¬fr) × P(¬fr)

= (4/14) × (12/14) ×(2/14) × (14/20) = 0.02448

Decision: predict Fraud = false

Maximum a posteriori hypothesis:
Compare: P(**x**|A) P(A) and P(**x**|B) P(B)

3. Naïve Bayes: smoothing

Laplace smoothing

- The assumption of conditional independence extends the coverage of a Naïve Bayes model and allows it to generalize beyond the contents of the training data.

- The model still does not have complete coverage of the set of all possible queries, e.g.

  $P(CH = none \mid \neg fr) = 0.$

$P(fr) = 6/20$	$P(\neg fr) = 14/20$
$P(CH = paid \mid fr) = 1/6$	$P(CH{=}paid \mid \neg fr) = 4/14$
$P(GC = guarantor \mid fr) = 1/6$	$P(GC{=}guarantor \mid \neg fr) = 0$
$P(ACC{=}free \mid fr) = 0$	$P(ACC{=}free \mid \neg fr) = 1/14$

Consider the query:

- Credit history = paid
- Guarantor/CoApplicant = guarantor
- Accommodation = free

- Then

$$P(CH{=}paid \mid fr) \times P(GC{=}guarantor \mid fr) \times P(ACC{=}free \mid fr) \times P(fr) = 0$$

$$P(CH{=}paid \mid \neg fr) \times P(GC{=}guarantor \mid \neg fr) \times P(ACC{=}free \mid \neg fr) \times P(\neg fr) = 0$$

3. Naïve Bayes: smoothing

Laplace smoothing

- Laplace smoothing for conditional probabilities is defined as:

$$P(f = \ell|t) = [\text{count}(f = \ell|t) + k]/[\text{count}(f|t) + (k \times |\text{Domain}(f)|)]$$

where

- count($f = \ell|t$): how often the event $f = \ell$ occurs when the target level is ℓ
- count($f|t$): how often the feature f took any level in the subset of data when the target level is ℓ
- $|\text{Domain}(f)|$: the number of levels in the domain of the feature
- k is a predetermined parameter. Larger values of k mean more smoothing occurs, more probability taken from larger probabilities to the small probabilities.

3. Naïve Bayes: smoothing

Laplace smoothing

- Raw probabilities:
  - P(GC = none | ¬fr) = 12/14 , P(GC=guarantor | ¬fr) = 0
  - P(GC=coapplicant | ¬fr) = 2/14
- Smoothing:
  - k = 3
  - count(GC | ¬fr) = 14
  - count(GC = none | ¬fr) = 12
  - count(GC = guarantor | ¬fr) = 0
  - count(GC = coapplicant | ¬fr) = 2
  - |Domain(GC)| = 3
- Smoothed probabilities:
  - P(GC = none | ¬fr) = (12 + 3)/[14 + (3 × 3)] = 15/23 = 0.6522
  - **P(GC=guarantor | ¬fr) = (0 + 3)/[14 + (3 × 3)] = 3/23 = 0.1304**
  - P(GC=coapplicant | ¬fr) = (2 + 3)/[14 + (3 × 3)] = 5/23 = 0.2174

$P(GC = \text{'none'} \mid fr)$	=	0.8334	$P(GC = \text{'none'} \mid \neg fr)$	=	0.8571
$P(GC = \text{'guarantor'} \mid fr)$	=	0.1666	$P(GC = \text{'guarantor'} \mid \neg fr)$	=	0
$P(GC = \text{'coapplicant'} \mid fr)$	=	0	$P(GC = \text{'coapplicant'} \mid \neg fr)$	=	0.1429

P(f = ℓ | t) =

[count(f = ℓ | t) + k]/[count(f | t) + (k × |Domain(f)|)]

k = 3

Domain(f) = {none, guarantor, coapplicant)

|Domain(f)| = 3

Before: (12,0,2)
After: (15,3,5)

3. Naïve Bayes: smoothing

Laplace smoothing

- Before smoothing the original conditional probabilities are:

P(fr) = 6/20	**P(¬fr) = 14/20**
P(CH = paid|fr) = 1/6	P(CH=paid|¬fr) = 4/14
P(GC = guarantor|fr) =1/6	P(GC=guarantor|¬fr) = 0
P(ACC=free|fr) = 0	P(ACC=free|¬fr) = 1/14

Consider the query:

- Credit history = paid
- Guarantor/CoApplicant = guarantor
- Accommodation = free

- After smoothing:

P(fr) = 6/20	**P(¬fr) = 14/20**
P(CH = paid|fr) = (1 + 3)/(6 + 3×4) = 4/18	P(CH=paid|¬fr) = (4 + 3)/(14 + 3×4) = 7/26
P(GC = guarantor|fr) = (1 + 3)/(6 + 3×3) = 4/15	**P(GC=guarantor|¬fr) = (0 + 3)/(14 + 3×3) = 3/23**
P(ACC=free|fr) = (0 + 3)/(6 + 3×3) = 3/15	P(ACC=free|¬fr) = (1 + 3)/(14 + 3×3) = 4/23

- For Fraud = true, compute:

  P(CH=paid|fr) × P(GC=guarantor|fr) × P(ACC=free|fr) × P(fr) = (4/18) × (4/15) × (3/15) × (6/20) = 0.0036

- For Fraud = false, compute:

  P(CH=paid|¬fr) × P(GC=guarantor|¬fr) × P(ACC=free|¬fr) × P(¬fr) = (7/26) × (3/23) ×(4/23) × (14/20) = 0.0043 ⇨ Not fraud

4. Naïve Bayes: continuous inputs

Continuous features: Probability density function

- A continuous feature can have an infinite number of values in its domain.
- Any particular value will occur a negligible amount of time, indistinguishable from 0 in a large dataset.
- Check how the probability of a continuous feature taking a value across the range of values it can take.
- A probability density function (PDF) represents the probability distribution of a continuous feature using a mathematical function.
- Some well known standard probability functions: normal, student t, exponential, mixture of Gaussians distributions.
- We need to select which probability distribution function best fits the distribution of the values of the feature. This is often done by plotting a density histogram, choose distribution best matches the shape of the histogram to model the feature. Finally, fit the parameters of the selected distribution to the feature values in the data set.

