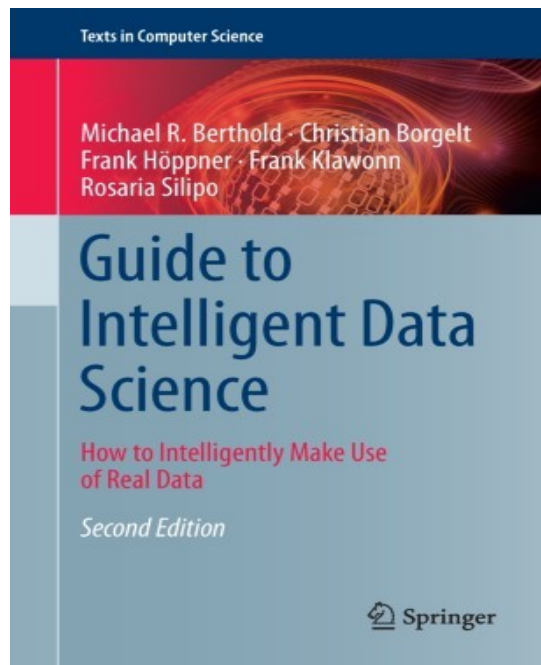




Big Data ACUA6002 Lecture 7

Bayes Classifiers

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

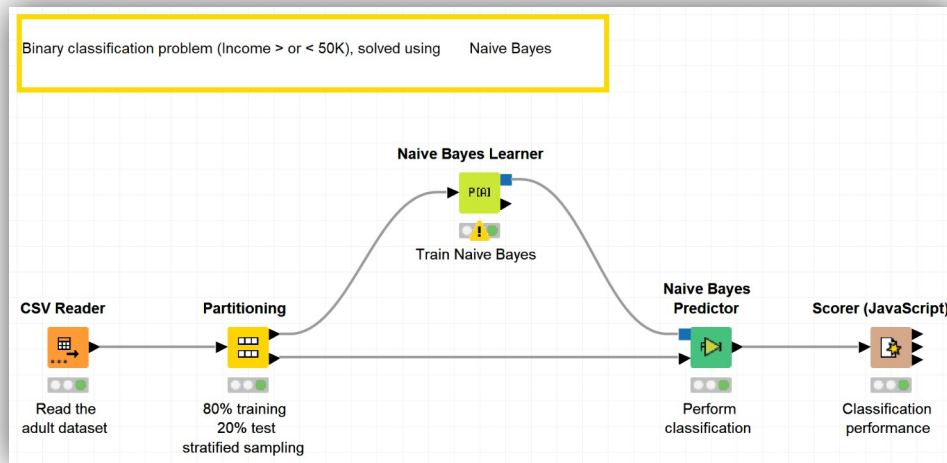
What is the simplest classifier?

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

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

Datasets

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



Bayes Classifiers

Given data $= \{ (\quad , \quad) \mid = 1, 2, \dots, \quad \}$

: Object description

: Target attribute

- Instead of finding structure in a data set, let's focus on (unknown) 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 , hypothesis is true given event

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

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

- We want the most probable hypothesis for a given event

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

$$h = \arg \max_{h \in \mathcal{H}} P(h | \mathcal{D})$$

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

Maximum Likelihood Hypothesis

- If we can assume that every hypothesis is equally likely
- In other words, for all ,
- Then we can get the **maximum likelihood hypothesis**

$$h = \arg \max_{h \in \mathcal{H}} (P(h))$$

Naïve Bayes Classifiers

- Probability can be estimated based given data

$$P(h) = \frac{N(h)}{N}$$

- Probability can be determined based on attributes being

$$P(h | x_1, x_2, \dots, x_n) = \frac{P(h) \prod_{i=1}^n P(x_i | h)}{P(x_1, x_2, \dots, x_n)}$$

Problem:

- Not all combinations of $\mathbf{x}$ may be observed
 - For 10 nominal attributes with 3 possible values for each attribute, there are 3^{10} possible combinations!

Solution:

- Naïve, unrealistic assumption that attributes are independent given the class
- Where $P(\mathbf{x} | h)$ can be computed easily as

$$P(\mathbf{x} | h) = \prod_{i=1}^n \frac{P(x_i | h)}{P(x_i)}$$

Given a data set with only nominal attributes

For attributes , the predicted class is derived:

- Compute the likelihood under the assumption that are independent given the class

=

- Assign E to the class with the highest likelihood

=

- This classifier is called naïve because of the conditional independence assumption among
- 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

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 (male or female) of a person x with the following attribute values:

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

Example

We need to calculate

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

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

and

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

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

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

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

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

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

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

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

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

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

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

⇒ the likelihood of person 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

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

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

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

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

$\Rightarrow$ the likelihood of person 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.

