Finding Explanations

- Instead of finding structure in a data set, we are now focusing on methods that find explanations for an unknown dependency within the data.
- Given: Dataset $\mathcal{D} = \{(\mathbf{x}_i, Y_i) | i = 1, ..., n\}$ with n tuples
  - $\mathbf{x}$: Object description
  - Y: Target attribute
    - nominal: classification problem
    - numerical: regression problem
- Data analysis
  - Supervised (because we know the desired outcome)
  - Descriptive (because we care about explanation)

Bayes Classifiers

- Given: Dataset $\mathcal{D} = \{(\mathbf{x}_i, Y_i) | i = 1, ..., n\}$ with n tuples
  - $\mathbf{x}$: Object description
  - Y: Nominal target attribute $\Rightarrow$ classification problem
- Bayes classifiers express their model in terms of simple probabilities.
- Provide 'gold standard' for evaluating other learning algorithms.
- Any other model should at least perform as well as the naive Bayes classifier.

Suggestion

Before trying to apply more complex models, a quick look at a Bayes classifier can be helpful to get a feeling for realistic accuracy expectations and simple dependencies in the data.

Bayes' theorem

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

Interpretation

The probability $P(h|E)$ that a hypothesis h is true given event E has occurred, can be derived from

- $P(h)$ the probability of the hypothesis h itself,
- $P(E)$ the probability of the event E and
- $P(E|h)$ the conditional probability of the event E given the hypothesis h.

Choosing Hypotheses

- We want the most probable hypothesis $h \in H$ for a given event E
- Maximum a posteriori hypothesis:

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h|E) \\ &= \arg\max_{h \in H} \frac{P(E|h)P(h)}{P(E)} \\ &= \arg\max_{h \in H} P(E|h)P(h) \end{aligned}$$

Maximum likelihood

If we assume that every hypothesis $h \in H$ is equally probable a priori ($P(h_i) = P(h_j)$ for all $h_i, h_j \in H$) we can further simplify the equation and get the maximum likelihood hypothesis:

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

Bayes classifiers

The probability $P(h)$ can be estimated easily based on a given data set D:

$$P(h) = \frac{\text{no. of data from class } h}{\text{no. of data}}$$

In principle, the probability $P(E|h)$ could be determined analogously based on the values of the attributes $A_1, \ldots, A_m$, i.e. the attribute vector $E = (a_1, \ldots, a_m)$.

$$P(E|h) = \frac{\text{no. of data from class } h \text{ with values } (a_1, \ldots, a_m)}{\text{no. of data from class } h}$$

Bayes classifiers

Problem

For $n = 10$ nominal attributes $A_1, \ldots, A_{10}$, each having three possible values, we would need $3^{10} = 59049$ data objects to have at least one example per combination.

Therefore, the computation is carried out under the (naïve, unrealistic) assumption that the attributes $A_1, \ldots, A_m$ are independent given the class, i.e.

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

$P(a_i|h)$ can be computed easily:

$$P(a_i|h) \;=\; \frac{\text{no. of data from class } h \text{ with } A_i = a_i}{\text{no. of data from class } h}$$

Naïve Bayes classifier

Given: A data set with only nominal attributes.

Based on the values $a_1, \ldots, a_m$ of the attributes $A_1, \ldots, A_m$ a prediction for the value of the attribute H should be derived:

- For each class $h \in H$ compute the likelihood $L(h|E)$ under the assumption that the $A_1, \ldots, A_m$ 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

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

This Bayes classifier is called naïve because of the (conditional) independence assumption for the attributes $A_1, \ldots, A_m$.

Although this assumption is unrealistic in most cases, the classifier often yields good results, 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 ($\underline{m}ale$ 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

$P(\text{Height} = t|\text{Sex} = m)$

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

$P(\text{Height} = t | \text{Sex} = m)$

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

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

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

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

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

$\Rightarrow$ 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

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

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

Example

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

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

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

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

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

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

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

$\Rightarrow$ the likelihood of person $\mathbf{x}$ 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.

More examples

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

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

More examples

Input	$L(m\vert\ldots)$	$L(f\vert\ldots)$	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)	$\frac{2}{4}\cdot\frac{2}{4}\cdot\frac{0}{4}\cdot\frac{4}{10}=0$	$\frac{1}{6}\cdot\frac{0}{6}\cdot\frac{4}{6}\cdot\frac{6}{10}=0$	?

