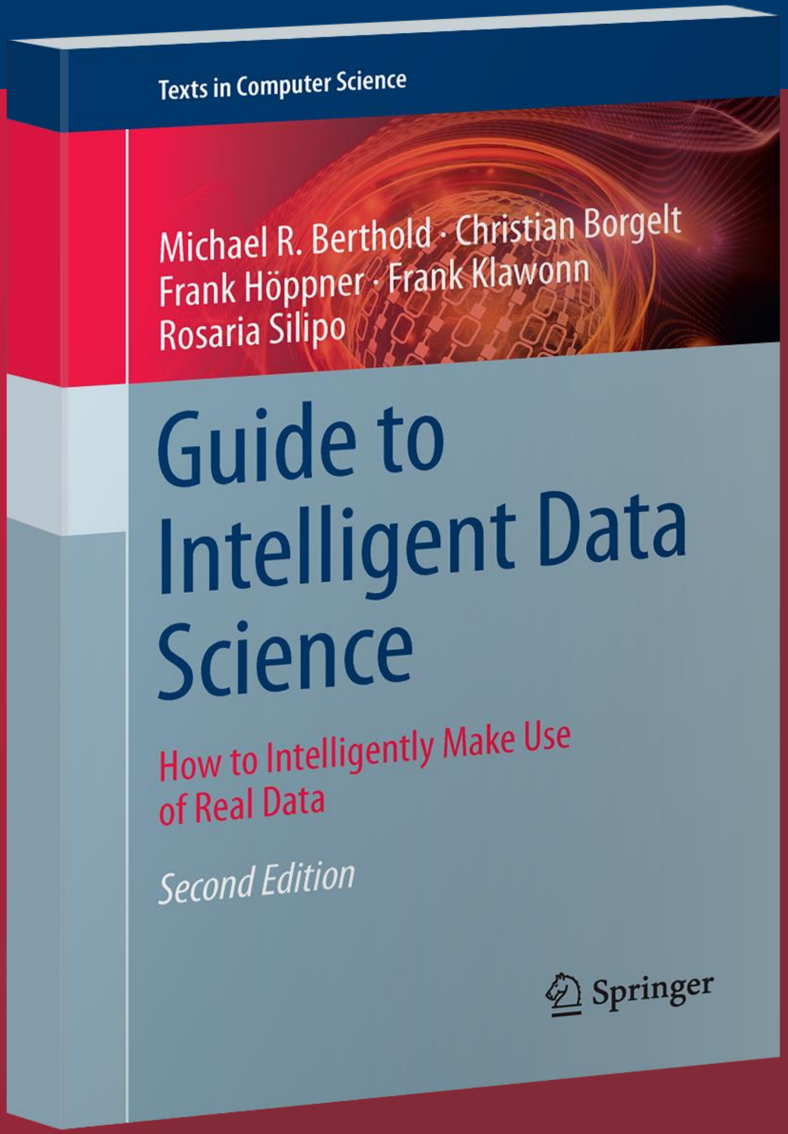
Bayes Classifiers



“Science is the systematic classification of experience”
-George Henry Lewes

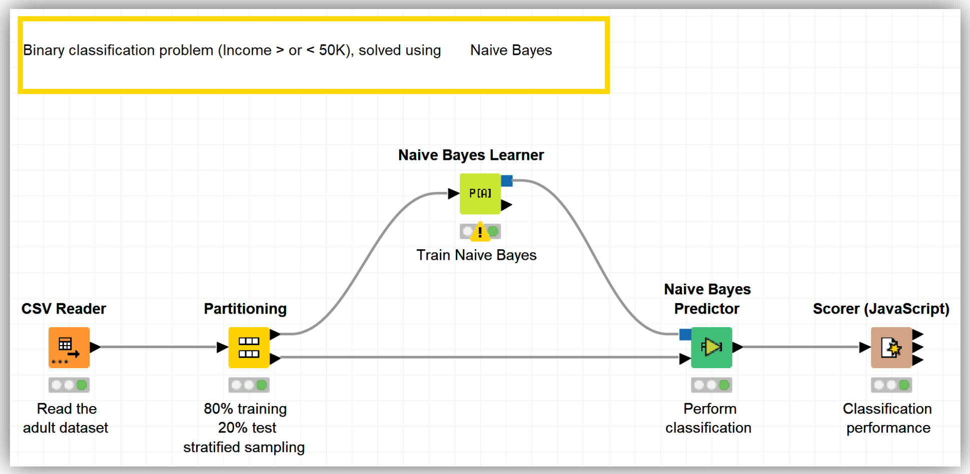
What is the simplest classifier?