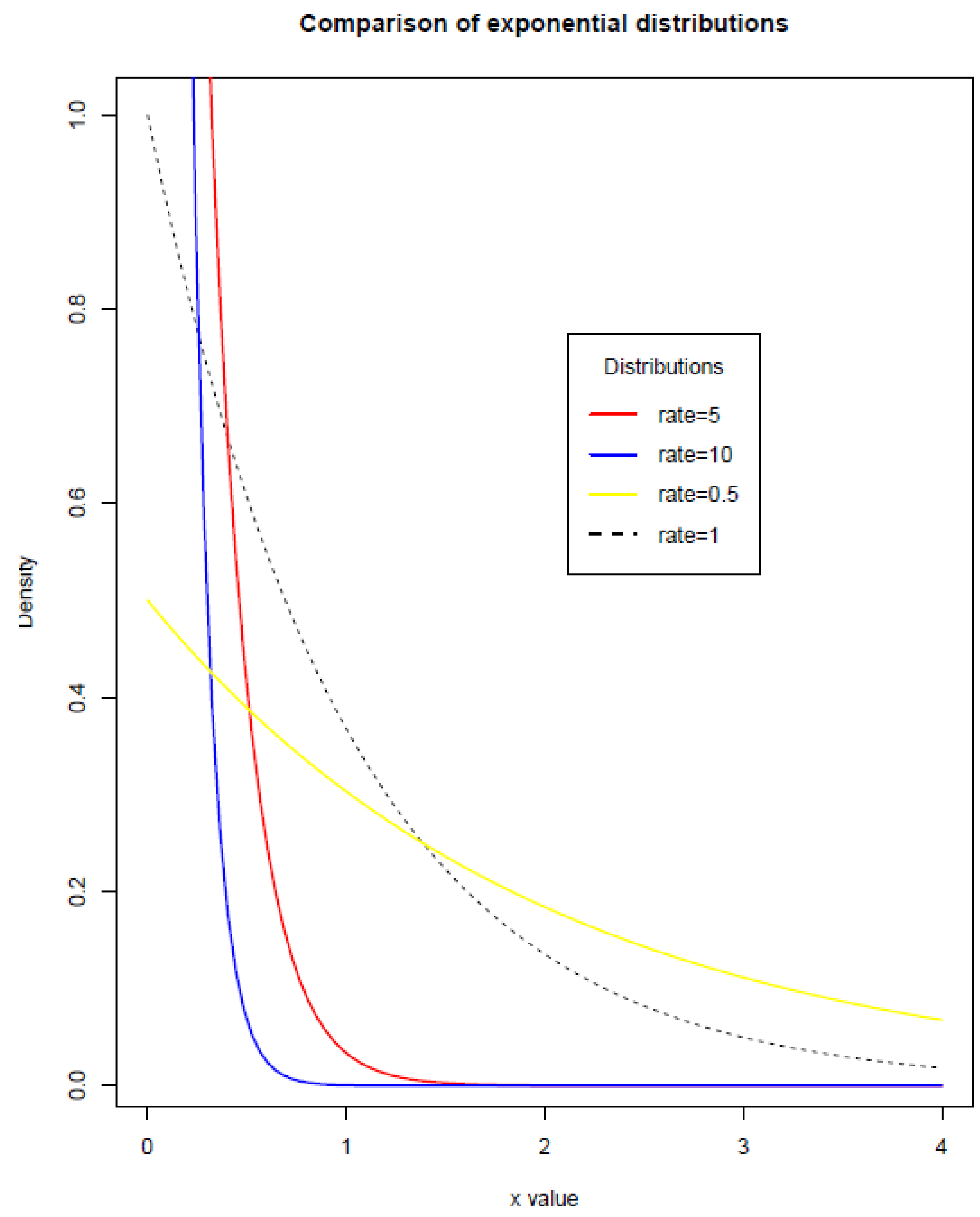
4. Naïve Bayes: continuous inputs

Continuous features: Exponential distribution density function

- For x > 0, the **exponential distribution** is defined by the probability distribution function:

  $E(x,\lambda) = \lambda\ e^{-\lambda x}$

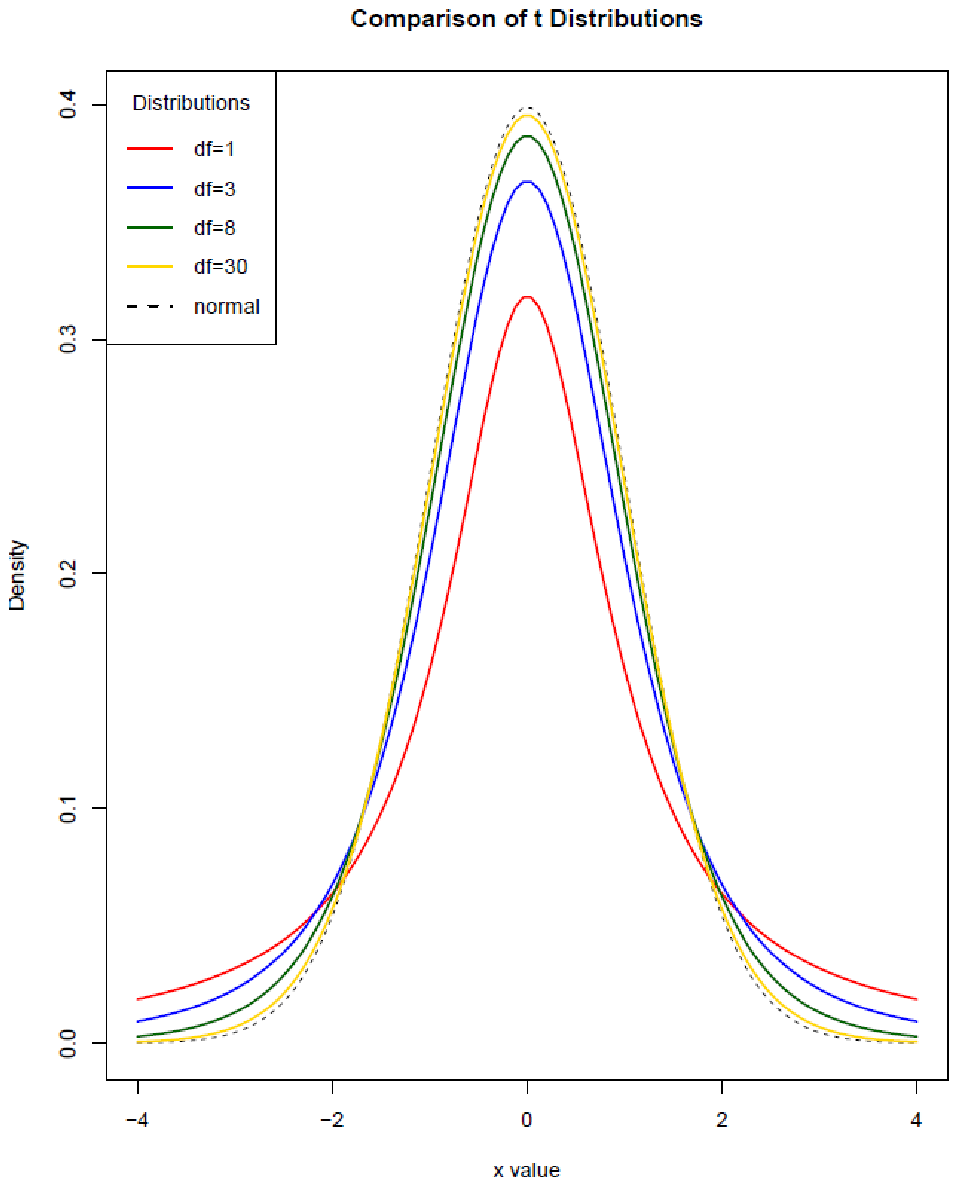
- It takes one parameter, λ known as rate.
- As λ gets larger, the peak of the distribution (on the left) gets larger and the drop-off in density gets steeper.
- To fit an exponential distribution to a continuous feature, we set λ equal to 1 divided by the mean of the feature.



4. Naïve Bayes: continuous inputs

Continuous features: Student t-distribution density function

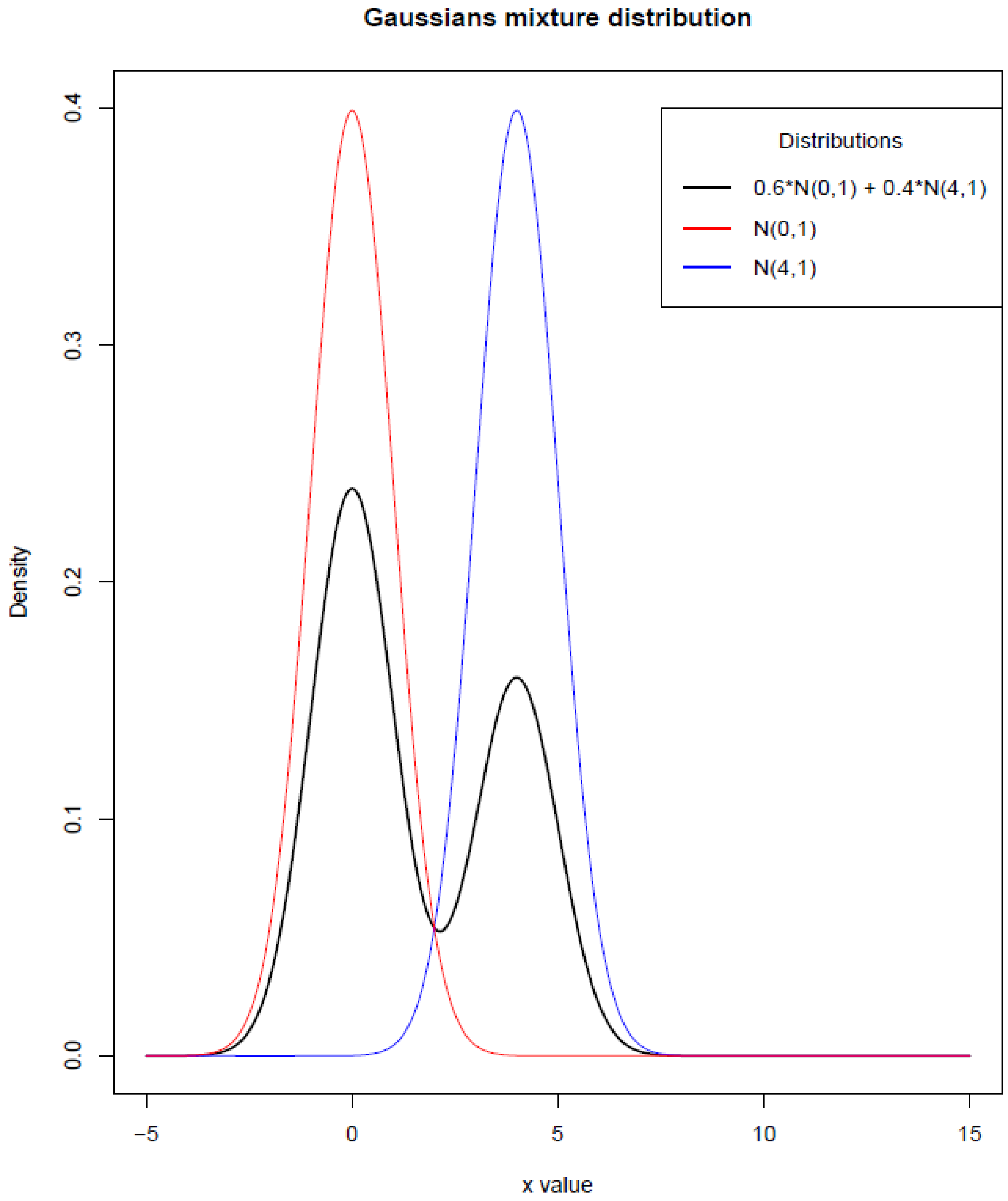
- The **student t-distribution** is symmetric around its peak.
- The parameters in a student-t distribution function:
  - ϕ: specifies the location of the peak
  - ρ: how spread out the distribution is.
  - κ: the degrees of freedom, the number of variables in the calculation of the statistic that are free to vary. For student t-distribution, df = the sample size – 1.
  - Try fitting the unimodal continuous feature using the normal distribution first. If it is not a good fit, consider the student-t distribution.