The object (t,h,y) can not 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

Laplace correction

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)	$\frac{2}{4} \cdot \frac{2}{4} \cdot \frac{0}{4} \cdot \frac{4}{10} = 0$	$\frac{1}{6} \cdot \frac{0}{6} \cdot \frac{4}{6} \cdot \frac{6}{10} = 0$	?

A solution is the usage of the Laplace correction γ:

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

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

- 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

The counting of the frequencies should be carried out once when the naïve Bayes classifier is constructed.

The probability distribution for the single attributes should be stored in a table.

When the naïve Bayes classifier is applied to new data, only the corresponding values in the table need to be multiplied.

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

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

- Estimation of the mean value

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

- Estimation of the variance

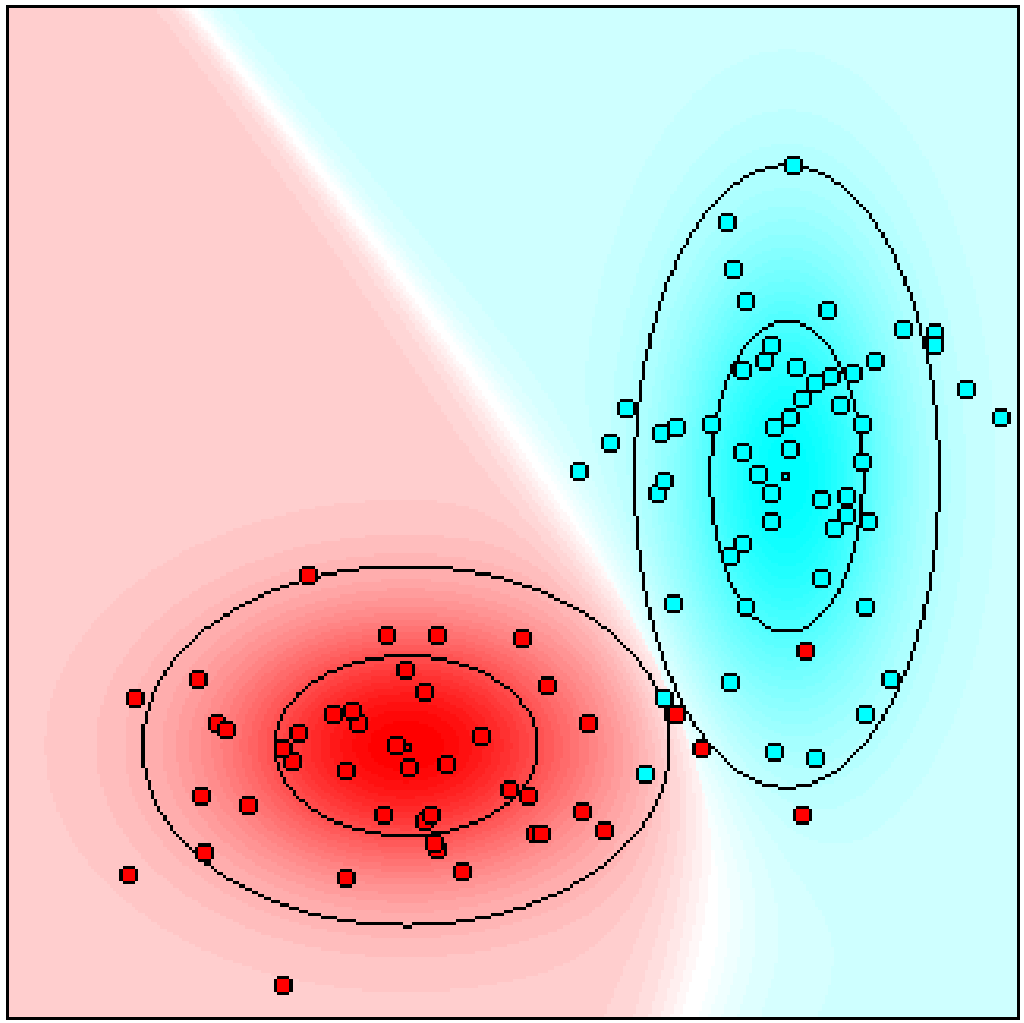
$$\hat{\sigma}_{X|y}^2 = \frac{1}{n'_y} \sum_{i=1}^{n} \tau(y_i = y) \cdot \left(\mathbf{x}_i[X] - \hat{\mu}_{X|y}\right)^2$$