**This lesson refers to chapter 8 of the GIDS book*

Content of this lesson

- Bayes Classifiers
  - Motivation
  - Naive Bayes classifiers
  - Full Bayes classifiers
  - Naive vs. Full Bayes classifiers

- Datasets used : adult dataset
- Example Workflows:
  - „Naive Bayes“ https://kni.me/w/0oyhMdWYK5w19xGj
    - Naive Bayes classifier



Bayes Classifiers

Given data $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) | i = 1,2, \dots, n\}$

$\boldsymbol{x}_i$: Object description

Y_i: Target attribute

- Instead of finding structure in a data set, let's focus on (unknow) dependency among attributes
- Bayes classifiers express their model as simple probabilities
- Can be used as a gold standard for evaluating other learning methods

➔ Any model should perform the same or better than a Naïve Bayes classifier

- The conditional probability $P(h|E)$, hypothesis h is true given event E

$$P(h|E) = \frac{P(E|h) \cdot P(h)}{P(E)}$$

- $P(h)$: Probability of hypothesis h
- $P(E)$: Probability of event E
- $P(E|h)$: Conditional probability of event E given hypothesis h

– We want the most probable hypothesis $h \in H$ for a given event E

➔ ***Maximum a posteriori hypothesis*** (***MAP***):

$$h_{MAP} = \arg\max_{h \in H} P(h|E)$$

$$= \arg\max_{h \in H} \frac{P(E|h) \cdot P(h)}{P(E)} = \arg\max_{h \in H} P(E|h) \cdot P(h)$$

Maximum Likelihood Hypothesis

- If we can assume that every hypothesis $h \in H$ is equally likely
- In other words, $P(h_i) = P(h_j)$ for all $h_i, h_j \in H$
- Then we can get the **maximum likelihood hypothesis**

$$h_{ML} = \arg\max_{h \in H} P(E|h)$$

Naïve Bayes Classifiers

- Probability $P(h)$ can be estimated based given data $\mathcal{D}$

$$P(h) = \frac{\#\ class\ h}{\#\ total}$$

- Probability $P(E|h)$ can be determined based on attributes $A_1, A_2, \cdots, A_m$ being $E = (a_1, a_2, \cdots, a_m)$

$$P(E|h) = \frac{\#\ class\ h\ with\ attributes(a_1, a_2, \cdots, a_m)}{\#\ class\ h}$$

Problem:

- Not all combinations of $A_1, A_2, \cdots, A_m$ may be observed
  - For 10 nominal attributes with 3 possible values for each attribute, there are $3^{10} = 59049$ possible combinations!

Solution:

- Naïve, unrealistic assumption that attributes are independent given the class

$$P(E = (a_1, a_2, \cdots, a_m)|h) = P(a_1|h) \cdot \cdots \cdot P(a_1|h) = \prod_{a_i \in E} P(a_i|h)$$

- Where $P(a_i|h)$ can be computed easily as

$$P(a_i|h) = \frac{\#\, class\ h\ with\ A_i = a_i}{\#\, class\ h}$$

Given a data set with only *nominal* attributes

For attributes $E = (a_1, a_2, \cdots, a_m)$, the predicted class $h \in H$ is derived:

- Compute the likelihood $L(h|E)$ under the assumption that $A_1, A_2, \cdots, A_m$ are independent given the class

$$L(h|E) = \prod_{a_i \in E} P(a_i|h) \cdot P(h)$$

- Assign E to the class $h \in H$ with the highest likelihood

$$pred(E) = \arg\max_{h \in H} L(E|h)$$

- This classifier is called ***naïve*** because of the conditional independence assumption among $A_1, A_2, \cdots, A_m$
- Needless to say, this is an unrealistic assumption in most cases
- But a naïve Bayes classifier often yields good results
- Especially when not too many attributes are correlated

Example: Naïve Bayes Classifier

Example

Given the dataset $\mathcal{D}$:

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

we want to predict the sex ($\underline{m}ale$ or $\underline{f}emale$) of a person $\mathbf{x}$ with the following attribute values:

$$\mathbf{x} = (\text{Height} = \underline{t}all, \text{Weight} = \underline{l}ow, \text{Long hair} = \underline{y}es)$$

Example

We need to calculate

$$L(\text{Sex} = m | \text{Height} = t, \text{Weight} = l, \ \text{Long hair} = y)$$

$$\begin{aligned} = \quad & P(\text{Height} = t | \text{Sex} = m) \cdot \\ & P(\text{Weight} = l | \text{Sex} = m) \cdot \\ & P(\text{Long hair} = y | \text{Sex} = m) \cdot \\ & P(\text{Sex} = m) \end{aligned}$$

and

$$L(\text{Sex} = f | \text{Height} = t, \text{Weight} = l, \ \text{Long hair} = y)$$

$$\begin{aligned} = \quad & P(\text{Height} = t | \text{Sex} = f) \cdot \\ & P(\text{Weight} = l | \text{Sex} = f) \cdot \\ & P(\text{Long hair} = y | \text{Sex} = f) \cdot \\ & P(\text{Sex} = f). \end{aligned}$$

Example

$P(\text{Sex} = m) = 4/10 = 2/5$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example: Naïve Bayes Classifier

Example

$P(\text{Height} = t | \text{Sex} = m) = 2/4 = 1/2$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example: Naïve Bayes Classifier

