Bayes Classifiers

- **Probabilistic Classification and Bayes' Rule**
- **Naive Bayes Classifiers**
  - Derivation of the classification formula
  - Probability estimation and Laplace correction
  - Simple examples of naive Bayes classifiers
  - A naive Bayes classifier for the Iris data
- **Full Bayes Classifiers**
  - Derivation of the classification formula
  - Comparison to naive Bayes classifiers
  - A simple example of a full Bayes classifier
  - A full Bayes classifier for the Iris data
- **Summary**

Probabilistic Classification

- A classifier is an algorithm that assigns a class from a predefined set to a case or object, based on the values of descriptive attributes.
- An optimal classifier maximizes the probability of a correct class assignment.
  - Let C be a class attribute with $\mathrm{dom}(C) = \{c_1, \ldots, c_{n_C}\}$, which occur with probabilities p_i, $1 \leq i \leq n_C$.
  - Let q_i be the probability with which a classifier assigns class c_i. ($q_i \in \{0, 1\}$ for a deterministic classifier)
  - The probability of a correct assignment is

$$P(\text{correct assignment}) = \sum_{i=1}^{n_C} p_i q_i.$$

  - Therefore the best choice for the q_i is

$$q_i = \begin{cases} 1, & \text{if } p_i = \max_{k=1}^{n_C} p_k, \\ 0, & \text{otherwise.} \end{cases}$$

Probabilistic Classification

- Consequence: An optimal classifier should assign the **most probable class**.
- This argument does not change if we take descriptive attributes into account.
  - Let $U = \{A_1, \ldots, A_m\}$ be a set of descriptive attributes with domains $\mathrm{dom}(A_k)$, $1 \leq k \leq m$.
  - Let $A_1 = a_1, \ldots, A_m = a_m$ be an instantiation of the descriptive attributes.
  - An optimal classifier should assign the class c_i for which

$$P(C = c_i \mid A_1 = a_1, \ldots, A_m = a_m) =$$
$$\max_{j=1}^{n_C} \; P(C = c_j \mid A_1 = a_1, \ldots, A_m = a_m)$$

- **Problem:** We cannot store a class (or the class probabilities) for every possible instantiation $A_1 = a_1, \ldots, A_m = a_m$ of the descriptive attributes. (The table size grows exponentially with the number of attributes.)
- Therefore: **Simplifying assumptions are necessary.**

Bayes' Rule and Bayes' Classifiers

- Bayes' rule is a formula that can be used to "invert" conditional probabilities: Let X and Y be events, $P(X) > 0$. Then

$$P(Y \mid X) = \frac{P(X \mid Y) \cdot P(Y)}{P(X)}.$$

- Bayes' rule follows directly from the definition of conditional probability:

$$P(Y \mid X) = \frac{P(X \cap Y)}{P(X)} \qquad \text{and} \qquad P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}.$$

- Bayes' classifiers: Compute the class probabilities as

$$P(C = c_i \mid A_1 = a_1, \ldots, A_m = a_m) =$$
$$\frac{P(A_1 = a_1, \ldots, A_m = a_m \mid C = c_i) \cdot P(C = c_i)}{P(A_1 = a_1, \ldots, A_m = a_m)}.$$

- Looks unreasonable at first sight: Even more probabilities to store.

Naive Bayes Classifiers

Naive Assumption:
The descriptive attributes are conditionally independent given the class.

Bayes' Rule:

$$P(C = c_i \mid \omega) = \frac{P(A_1 = a_1, \ldots, A_m = a_m \mid C = c_i) \cdot P(C = c_i)}{P(A_1 = a_1, \ldots, A_m = a_m)} \quad \leftarrow p_0$$

abbrev. for the normalizing constant

Chain Rule of Probability:

$$P(C = c_i \mid \omega) = \frac{P(C = c_i)}{p_0} \cdot \prod_{k=1}^{m} P(A_k = a_k \mid A_1 = a_1, \ldots, A_{k-1} = a_{k-1}, C = c_i)$$

Conditional Independence Assumption:

$$P(C = c_i \mid \omega) = \frac{P(C = c_i)}{p_0} \cdot \prod_{k=1}^{m} P(A_k = a_k \mid C = c_i)$$

Reminder: Chain Rule of Probability

- Based on the **product rule** of probability:

$$P(A \wedge B) = P(A \mid B) \cdot P(B)$$

(Multiply definition of conditional probability with $P(B)$.)