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

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

Example

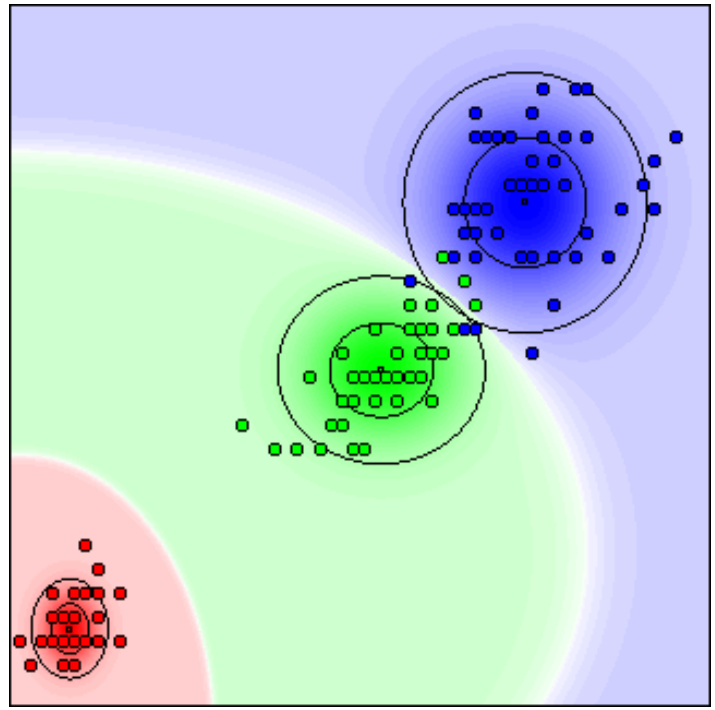
- 100 data points, 2 classes
- Small squares: mean values
- Inner ellipses:
  one standard deviation
- Outer ellipses:
  two standard deviations
- Classes overlap:
  classification is not perfect



Naïve Bayes classifier

Naïve Bayes classifier: Iris data

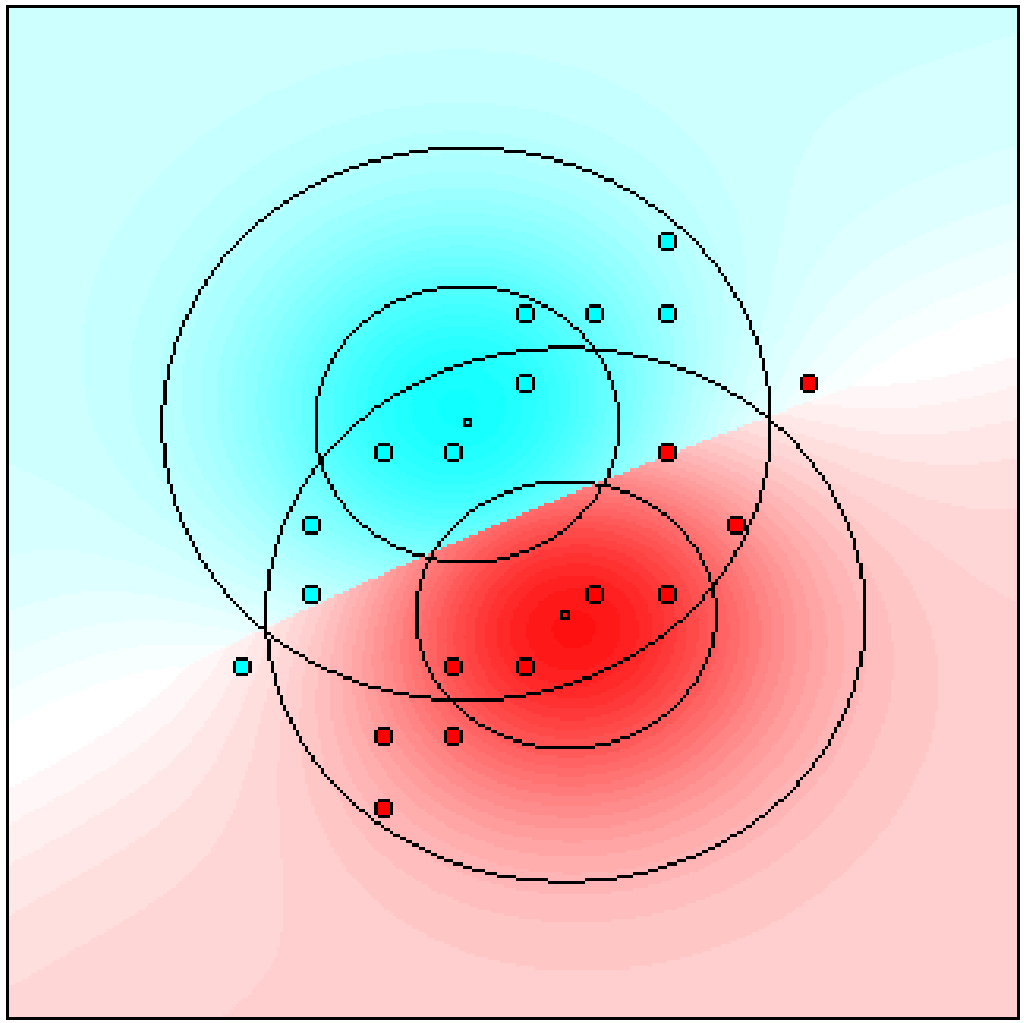
- 150 data points, 3 classes
  Iris setosa (red)
  Iris versicolor (green)
  Iris virginica (blue)
- Shown: 2 out of 4 attributes
  sepal length
  sepal width
  petal length (horizontal)
  petal width (vertical)
- 6 misclassifications on the training data
  (with all 4 attributes)