Example

$$P(\text{Weight} = l | \text{Sex} = m) = 0/4 = 0$$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example: Naïve Bayes Classifier

Example

$P(\text{Long hair} = y | \text{Sex} = m) = 0/4 = 0$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example: Naïve Bayes Classifier

Example

$$L(\text{Sex} = m | \text{Height} = t, \text{Weight} = l,\ \text{Long hair} = y)$$

$$= \frac{2}{4} \cdot \frac{0}{4} \cdot \frac{0}{4} \cdot \frac{4}{10} = \frac{1}{2} \cdot 0 \cdot 0 \cdot \frac{2}{5} = 0$$

$\Rightarrow$ the likelihood of person $\mathbf{x}$ being a men is 0.

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example

$P(\text{Sex} = f) = 6/10 = 3/5$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example: Naïve Bayes Classifier

Example

$P(\text{Height} = t | \text{Sex} = f) = 1/6$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example

$$P(\text{Weight} = l | \text{Sex} = f) = 3/6 = 1/2$$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	g	n	n	m

Example

$$P(\text{Long hair} = y | \text{Sex} = f) = 4/6 = 2/3$$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example: Naïve Bayes Classifier

Example

$$L(\text{Sex} = f \mid \text{Height} = t, \text{Weight} = l,\ \text{Long hair} = y)$$

$$= \frac{1}{6} \cdot \frac{3}{6} \cdot \frac{4}{6} \cdot \frac{6}{10} = \frac{1}{6} \cdot \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{5} = \frac{1}{30} > 0$$

$\Rightarrow$ the likelihood of person $\mathbf{x}$ being a female is $\frac{1}{30}$.

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example

$$L(\text{Sex} = f | \text{Height} = t, \text{Weight} = l,\ \text{Long hair} = y) = \frac{1}{30}$$

$$L(\text{Sex} = m | \text{Height} = t, \text{Weight} = l,\ \text{Long hair} = y) = 0$$

Classification of person

$$\mathbf{x} = (\text{Height} = \underline{t}all, \text{Weight} = \underline{l}ow, \text{Long hair} = \underline{y}es)$$

as female (f).

Notice

The data set $\mathcal{D}$ does not contain any object with this combination of values.

$\Rightarrow$ A full Bayes classifier would not be able to classify this object.

Example: Naïve Bayes Classifier

- The object (m, n, n) is classified as m although the data sets contains two such objects, one from class m and one from class f.
- The main impact comes from the attribute *Long hair* = n, having probability 1 in class m, but a low probability in class f.

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Input	$L(m\|\ldots)$	$L(f\|\ldots)$	Class
(m, n, n)	$\frac{1}{4} \cdot \frac{2}{4} \cdot \frac{4}{4} \cdot \frac{4}{10} = \frac{1}{20}$	$\frac{2}{6} \cdot \frac{3}{6} \cdot \frac{2}{6} \cdot \frac{6}{10} = \frac{1}{30}$	m

Example: Naïve Bayes Classifier

- The object (t, h, y) cannot be classified since the likelihood is zero for both classes

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Input	$L(m\|\ldots)$	$L(f\|\ldots)$	Class
(t,h,n)	$\frac{2}{4}\cdot\frac{2}{4}\cdot\frac{4}{4}\cdot\frac{4}{10}=\frac{1}{10}$	$\frac{1}{6}\cdot\frac{0}{6}\cdot\frac{2}{6}\cdot\frac{6}{10}=0$	m
(t,h,y)	$\frac{2}{4}\cdot\frac{2}{4}\cdot\frac{0}{4}\cdot\frac{4}{10}=0$	$\frac{1}{6}\cdot\frac{0}{6}\cdot\frac{4}{6}\cdot\frac{6}{10}=0$	?

- If a single likelihood is zero, then the overall likelihood is zero automatically, even then when the other likelihoods are high

Input	$L(m\|\ldots)$	$L(f\|\ldots)$	Class
(t,h,y)	$\frac{2}{4}\cdot\frac{2}{4}\cdot\frac{0}{4}\cdot\frac{4}{10}=0$	$\frac{1}{6}\cdot\frac{0}{6}\cdot\frac{4}{6}\cdot\frac{6}{10}=0$	?