⇒ A full Bayes classifier would not be able to classify this object.

Example: Naïve Bayes Classifier

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

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	$L(m ...)$	$L(f ...)$	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 () 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	$L(m ...)$	$L(f ...)$	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	$L(m \dots)$	$L(f \dots)$	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**

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

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

no. of data

no of data from class

no. of data from class with value for attribute

no. of distinct values in

Example

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

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

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 $\gamma = \frac{1}{2}$

- 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

- *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.

- Assume a normal distribution for a numerical attribute

$$p(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

- Estimation of the mean value

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n (x_i = \mu) \cdot [x_i]$$

- Estimation of the variance

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i = \mu) \cdot ([x_i] - \hat{\mu})^2$$

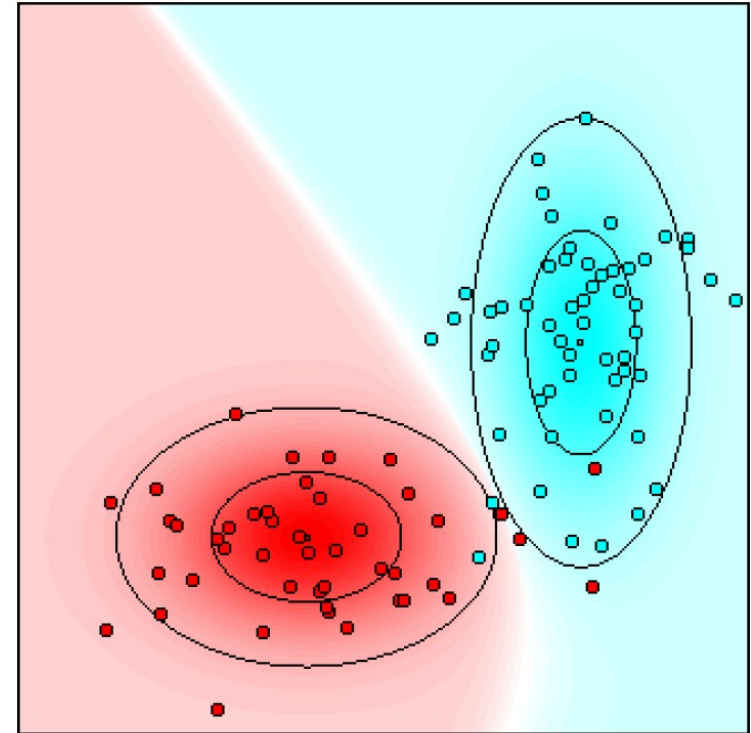
: Maximum likelihood estimation

: Unbiased estimation

$$(x_i = \mu) = \begin{cases} 1 & \text{if } x_i = \mu \\ 0 & \text{otherwise} \end{cases}$$

Example: Numerical Attributes

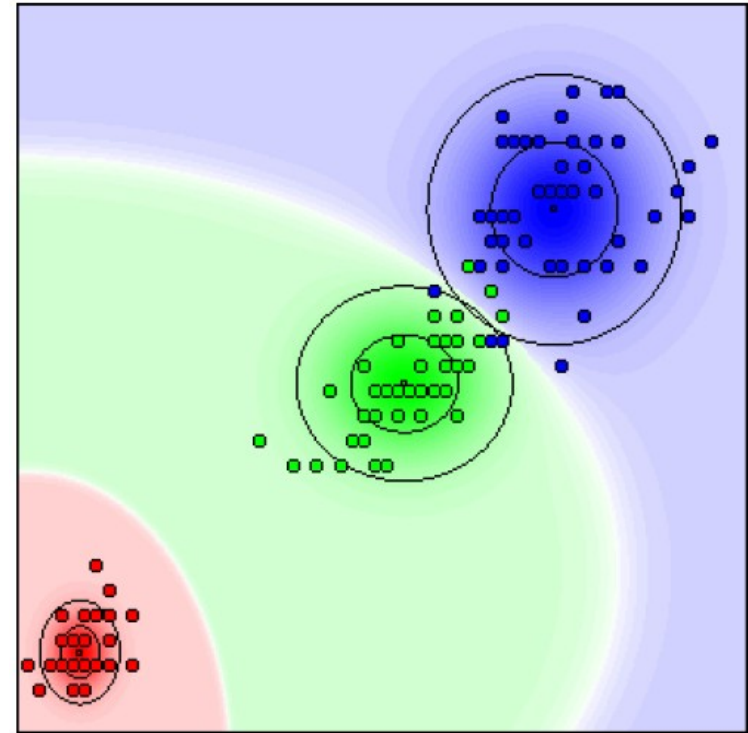
- 100 data points, 2 classes
- Small squares: class mean
- Inner ellipses: 1 s.d. from the mean
- Outer ellipses: 2 s.d. from the mean
- Classes overlap → classification is not perfect