Naïve Bayes classifier

Example

- 20 data points, 2 classes
- Small squares: mean values
- Inner ellipses:
  one standard deviation
- Outer ellipses:
  two standard deviations
- Attributes are not
  conditionally independent
  given the class



Naïve Bayes classifier

Full Bayes classifiers

Restricted to metric/numeric attributes (only the class is nominal/symbolic).

Simplifying Assumption:
Each class can be described by a multivariate normal distribution

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

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

Intuitively

Each class has a bell-shaped probability density.

Estimation of Probabilities:

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

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

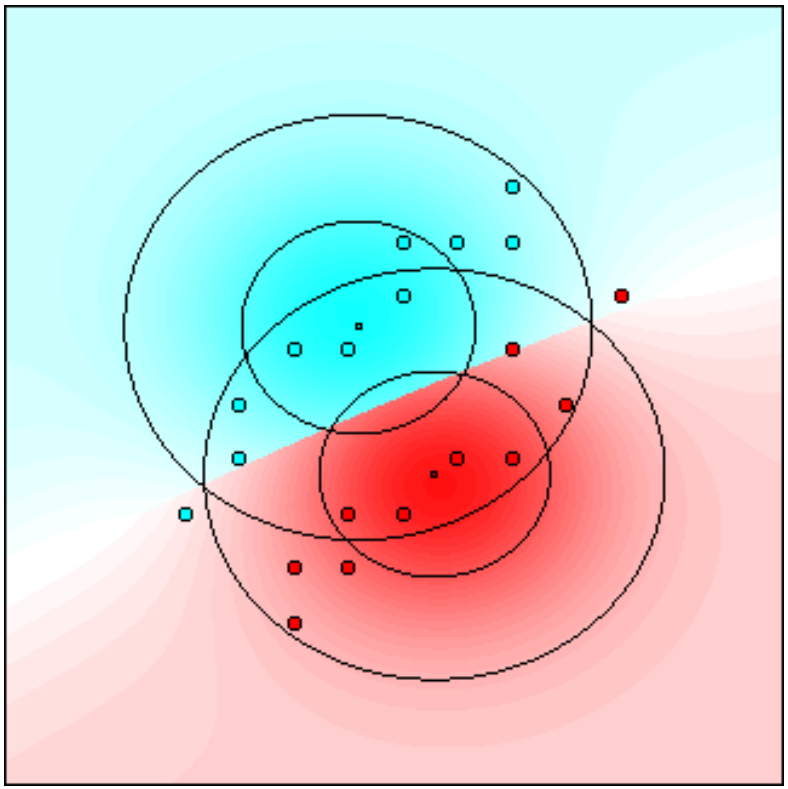
$\mathbf{x}_i[\mathbf{X}_M]$: attribute vector $\mathbf{x}$ at position i that contains the values of all metric attributes $\mathbf{X}_M$