- Solution: **Laplace correction γ**

$$P(y)=\frac{n_y}{n}\Longrightarrow \hat{P}(y)=\frac{\gamma+n_y}{\gamma\cdot|dom(Y)|+n}$$

$$P(x|y)=\frac{n_{yx}}{n_y}\Longrightarrow \hat{P}(x|y)=\frac{\gamma+n_{yx}}{\gamma\cdot|dom(X)|+n_y}$$

$$P(y)=\frac{n_y}{n}\Rightarrow$$

$$x|y)=\frac{n_{hx}}{n_y}\Rightarrow$$

n	no. of data
n_y	no of data from class y
n_{yx}	no. of data from class y with value x for attribute X
$dom(X)$	no. of distinct values in X

Laplace Correction

Example

Laplace correction for $P(\text{Height} = \ldots | \text{Sex} = m)$ with $\gamma = 1$

$$\hat{P}(s|m) = \frac{\gamma + n_{ms}}{\gamma \cdot |dom(Height)| + n_m} = \frac{1+1}{1 \cdot 3 + 4} = \frac{2}{7}$$

Height	$\#$	$\#_{Laplace}$	P	$\hat{P}$
s	1	2	1/4	2/7
m	1	2	1/4	2/7
t	2	3	2/4	3/7

Notice

- $\gamma = 0$: Maximum likelihood estimation
- Common choices: $\gamma = 1$ or $\gamma = \frac{1}{2}$

Naïve Bayes Classifier: Implementation

- Frequency tables are generated when constructing a naïve Bayes classifier
- Probability distribution of each attribute can be obtained from the frequency table
- To learn from a naïve Bayes classifier, corresponding frequencies are multiplied from the tables

Treatment of Missing Values

- *During learning*: The missing values are simply not counted for the frequencies of the corresponding attribute.

- *During classification*: Only the probabilities (likelihoods) of those attributes are multiplied for which a value is available.

Numerical Attributes

- Assume a normal distribution for a numerical attribute X

$$f(x|y) = \frac{1}{\sqrt{2\pi}\sigma_{X|y}} \exp\left(-\frac{\left(x-\mu_{X|y}\right)^2}{2\sigma_{X|y}^2}\right)$$

- Estimation of the mean value

$$\hat{\mu}_{X|y} = \frac{1}{n_y}\sum_{i=1}^{n} \tau(y_i = y) \cdot \boldsymbol{x}_i[X]$$

- Estimation of the variance

$$\hat{\sigma}_{X|y}^2 = \frac{1}{n_y'}\sum_{i=1}^{n} \tau(y_i = y) \cdot \left(\boldsymbol{x}_i[X] - \hat{\mu}_{X|y}\right)^2$$

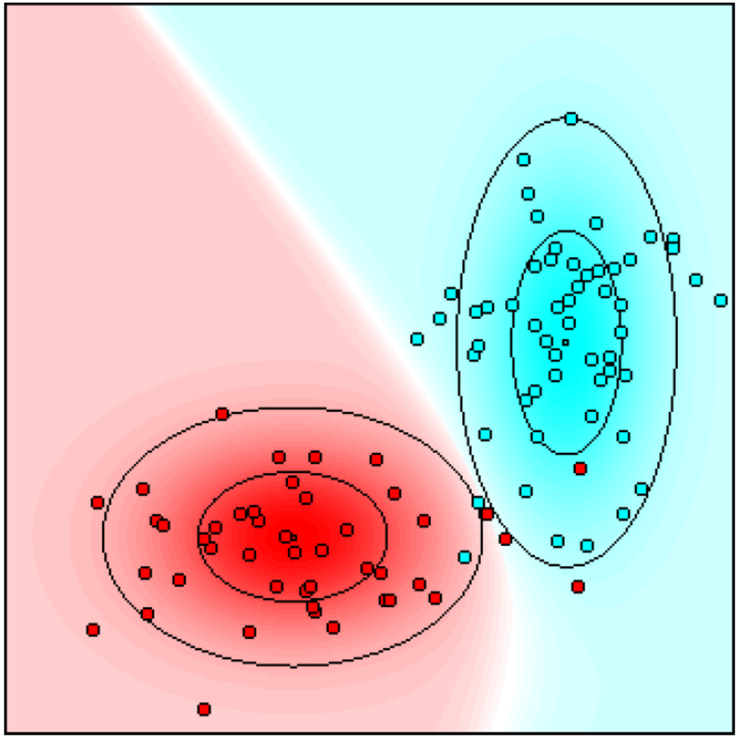
$n_y' = n_y$: Maximum likelihood estimation