Naïve Bayes classifier

Naïve Bayes Classifier: Iris Data

- 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

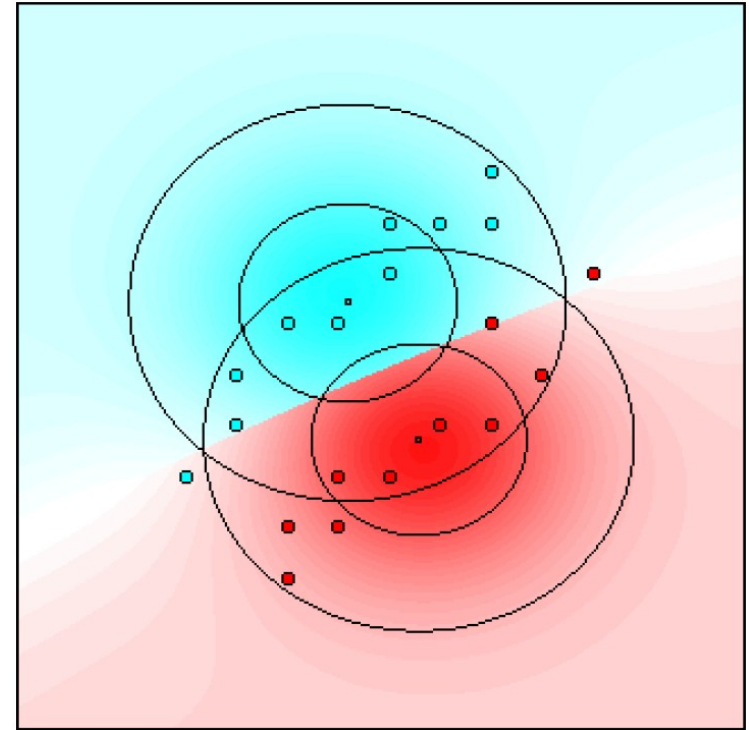


Naïve Bayes classifier

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:

$$p(\mathbf{v}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{(\mathbf{v} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{v} - \boldsymbol{\mu})}{2}\right)$$

- : set of metric attributes
- : attribute vector
- : mean value vector for class
- : covariance matrix for class

Joint distribution with covariance among attributes

→ Conditional independence no longer holds

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

$$\hat{\mu}_v = \frac{1}{N_v} \sum_{i=1}^{N_v} (\mathbf{x}_i = v) \cdot \mathbf{x}_i$$

- Estimation of the (class-conditional) covariance matrix

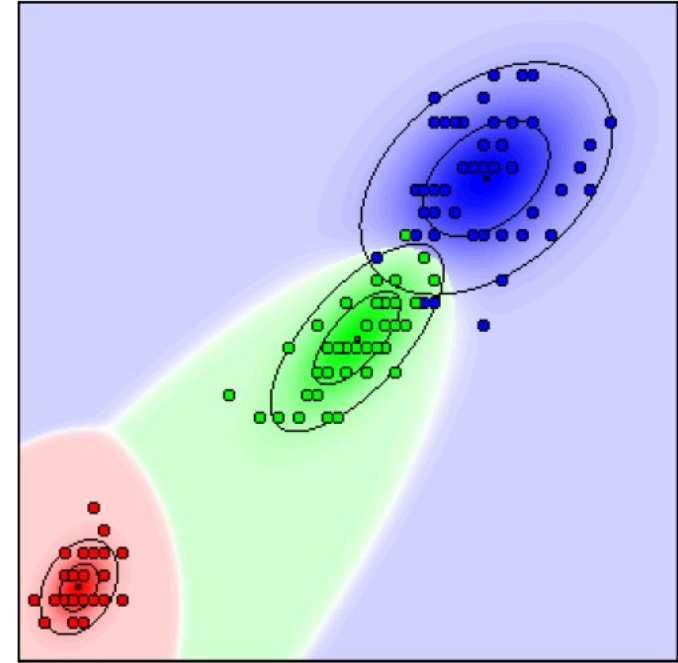
$$\hat{\Sigma}_v = \frac{1}{N_v} \sum_{i=1}^{N_v} (\mathbf{x}_i = v) \times (\mathbf{x}_i - \hat{\mu}_v)(\mathbf{x}_i - \hat{\mu}_v)^T$$