- **Multiple application** of the product rule yields:

$$\begin{aligned} P(A_1, \ldots, A_m) &= P(A_m \mid A_1, \ldots, A_{m-1}) \cdot P(A_1, \ldots, A_{m-1}) \\ &= P(A_m \mid A_1, \ldots, A_{m-1}) \\ &\cdot\; P(A_{m-1} \mid A_1, \ldots, A_{m-2}) \cdot P(A_1, \ldots, A_{m-2}) \\ &= \vdots \\ &= \prod_{k=1}^{m} P(A_k \mid A_1, \ldots, A_{k-1}) \end{aligned}$$

- The scheme works also if there is already a condition in the original expression:

$$P(A_1, \ldots, A_m \mid C) = \prod_{i=1}^{m} P(A_k \mid A_1, \ldots, A_{k-1}, C)$$

Conditional Independence

- Reminder: **stochastic independence** (unconditional)

$$P(A \wedge B) = P(A) \cdot P(B)$$

(Joint probability is the product of the individual probabilities.)

- Comparison to the **product rule**

$$P(A \wedge B) = P(A \mid B) \cdot P(B)$$

shows that this is equivalent to

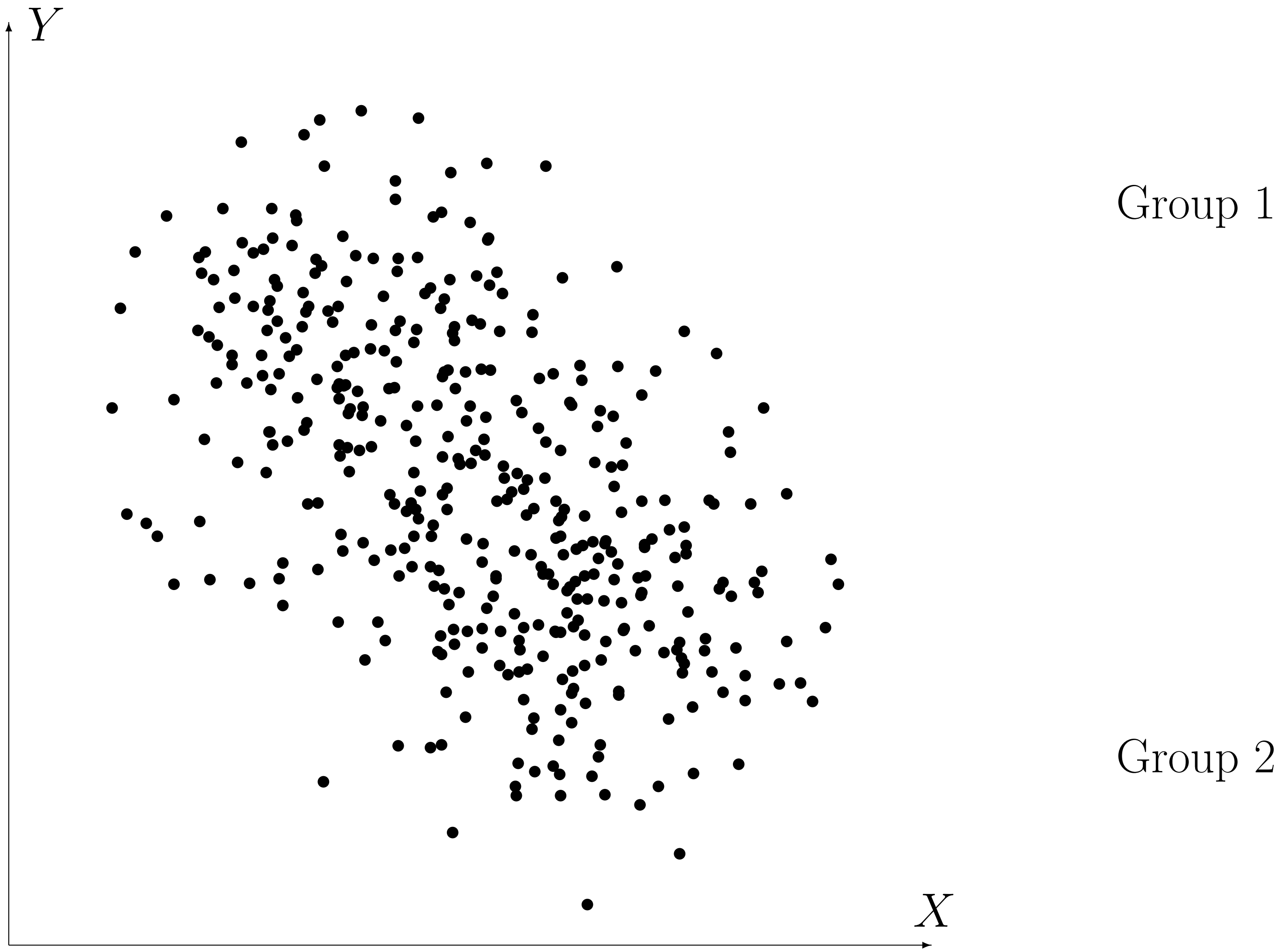
$$P(A \mid B) = P(A)$$

- The same formulae hold conditionally, i.e.

$$\begin{aligned} P(A \wedge B \mid C) &= P(A \mid C) \cdot P(B \mid C) \qquad \text{and} \\ P(A \mid B, C) &= P(A \mid C). \end{aligned}$$

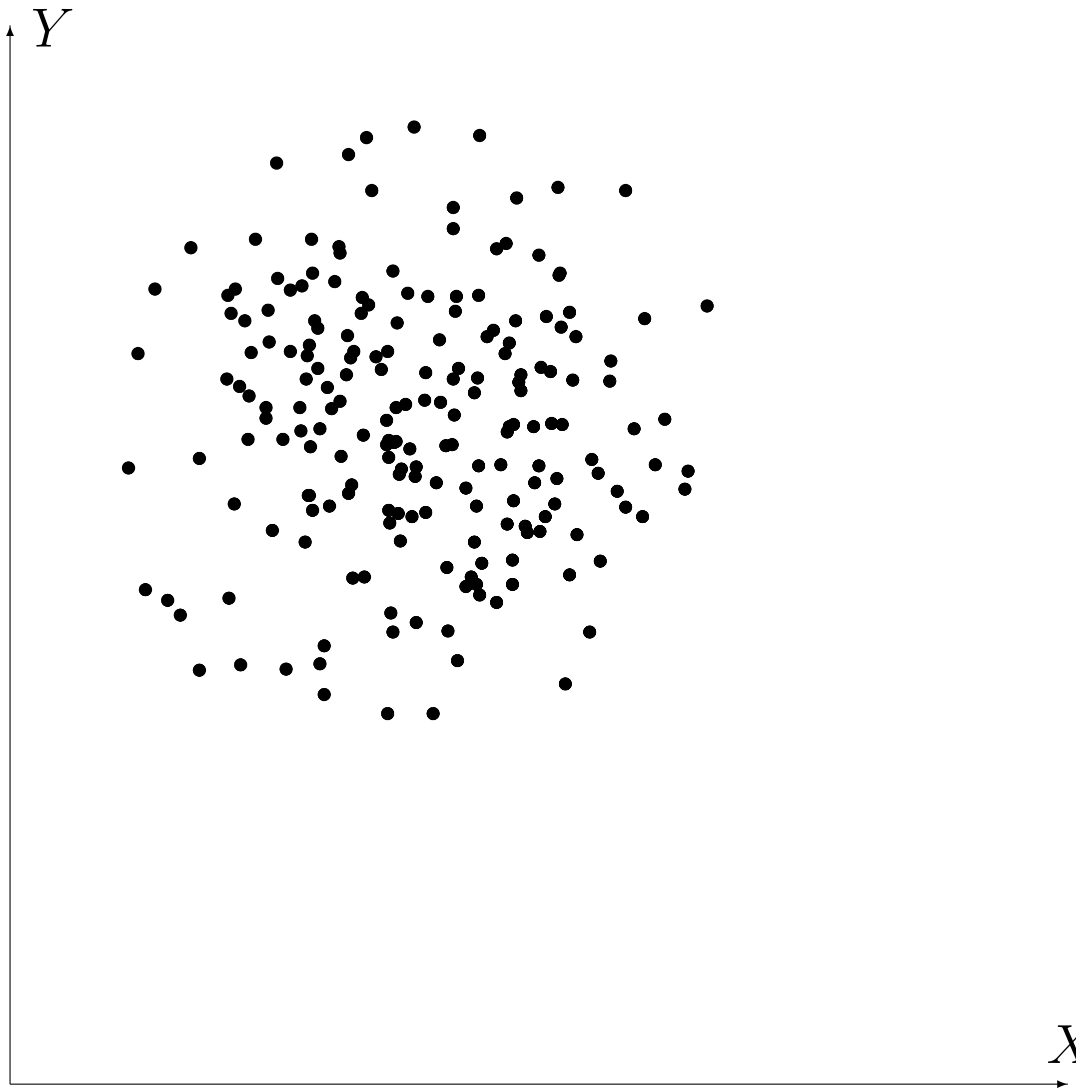
- **Conditional independence allows us to cancel some conditions**.