4. Naïve Bayes: continuous inputs

Continuous features: Gaussian mixture distribution density function

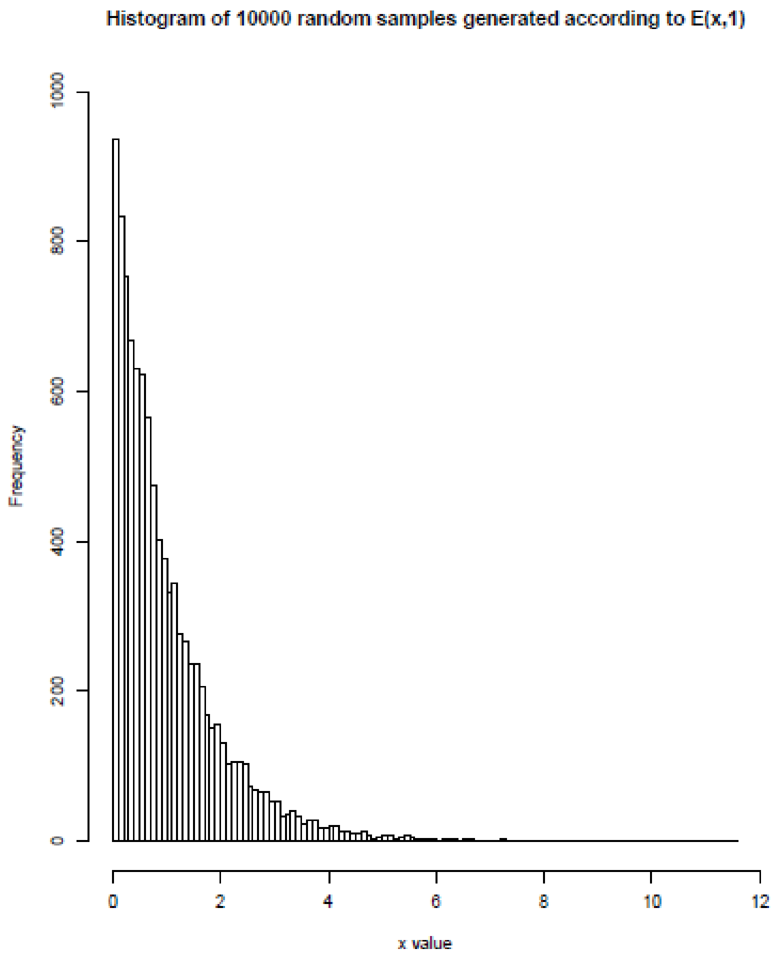
- The mixture of **Gaussians distribution** is the distribution that results when a number of Gaussian (normal) distribution are merged.
- It is used to represent data that is composed of multiple subpopulations.
- **Multimodal distribution**: the multiple peaks in the density curve arise from the different subpopulations.
- A mixture of Gaussians distribution is defined by 3 parameters for each component: a mean μ, a standard deviation σ and a weight ω. The sum of all weights must be equal to 1.

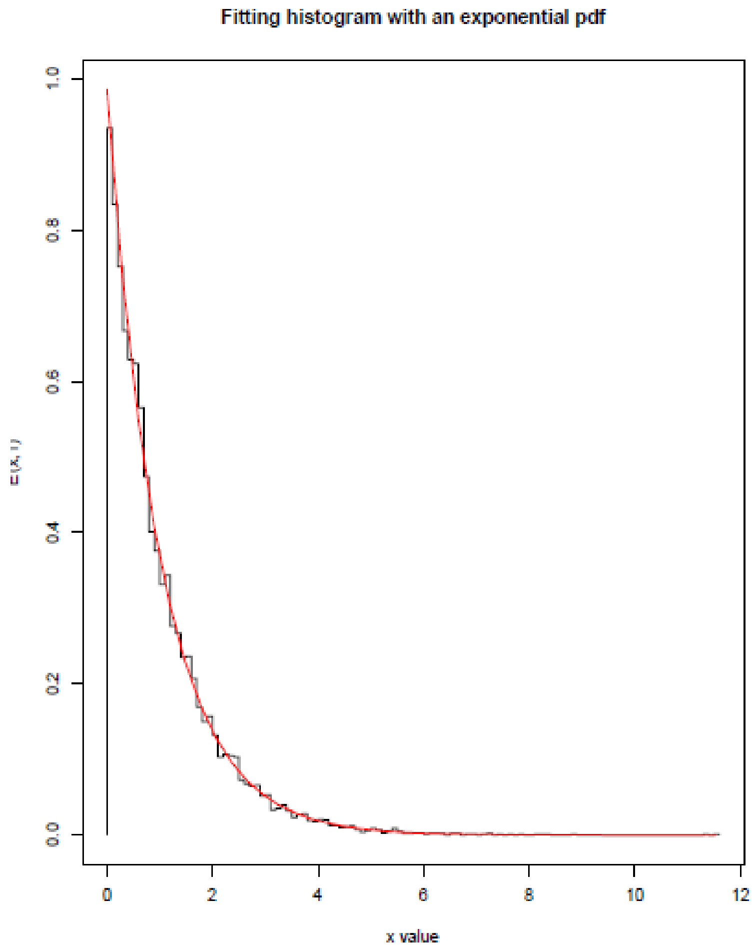


4. Naïve Bayes: continuous inputs

Continuous features: Probability density function

- Approximating the pdf from data:

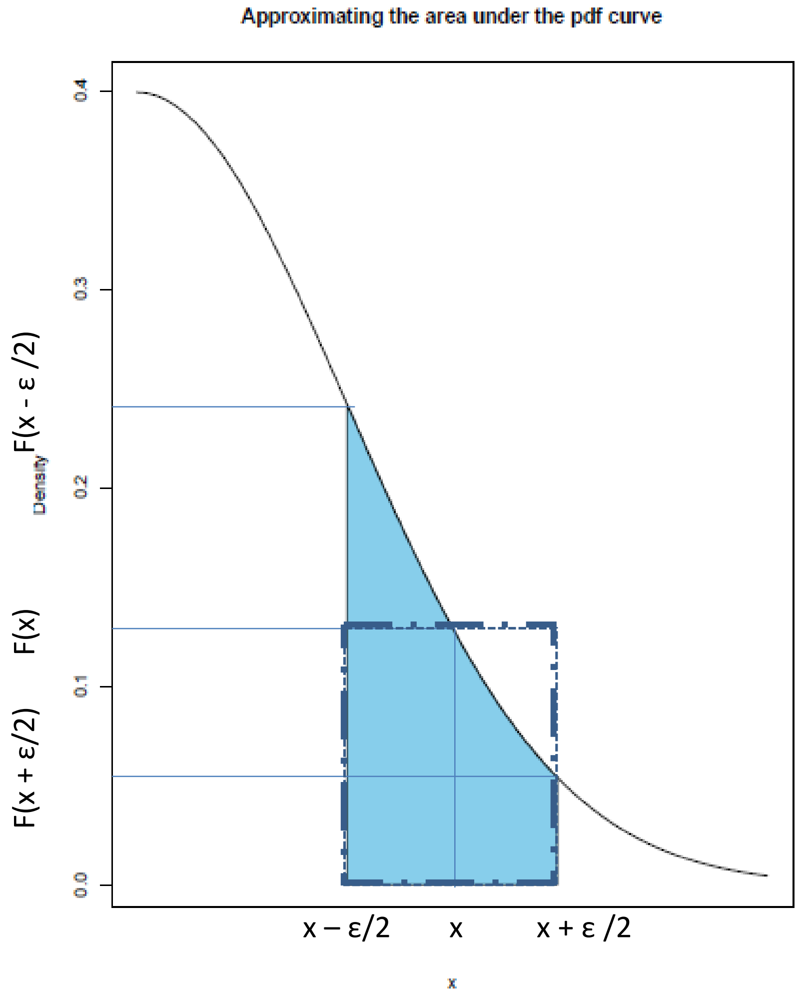




4. Naïve Bayes: continuous inputs

Continuous features: Probability density function

- Calculating a probability with a PDF



- The area under the curve where x is from x – ε/2 to x + ε /2 is approximated as the area of the dotted rectangle F(x) × ε
- The interval size ε is made on a case by case basis. For example, if the feature is 'temperature', the interval size may be 1 degree. If it is a financial feature, intervals may represent cents.
- In Naïve Bayes model, we <u>do not need</u> to actually compute the exact probability.
- We only need to calculate the <u>relative likelihood</u> of a continuous feature given different levels of target feature.
- The actual probability <u>need not be computed</u>, we can just make use of the <u>height</u> of the density curve defined by the PDF.

4. Naïve Bayes: continuous inputs

Continuous features: Probability density function

- Example: Account balance (ACB) is a new continuous descriptive feature. What is the prediction for the query:

  CR = paid, GC = guarantor, ACC = free, ACB = 759.70?

ID	Credit History	Guarantor/ CoApplicant	Accommodation	Account Balance	Fraud
1	current	none	own	56.75	true
2	current	none	own	1,800.11	false
3	current	none	own	1,341.03	false
4	paid	guarantor	rent	749.50	true
5	arrears	none	own	1,150.00	false
6	arrears	none	own	928.30	true
7	current	none	own	250.90	false
8	arrears	none	own	806.15	false
9	current	none	rent	1,209.02	false
10	none	none	own	405.72	true
11	current	coapplicant	own	550.00	false
12	current	none	free	223.89	true
13	current	none	rent	103.23	true
14	paid	none	own	758.22	false
15	arrears	none	own	430.79	false
16	current	none	own	675.11	false
17	arrears	coapplicant	rent	1,657.20	false
18	arrears	none	free	1,405.18	false
19	arrears	none	own	760.51	false
20	current	none	own	985.41	false