: Maximum likelihood estimation

: 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

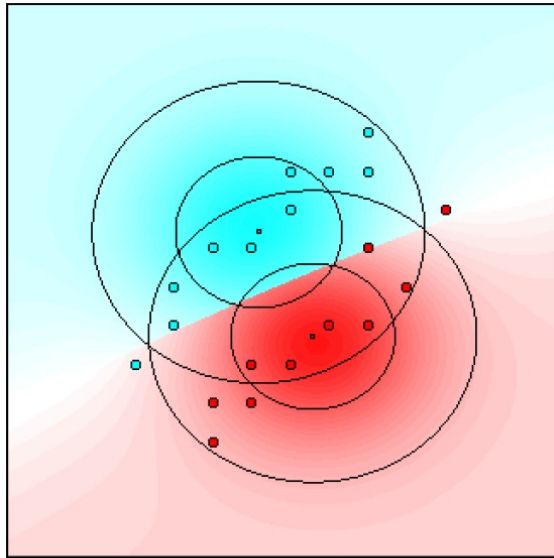


Full Bayes classifier

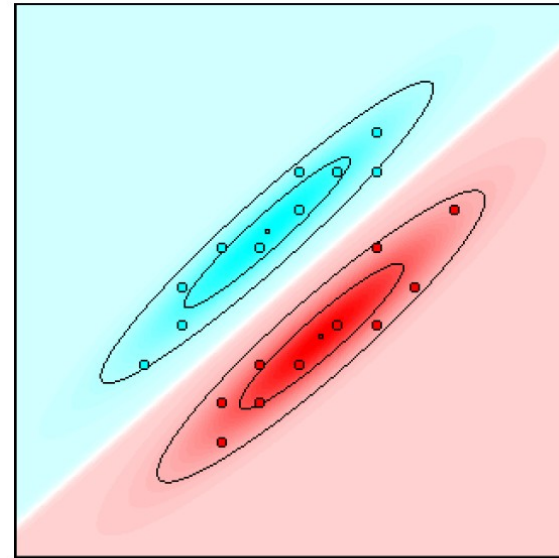
Naive vs. Full Bayes Classifiers

Naïve vs. Full Bayes Classifiers

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



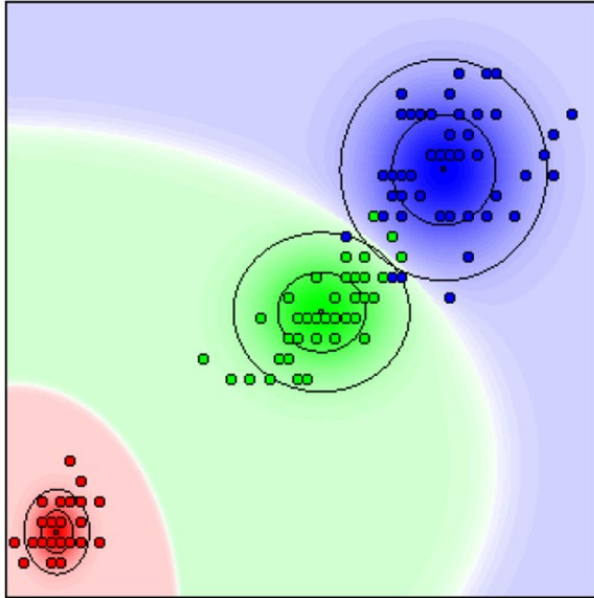
Naïve Bayes classifier



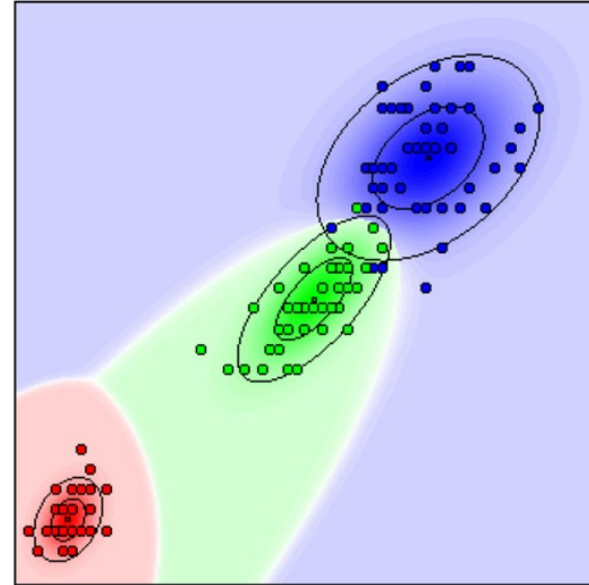
Full Bayes classifier

Naïve vs. Full Bayes Classifiers

– Iris data



Naïve Bayes classifier



Full Bayes classifier

Pros:

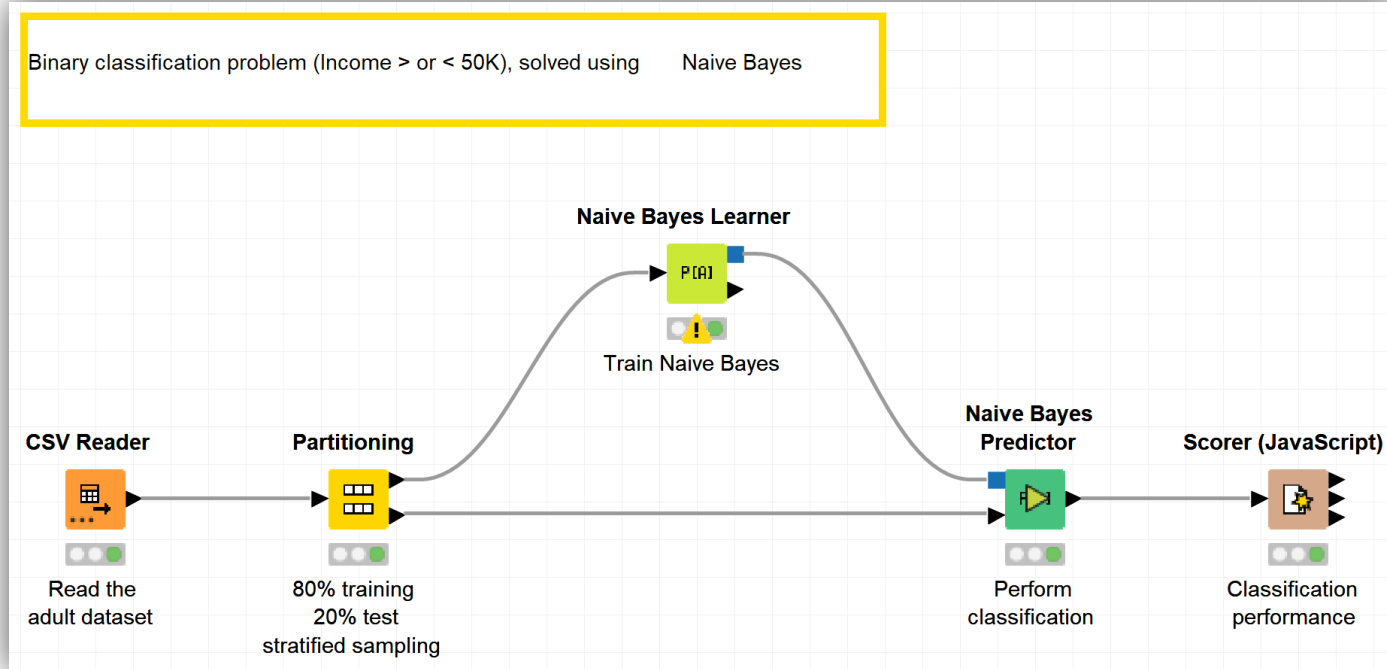
- Gold standard for comparison with other classifiers
- High classification accuracy in many applications
- Classifier can easily be adapted to new training objects
- Integration of domain knowledge

Cons:

- The conditional probabilities may 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%

Berthold, M.R., Borgelt, C., Höppner, F., Klawonn, F. and Silipo, R. (2020). *Guide to Intelligent Data Science: How to Intelligently Make Use of Real Data*. 2nd ed. [online] www.springer.com. Springer International Publishing.

Available at:

<https://www.springer.com/gp/book/9783030455736> [Accessed 17 Jun. 2021].

Questions?