Conditional Independence: An Example



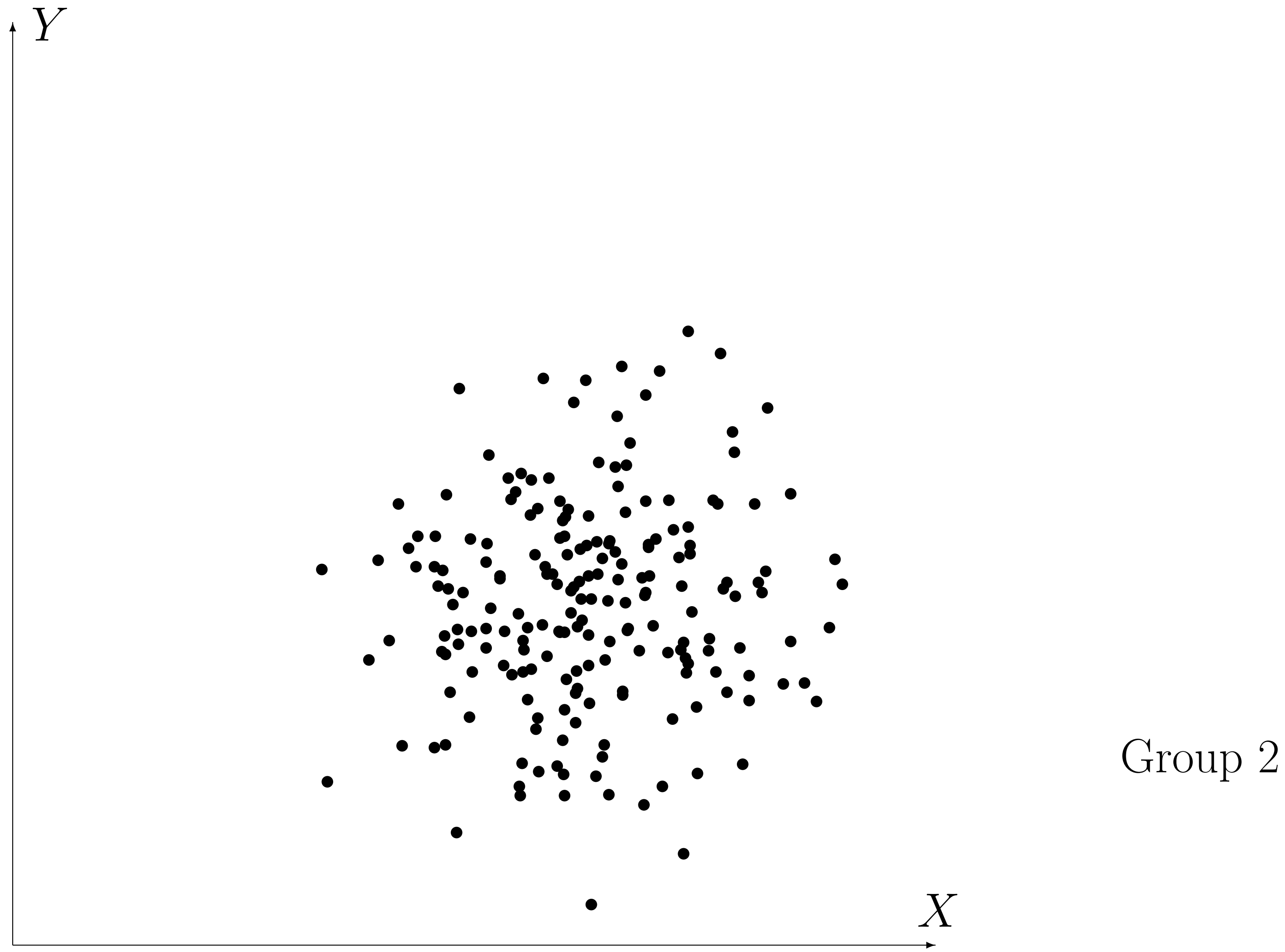
(Weak) Dependence in the entire dataset: X and Y dependent.