$n_y' = n_y - 1$: Unbiased estimation

$$\tau(y_i = y) = \begin{cases} 1 & if\ true \\ 0 & else \end{cases}$$

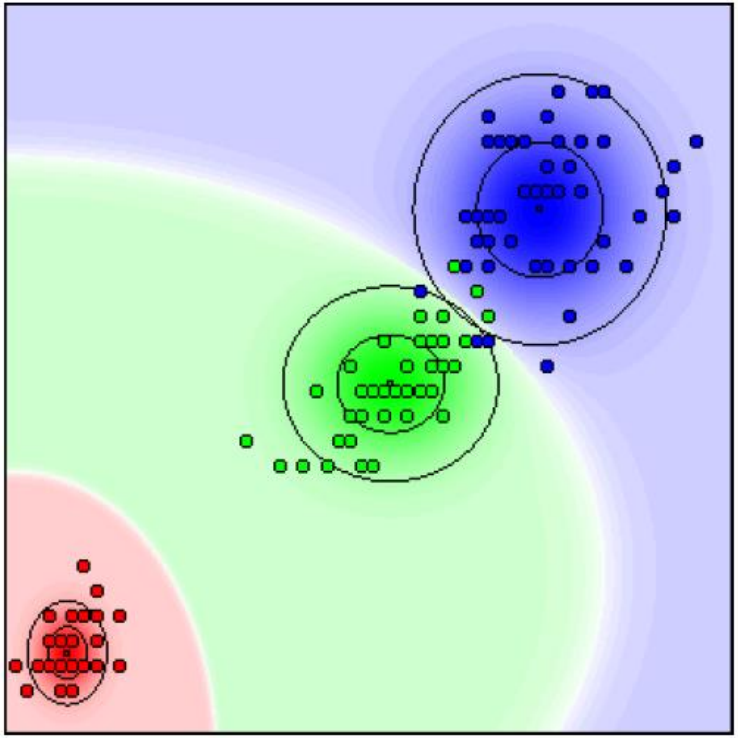
Example: Numerical Attributes

- 100 data points, 2 classes
- Small squares: class mean
- Inner ellipses: 1 s.d. from the mean
- Outer ellipses: 2 s.d. from the mean
- Classes overlap ➔ classification is not perfect



Naïve Bayes classifier

- 150 data points, 3 classes
  - Iris setosa (red)
  - Iris versicolor (green)
  - Iris virginica (blue)
- 4 numerical attributes
  - Sepal length
  - Sepal width
  - Petal length (shown on x-axis)
  - Petal width (shown on y-axis)
- 6 mis-classification on the training data

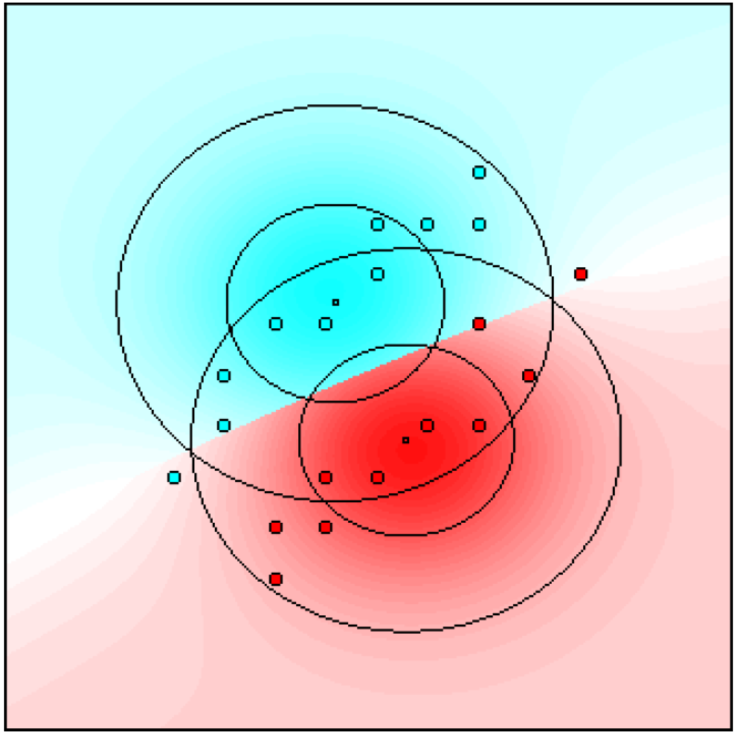


Naïve Bayes classifier

Full Bayes Classifiers

Example: Numerical Attributes

- 20 data points, 2 classes
- Small squares: class mean
- Inner ellipses: 1 s.d. from the mean
- Outer ellipses: 2 s.d. from the mean
- Attributes are not conditionally independent given the class



Naïve Bayes classifier

- Restricted to numeric or metric attributes – only the target is nominal
- Each class can be described by a multivariate normal distribution:

$$f(\boldsymbol{x}_M|y) = \frac{1}{\sqrt{(2\pi)^m \left|\boldsymbol{\Sigma}_{\boldsymbol{X}_M|y}\right|}} \exp\left(-\frac{\left(\boldsymbol{x}_M - \mu_{\boldsymbol{X}_M|y}\right)^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{X}_M|y}^{-1} \left(\boldsymbol{x}_M - \mu_{\boldsymbol{X}_M|y}\right)}{2}\right)$$

$\boldsymbol{X}_M$:	set of metric attributes
$\boldsymbol{x}_M$:	attribute vector
$\mu_{\boldsymbol{X}_M|y}$:	mean value vector for class y
$\boldsymbol{\Sigma}_{\boldsymbol{X}_M|y}$:	covariance matrix for class y

Joint distribution with covariance among attributes

→ Conditional independence no longer holds

- Estimation of the (class-conditional) mean value vector

$$\hat{\mu}_{\boldsymbol{X}|y} = \frac{1}{n_y} \sum_{i=1}^{n} \tau(y_i = y) \cdot \boldsymbol{x}_i[\boldsymbol{X}_M]$$