- Estimation of the (class-conditional) covariance matrix

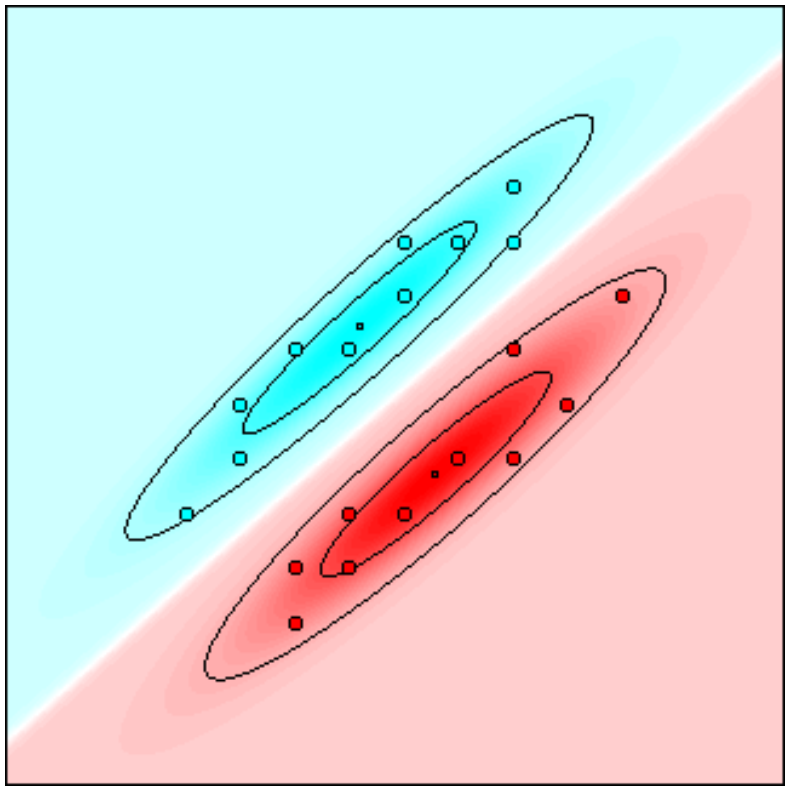
$$\widehat{\mathbf{\Sigma}}_{\mathbf{X}_M|y} = \frac{1}{n'_y} \sum_{i=1}^{n} \tau(y_i = y) \times \left(\mathbf{x}_i[\mathbf{X}_M] - \hat{\mu}_{\mathbf{X}_M|y}\right) \left(\mathbf{x}_i[\mathbf{X}_M] - \hat{\mu}_{\mathbf{X}_M|y}\right)^\top$$

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

Naïve vs. full Bayes classifiers



Naïve Bayes classifier



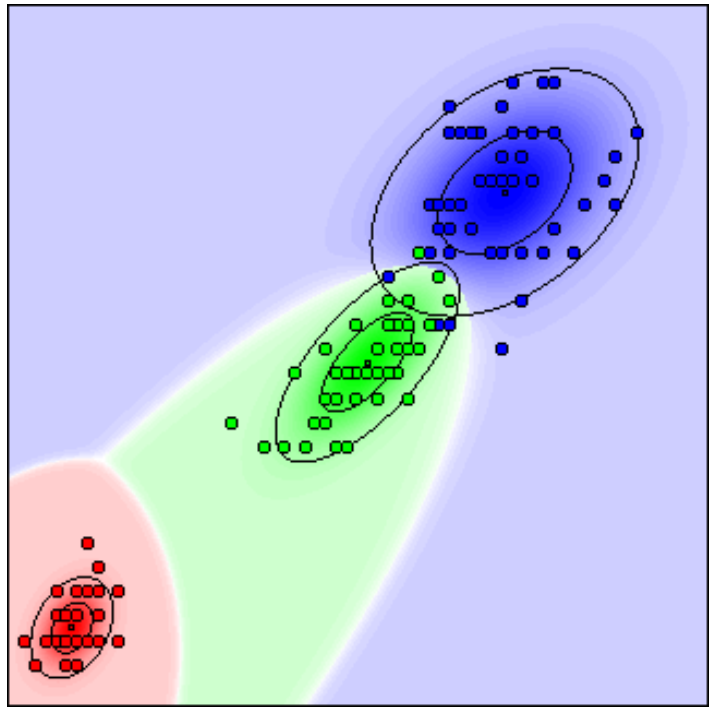
Full Bayes classifier

Notice

Naïve Bayes classifiers for numerical data are equivalent to full Bayes classifiers with diagonal covariance matrices.

Full Bayes classifier: Iris data

- 150 data points, 3 classes
  Iris setosa (red)
  Iris versicolor (green)
  Iris virginica (blue)
- Shown: 2 out of 4 attributes
  sepal length
  sepal width
  petal length (horizontal)
  petal width (vertical)
- 2 misclassifications on the training data
  (with all 4 attributes)



Full Bayes classifier

Summary

- **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