Conditional Independence: An Example



No Dependence in Group 1: X and Y conditionally independent given Group 1.

Conditional Independence: An Example



No Dependence in Group 2: X and Y conditionally independent given Group 2.

Marginal & Conditional Independence

- The next table shows four 3-dimensional probability distributions (one per row).
- The (in)dependencies are always w. r. t. A and B.
- The condition variable is C.

		a_1				a_2			
		b_1		b_2		b_1		b_2	
marginal	conditional	c_1	c_2	c_1	c_2	c_1	c_2	c_1	c_2
indep.	indep.	0.03	0.01	0.27	0.09	0.006	0.054	0.054	0.486
dep.	dep.	0.01	0.03	0.126	0.234	0.1275	0.3825	0.0315	0.0585
dep.	indep.	0.12	0.085	0.18	0.015	0.024	0.459	0.036	0.081
indep.	dep.	0.008	0.032	0.144	0.216	0.018	0.042	0.054	0.486

- All combinations are possible.

- Consequence: Manageable amount of data to store.
  Store distributions $P(C = c_i)$ and $\forall 1 \leq k \leq m : P(A_k = a_k \mid C = c_i)$.

- It is not necessary to compute p_0 explicitely, because it can be computed implicitly by normalizing the computed values to sum 1.

Estimation of Probabilities:

- **Nominal/Symbolic Attributes**

$$\hat{P}(A_k = a_k \mid C = c_i) = \frac{\#(A_k = a_k, C = c_i) + \gamma}{\#(C = c_i) + n_{A_k}\gamma}$$

  γ is called **Laplace correction**: Assume for every class c_i some number of hypothetical samples for every value of A_k to prevent the estimate to be 0 if $\#(A_k = a_k, C = c_i) = 0$.

  $\gamma = 0$: Maximum likelihood estimation.

  Common choices: $\gamma = 1$ or $\gamma = \frac{1}{2}$.

Estimation of Probabilities:

- **Metric/Numeric Attributes:** Assume a normal distribution.

$$P(A_k = a_k \mid C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_k(c_i)} \exp\left(-\frac{(a_k - \mu_k(c_i))^2}{2\sigma_k^2(c_i)}\right)$$

- Estimate of mean value

$$\hat{\mu}_k(c_i) = \frac{1}{\#(C = c_i)} \sum_{j=1}^{\#(C=c_i)} a_k(j)$$

- Estimate of variance

$$\hat{\sigma}_k^2(c_i) = \frac{1}{\xi} \sum_{j=1}^{\#(C=c_i)} \left(a_k(j) - \hat{\mu}_k(c_i)\right)^2$$

$\xi = \#(C = c_i)$: Maximum likelihood estimation
$\xi = \#(C = c_i) - 1$: Unbiased estimation

Naive Bayes Classifiers: Simple Example 1

No	Sex	Age	Blood pr.	Drug
1	male	20	normal	A
2	female	73	normal	B
3	female	37	high	A
4	male	33	low	B
5	female	48	high	A
6	male	29	normal	A
7	female	52	normal	B
8	male	42	low	B
9	male	61	normal	B
10	female	30	normal	A
11	female	26	low	B
12	male	54	high	A

$P(\text{Drug})$	A	B
	0.5	0.5

$P(\text{Sex} \mid \text{Drug})$	A	B
male	0.5	0.5
female	0.5	0.5

$P(\text{Age} \mid \text{Drug})$	A	B
μ	36.3	47.8
σ^2	161.9	311.0

$P(\text{Blood Pr.} \mid \text{Drug})$	A	B
low	0	0.5
normal	0.5	0.5
high	0.5	0

A simple database and estimated (conditional) probability distributions.