- Estimation of the (class-conditional) covariance matrix

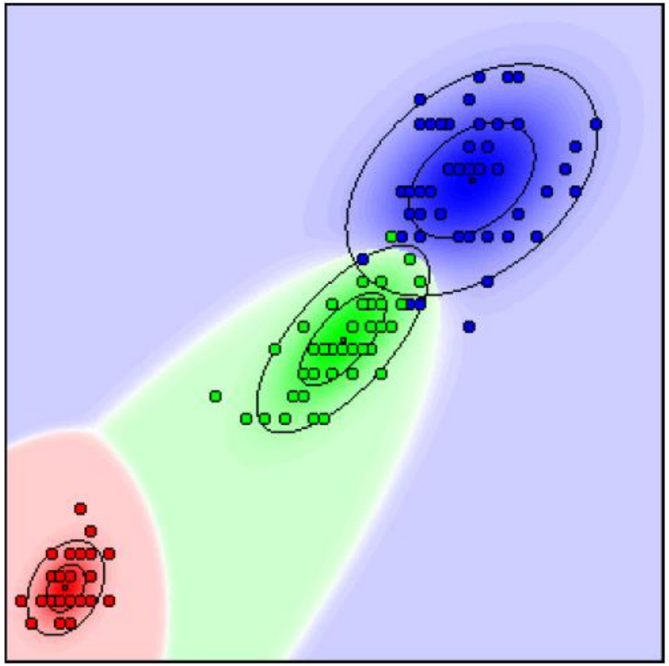
$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}_M|y} = \frac{1}{n'_y} \sum_{i=1}^{n} \tau(y_i = y) \times \left(\boldsymbol{x}_i[\boldsymbol{X}_M] - \hat{\mu}_{\boldsymbol{X}_M|y}\right)\left(\boldsymbol{x}_i[\boldsymbol{X}_M] - \hat{\mu}_{\boldsymbol{X}_M|y}\right)^T$$

$n'_y = n_y$: Maximum likelihood estimation

$n'_y = n_y - 1$: Unbiased estimation

Iris data revisited

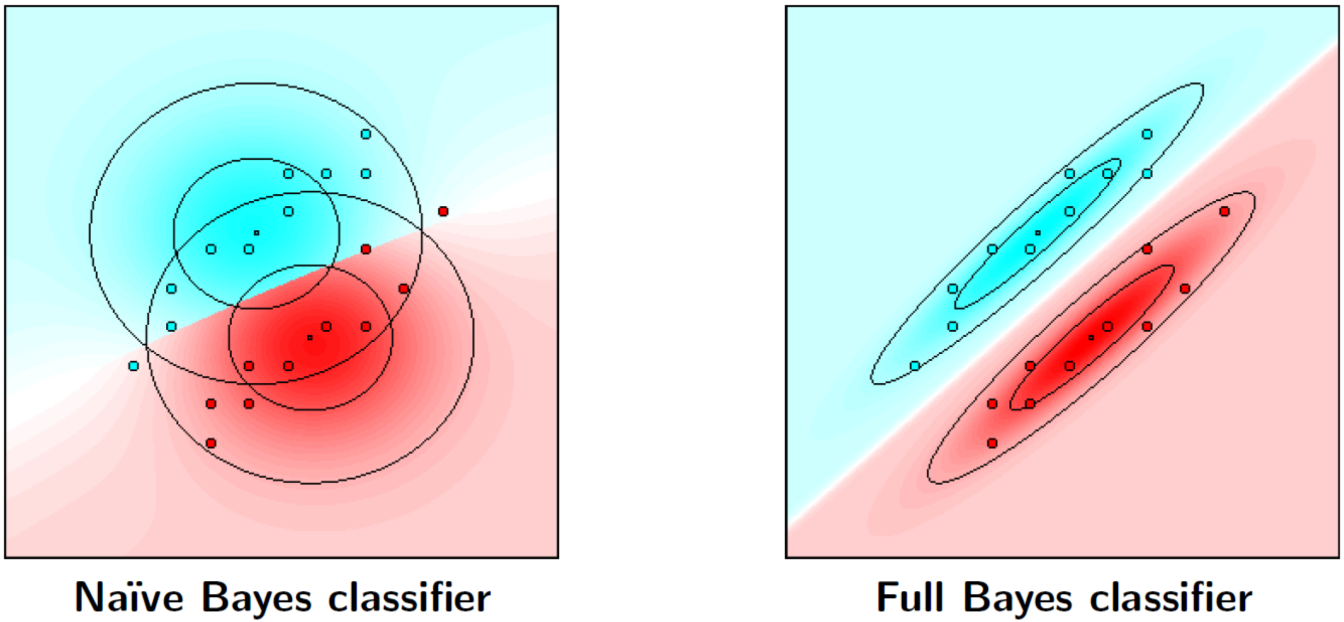
- 150 data points, 3 classes
  - Iris setosa (red)
  - Iris versicolor (green)
  - Iris virginica (blue)
- 4 numerical attributes
  - Sepal length
  - Sepal width
  - Petal length (shown on x-axis)
  - Petal width (shown on y-axis)
- 2 mis-classification on the training data