- Account balance is a new continuous descriptive feature. What is the prediction for the query: CR = paid, GC = guarantor, ACC = free, ACB = 759.70?
- First partition data into two groups according to target levels: fraud and not fraud.
- For the feature Account Balance (ACB), we define two PDFs:

  $$P(AB = X|fr) = PDF1(AB = X|fr)$$

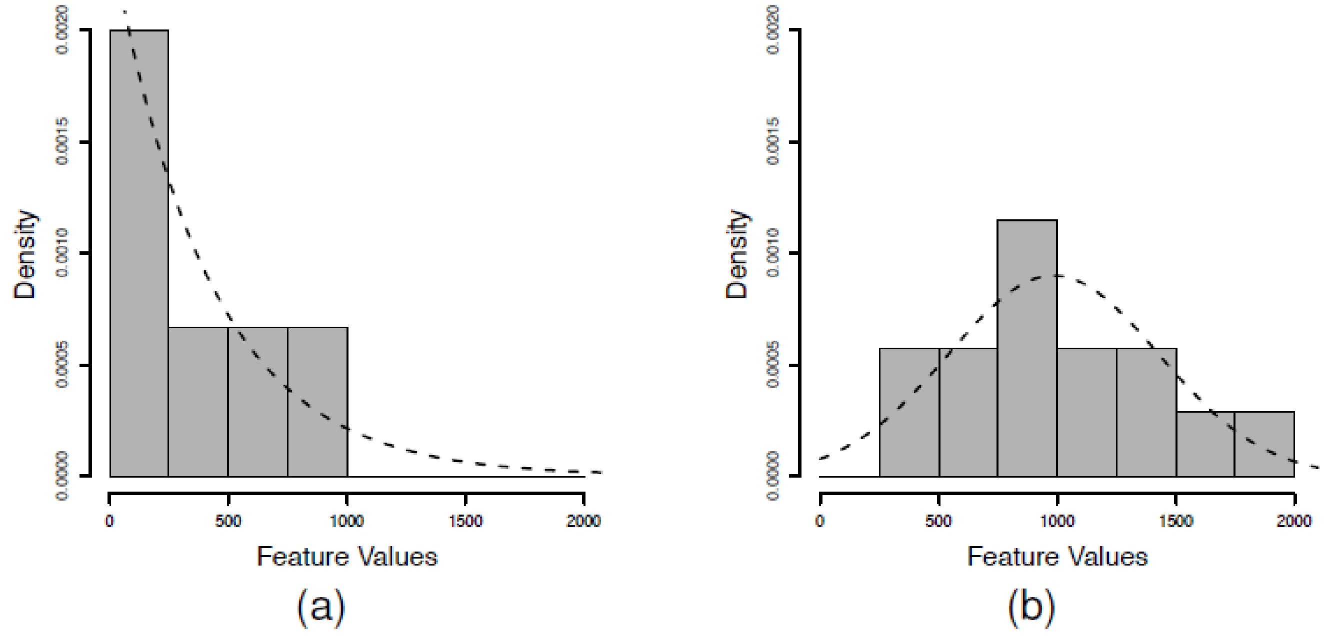
  $$P(AB = X|\neg fr) = PDF2(AB=X|\neg fr)$$

- These two distributions do not have to be the same.

4. Naïve Bayes: continuous inputs

Continuous features: Probability density function

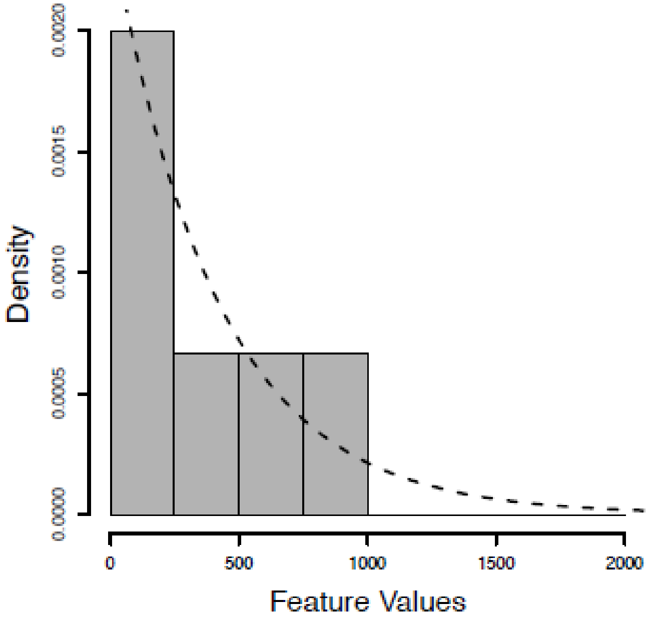
- Example: Account balance is a new continuous descriptive feature. What is the prediction for the query: CR = paid, GC = guarantor, ACC = free, ACB = 759.70?
- Histograms for the two subsets of data:



Bin size = 250
(a) Fraud cases, fitted with exponential distribution
(b) Non fraud cases, fitted with normal distribution

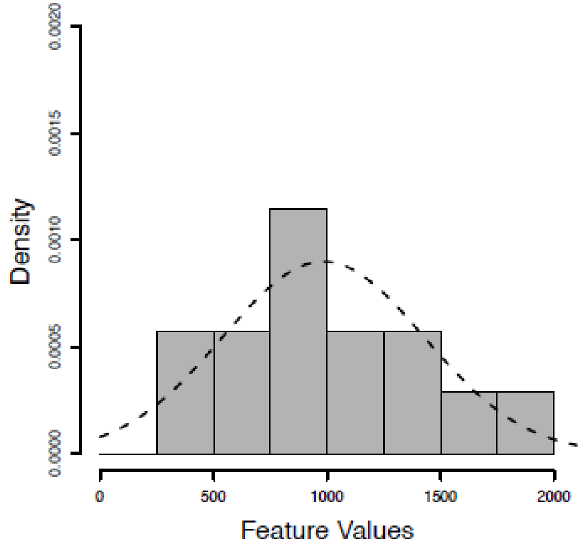
4. Naïve Bayes: continuous inputs

Continuous features: Probability density function



(a)

ID	...	Account Balance	Fraud
1		56.75	true
4		749.50	true
6		928.30	true
10	...	405.72	true
12		223.89	true
13		103.23	true
$\overline{AB}$		411.22	
$\lambda = 1/\overline{AB}$		0.0024	



ID	...	Account Balance	Fraud
2		1 800.11	false
3		1 341.03	false
5		1 150.00	false
7		250.90	false
8		806.15	false
9		1 209.02	false
11		550.00	false
14		758.22	false
15		430.79	false
16		675.11	false
17		1 657.20	false
18		1 405.18	false
19		760.51	false
20		985.41	false
$\overline{AB}$		984.26	
$sd(AB)$		460.94	

4. Naïve Bayes: continuous inputs

Continuous features: Probability density function

Table: The Laplace smoothed (with $k = 3$) probabilities needed by a naive Bayes prediction model calculated from the dataset in Table 5 [23], extended to include the conditional probabilities for the new ACCOUNT BALANCE feature, which are defined in terms of PDFs.