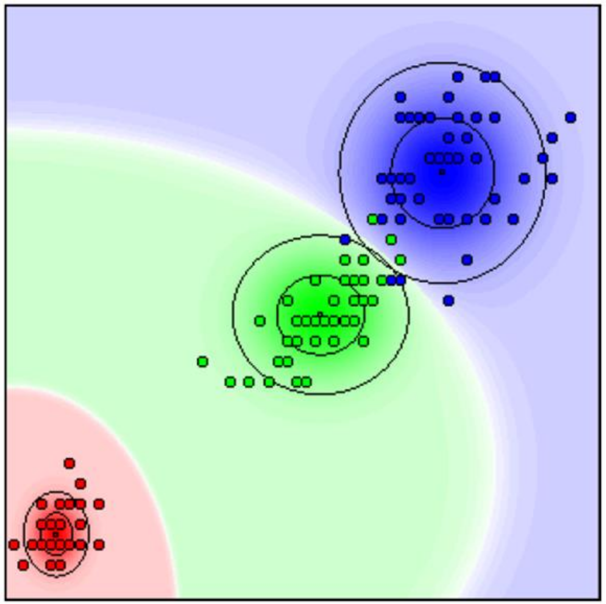
Full Bayes classifier

Naive vs. Full Bayes Classifiers

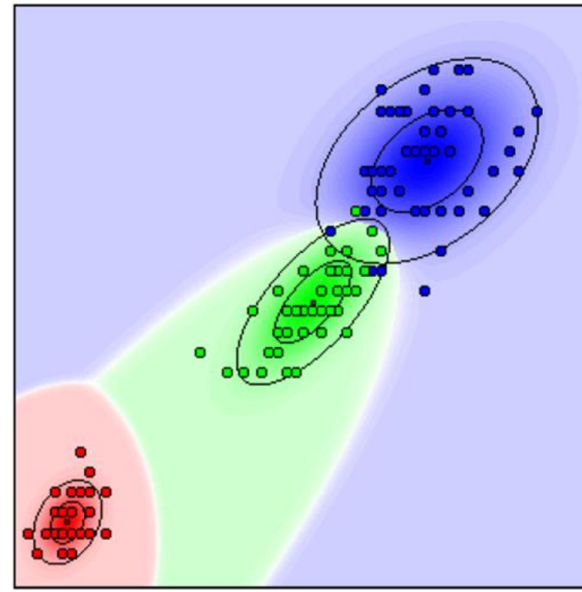
– Naïve Bayes classifiers for numerical data → full Bayes classifiers with diagonal covariance matrices



– Iris data



Naïve Bayes classifier



Full Bayes classifier

Pros:

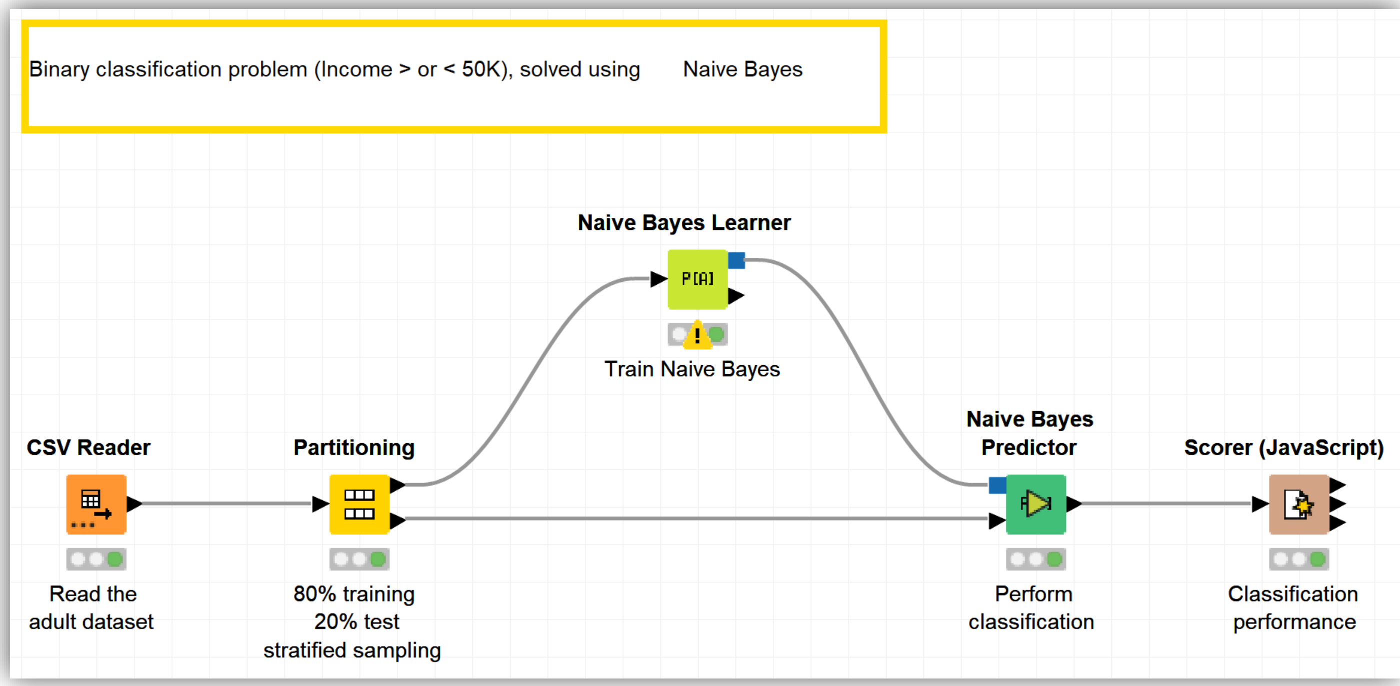
- Gold standard for comparison with other classifiers
- High classification accuracy in many applications
- Classifier can easily be adapted to new training objects
- Integration of domain knowledge