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = none \mid fr)$	$=$	0.2222	$P(CH = none \mid \neg fr)$	$=$	0.1154
$P(CH = paid \mid fr)$	$=$	0.2222	$P(CH = paid \mid \neg fr)$	$=$	0.2692
$P(CH = current \mid fr)$	$=$	0.3333	$P(CH = current \mid \neg fr)$	$=$	0.2692
$P(CH = arrears \mid fr)$	$=$	0.2222	$P(CH = arrears \mid \neg fr)$	$=$	0.3462
$P(GC = none \mid fr)$	$=$	0.5333	$P(GC = none \mid \neg fr)$	$=$	0.6522
$P(GC = guarantor \mid fr)$	$=$	0.2667	$P(GC = guarantor \mid \neg fr)$	$=$	0.1304
$P(GC = coapplicant \mid fr)$	$=$	0.2	$P(GC = coapplicant \mid \neg fr)$	$=$	0.2174
$P(ACC = own \mid fr)$	$=$	0.4667	$P(ACC = own \mid \neg fr)$	$=$	0.6087
$P(ACC = rent \mid fr)$	$=$	0.3333	$P(ACC = rent \mid \neg fr)$	$=$	0.2174
$P(ACC = free \mid fr)$	$=$	0.2	$P(ACC = free \mid \neg fr)$	$=$	0.1739
$P(AB = x \mid fr)$			$P(AB = x \mid \neg fr)$		
	$\approx$	$E\left(\begin{matrix} x, \\ \lambda = 0.0024 \end{matrix}\right)$		$\approx$	$N\left(\begin{matrix} x, \\ \mu = 984.26, \\ \sigma = 460.94 \end{matrix}\right)$

4. Naïve Bayes: continuous inputs

Continuous features: Probability density function

Table: The probabilities, from Table 7 [29], needed by the naive Bayes prediction model to make a prediction for the query $\langle \text{CH} = \textit{'paid'}, \text{GC} = \textit{'guarantor'}, \text{ACC} = \textit{'free'}, \text{AB} = 759.07 \rangle$ and the calculation of the scores for each candidate prediction.

$$
\begin{array}{rcl@{\qquad}rcl}
P(fr) & = & 0.3 & P(\neg fr) & = & 0.7 \\
P(CH = paid \mid fr) & = & 0.2222 & P(CH = paid \mid \neg fr) & = & 0.2692 \\
P(GC = guarantor \mid fr) & = & 0.2667 & P(GC = guarantor \mid \neg fr) & = & 0.1304 \\
P(ACC = free \mid fr) & = & 0.2 & P(ACC = free \mid \neg fr) & = & 0.1739 \\
P(AB = 759.07 \mid fr) & & & P(AB = 759.07 \mid \neg fr) & & \\
\approx E\begin{pmatrix} 759.07, \\ \lambda = 0.0024 \end{pmatrix} & = & 0.00039 & \approx N\begin{pmatrix} 759.07, \\ \mu = 984.26, \\ \sigma = 460.94 \end{pmatrix} & = & 0.00077
\end{array}
$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid fr)\right) \times P(fr) = 0.0000014$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid \neg fr)\right) \times P(\neg fr) = 0.0000033$$

Predict Fraud = false

4. Naïve Bayes: continuous inputs

Continuous features: Binning

- An alternative way to representing a continuous feature using a PDF is to convert the feature into categorical features using:
  - Equal width binning may result in bins with large number of instances and other bins empty.
  - Equal frequency binning is recommended for building Bayesian model.
- Dataset with additional continuous features:

ID	Credit History	Guarantor/ CoApplicant	Accommodation	Account Balance	Loan Amount	Fraud
1	current	none	own	56.75	900	true
2	current	none	own	1 800.11	150 000	false
3	current	none	own	1 341.03	48 000	false
4	paid	guarantor	rent	749.50	10 000	true
5	arrears	none	own	1 150.00	32 000	false
6	arrears	none	own	928.30	250 000	true
7	current	none	own	250.90	25 000	false
8	arrears	none	own	806.15	18 500	false
9	current	none	rent	1 209.02	20 000	false
10	none	none	own	405.72	9 500	true
11	current	coapplicant	own	550.00	16 750	false
12	current	none	free	223.89	9 850	true
13	current	none	rent	103.23	95 500	true
14	paid	none	own	758.22	65 000	false
15	arrears	none	own	430.79	500	false
16	current	none	own	675.11	16 000	false
17	arrears	coapplicant	rent	1 657.20	15 450	false
18	arrears	none	free	1 405.18	50 000	false
19	arrears	none	own	760.51	500	false
20	current	none	own	985.41	35 000	false

4. Naïve Bayes: continuous inputs

Continuous features: Probability density function

- The continuous feature Loan Amount is discretized using equal frequency bins:

ID	Loan Amount	Binned Loan Amount	Fraud	ID	Loan Amount	Binned Loan Amount	Fraud
15	500	bin1	false	9	20,000	bin3	false
19	500	bin1	false	7	25,000	bin3	false
1	900	bin1	true	5	32,000	bin3	false
10	9,500	bin1	true	20	35,000	bin3	false
12	9,850	bin1	true	3	48,000	bin3	false
4	10,000	bin2	true	18	50,000	bin4	false
17	15,450	bin2	false	14	65,000	bin4	false
16	16,000	bin2	false	13	95,500	bin4	true
11	16,750	bin2	false	2	150,000	bin4	false
8	18,500	bin2	false	6	250,000	bin4	true

- The corresponding thresholds:

Bin Thresholds		
	Bin1	$\leq 9,925$
$9,925 <$	Bin2	$\leq 19,250$
$19,225 <$	Bin3	$\leq 49,000$
$49,000 <$	Bin4	

4. Naïve Bayes: continuous inputs

Continuous features: Binning

- Query: CH = paid, GC = guarantor, ACC = free, AB = 759.07, LA = 8000
- Loan amount of 8000 ≤ 9925 will be placed in bin1.