Cons:

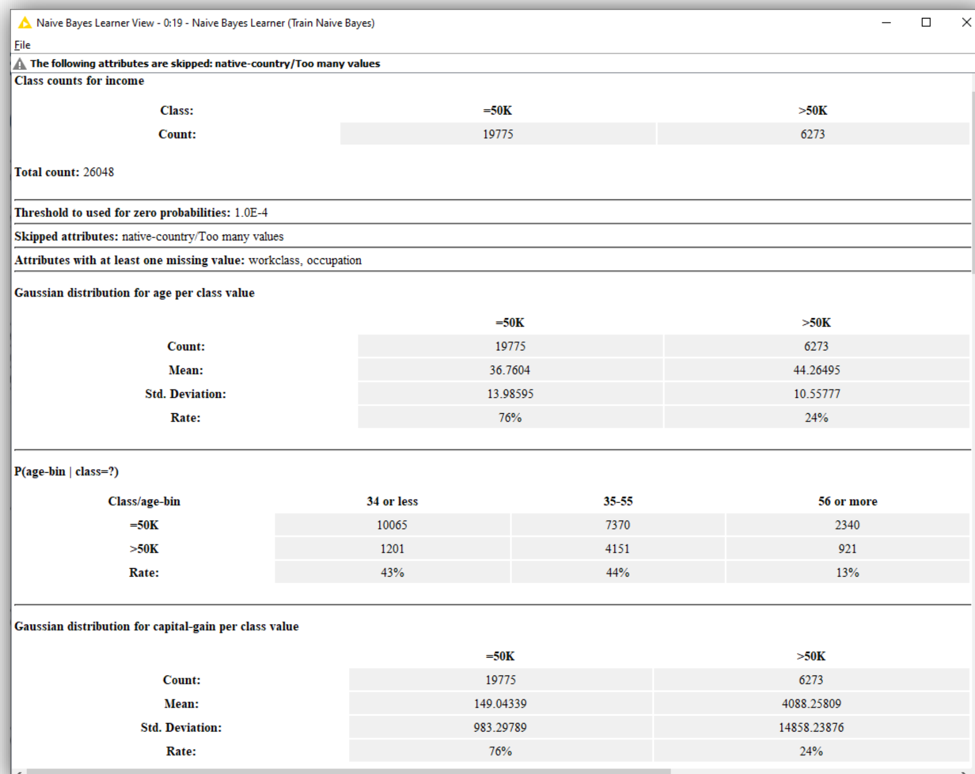
- The conditional probabilities my not be available
- Independence assumptions might not hold for data set

Practical Examples with KNIME Analytics Platform

– Naïve Bayes classification of the income on the adult data



– Naïve Bayes Learner node showing conditional probabilities and distributions involved in the decision process



Naive Bayes Learner View - 0:19 - Naive Bayes Learner (Train Naive Bayes)

File

The following attributes are skipped: native-country/Too many values

Class counts for income

Class:	=50K	>50K
Count:	19775	6273

Total count: 26048

Threshold to used for zero probabilities: 1.0E-4

Skipped attributes: native-country/Too many values

Attributes with at least one missing value: workclass, occupation

Gaussian distribution for age per class value

	=50K	>50K
Count:	19775	6273
Mean:	36.7604	44.26495
Std. Deviation:	13.98595	10.55777
Rate:	76%	24%

P(age-bin | class=?)

Class/age-bin	34 or less	35-55	56 or more
=50K	10065	7370	2340
>50K	1201	4151	921
Rate:	43%	44%	13%

Gaussian distribution for capital-gain per class value

	=50K	>50K
Count:	19775	6273
Mean:	149.04339	4088.25809
Std. Deviation:	983.29789	14858.23876
Rate:	76%	24%

Thank you

For any questions please contact: education@knime.com