ID	Loan Amount	Binned loan amount	Fraud
1	900	Bin1	True
10	9500	Bin1	True
12	9850	Bin1	True
4	10000	Bin2	True
13	95500	Bin4	True
6	250000	Bin4	True

Conditional probabilities:

- P(BLA = bin1|fr) = 3/6
- P(BLA = bin2|fr) = 1/6
- P(BLA = bin3|fr) = 0
- P(BLA = bin4|fr) = 2/6

Laplace smoothed (k=3) probabilities:

- P(BLA = bin1|fr) = (3 + 3)/(6 + 4× 3) = 0.3333
- P(BLA = bin2|fr) = (1 + 3)/(6 + 4× 3) = 0.2222
- P(BLA = bin3|fr) = (0 + 3)/(6 + 4× 3) = 0.1667
- P(BLA = bin4|fr) = (2 + 3)/(6 + 4× 3) = 0.2778

4. Naïve Bayes: continuous inputs

Continuous features: Binning

- Query: CH = paid, GC = guarantor, ACC = free, AB = 759.07, LA = 8000
- Loan amount of 8000 ≤ 9925 will be placed in bin1.

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = none \mid fr)$	$=$	0.2222	$P(CH = none \mid \neg fr)$	$=$	0.1154
$P(CH = paid \mid fr)$	$=$	0.2222	$P(CH = paid \mid \neg fr)$	$=$	0.2692
$P(CH = current \mid fr)$	$=$	0.3333	$P(CH = current \mid \neg fr)$	$=$	0.2692
$P(CH = arrears \mid fr)$	$=$	0.2222	$P(CH = arrears \mid \neg fr)$	$=$	0.3462
$P(GC = none \mid fr)$	$=$	0.5333	$P(GC = none \mid \neg fr)$	$=$	0.6522
$P(GC = guarantor \mid fr)$	$=$	0.2667	$P(GC = guarantor \mid \neg fr)$	$=$	0.1304
$P(GC = coapplicant \mid fr)$	$=$	0.2	$P(GC = coapplicant \mid \neg fr)$	$=$	0.2174
$P(ACC = own \mid fr)$	$=$	0.4667	$P(ACC = own \mid \neg fr)$	$=$	0.6087
$P(ACC = rent \mid fr)$	$=$	0.3333	$P(ACC = rent \mid \neg fr)$	$=$	0.2174
$P(ACC = free \mid fr)$	$=$	0.2	$P(ACC = free \mid \neg fr)$	$=$	0.1739
$P(AB = x \mid fr)$			$P(AB = x \mid \neg fr)$		
$\approx E\left(\begin{array}{c} x, \\ \lambda = 0.0024 \end{array}\right)$			$\approx N\left(\begin{array}{c} x, \\ \mu = 984.26, \\ \sigma = 460.94 \end{array}\right)$		
$P(BLA = bin1 \mid fr)$	$=$	0.3333	$P(BLA = bin1 \mid \neg fr)$	$=$	0.1923
$P(BLA = bin2 \mid fr)$	$=$	0.2222	$P(BLA = bin2 \mid \neg fr)$	$=$	0.2692
$P(BLA = bin3 \mid fr)$	$=$	0.1667	$P(BLA = bin3 \mid \neg fr)$	$=$	0.3077
$P(BLA = bin4 \mid fr)$	$=$	0.2778	$P(BLA = bin4 \mid \neg fr)$	$=$	0.2308

4. Naïve Bayes: continuous inputs

Continuous features: Binning

- Query: CH = paid, GC = guarantor, ACC = free, AB = 759.07, LA = 8000 = bin1

$$
\begin{array}{rclrcl}
P(fr) & = & 0.3 & P(\neg fr) & = & 0.7 \\
P(CH = paid \mid fr) & = & 0.2222 & P(CH = paid \mid \neg fr) & = & 0.2692 \\
P(GC = guarantor \mid fr) & = & 0.2667 & P(GC = guarantor \mid \neg fr) & = & 0.1304 \\
P(ACC = free \mid fr) & = & 0.2 & P(ACC = free \mid \neg fr) & = & 0.1739 \\
P(AB = 759.07 \mid fr) & & & P(AB = 759.07 \mid \neg fr) & & \\
\approx E\left(\begin{array}{c} 759.07, \\ \lambda = 0.0024 \end{array}\right) & = & 0.00039 & \approx N\left(\begin{array}{c} 759.07, \\ \mu = 984.26, \\ \sigma = 460.94 \end{array}\right) & = & 0.00077 \\
P(BLA = bin1 \mid fr) & = & 0.3333 & P(BLA = bin1 \mid \neg fr) & = & 0.1923
\end{array}
$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid fr)\right) \times P(fr) = 0.000000462$$

$$\left(\prod_{k=1}^{n} P(\mathbf{q}[k] \mid \neg fr)\right) \times P(\neg fr) = 0.000000633$$

Predict Fraud = false

References:

Mitchell, T. Machine Learning, Chapter 6, McGraw Hill, 1997.

Kelleher, J.D., Mac Namee, B., D'Arcy, A. Machine Learning for Predictive Data Analytics, Chapter 6, MIT Press, 2015.