Naive Bayes Classifiers: Simple Example 1

$$
\begin{aligned}
&P(\text{Drug A} \mid \text{male, 61, normal}) \\
&\quad = \; c_1 \cdot P(\text{Drug A}) \cdot P(\text{male} \mid \text{Drug A}) \cdot P(61 \mid \text{Drug A}) \cdot P(\text{normal} \mid \text{Drug A}) \\
&\quad \approx \; c_1 \cdot 0.5 \cdot 0.5 \cdot 0.004787 \cdot 0.5 \; = \; c_1 \cdot 5.984 \cdot 10^{-4} \; = \; 0.219
\end{aligned}
$$

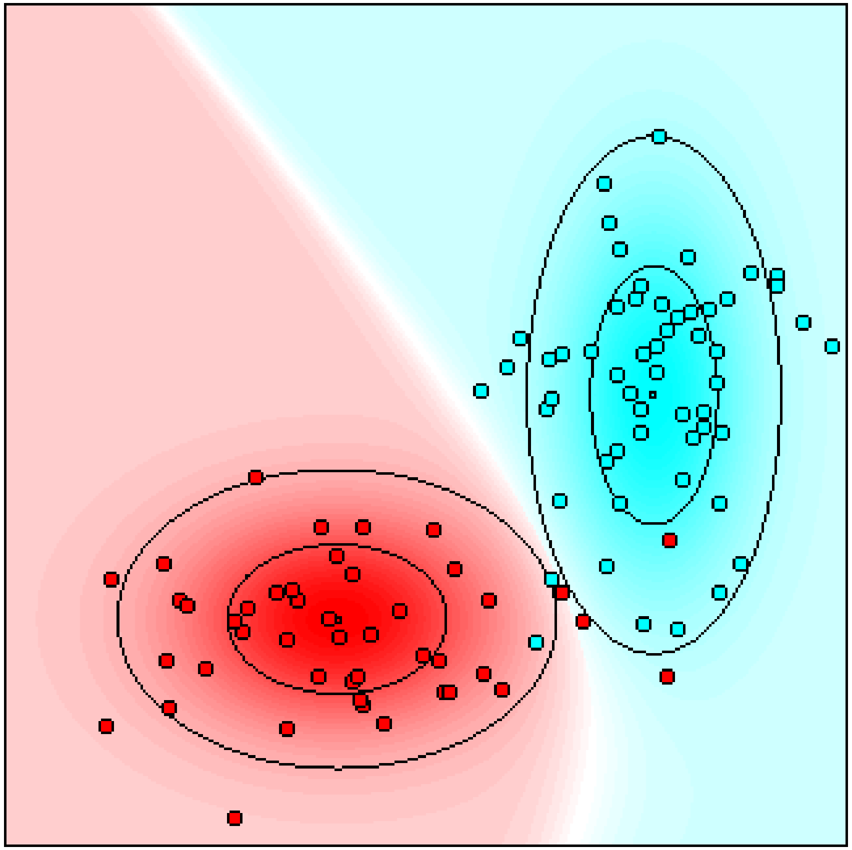
$$
\begin{aligned}
&P(\text{Drug B} \mid \text{male, 61, normal}) \\
&\quad = \; c_1 \cdot P(\text{Drug B}) \cdot P(\text{male} \mid \text{Drug B}) \cdot P(61 \mid \text{Drug B}) \cdot P(\text{normal} \mid \text{Drug B}) \\
&\quad \approx \; c_1 \cdot 0.5 \cdot 0.5 \cdot 0.017120 \cdot 0.5 \; = \; c_1 \cdot 2.140 \cdot 10^{-3} \; = \; 0.781
\end{aligned}
$$

$$
\begin{aligned}
&P(\text{Drug A} \mid \text{female, 30, normal}) \\
&\quad = \; c_2 \cdot P(\text{Drug A}) \cdot P(\text{female} \mid \text{Drug A}) \cdot P(30 \mid \text{Drug A}) \cdot P(\text{normal} \mid \text{Drug A}) \\
&\quad \approx \; c_2 \cdot 0.5 \cdot 0.5 \cdot 0.027703 \cdot 0.5 \; = \; c_2 \cdot 3.471 \cdot 10^{-3} \; = \; 0.671
\end{aligned}
$$

$$
\begin{aligned}
&P(\text{Drug B} \mid \text{female, 30, normal}) \\
&\quad = \; c_2 \cdot P(\text{Drug B}) \cdot P(\text{female} \mid \text{Drug B}) \cdot P(30 \mid \text{Drug B}) \cdot P(\text{normal} \mid \text{Drug B}) \\
&\quad \approx \; c_2 \cdot 0.5 \cdot 0.5 \cdot 0.013567 \cdot 0.5 \; = \; c_2 \cdot 1.696 \cdot 10^{-3} \; = \; 0.329
\end{aligned}
$$

Naive Bayes Classifiers: Simple Example 2

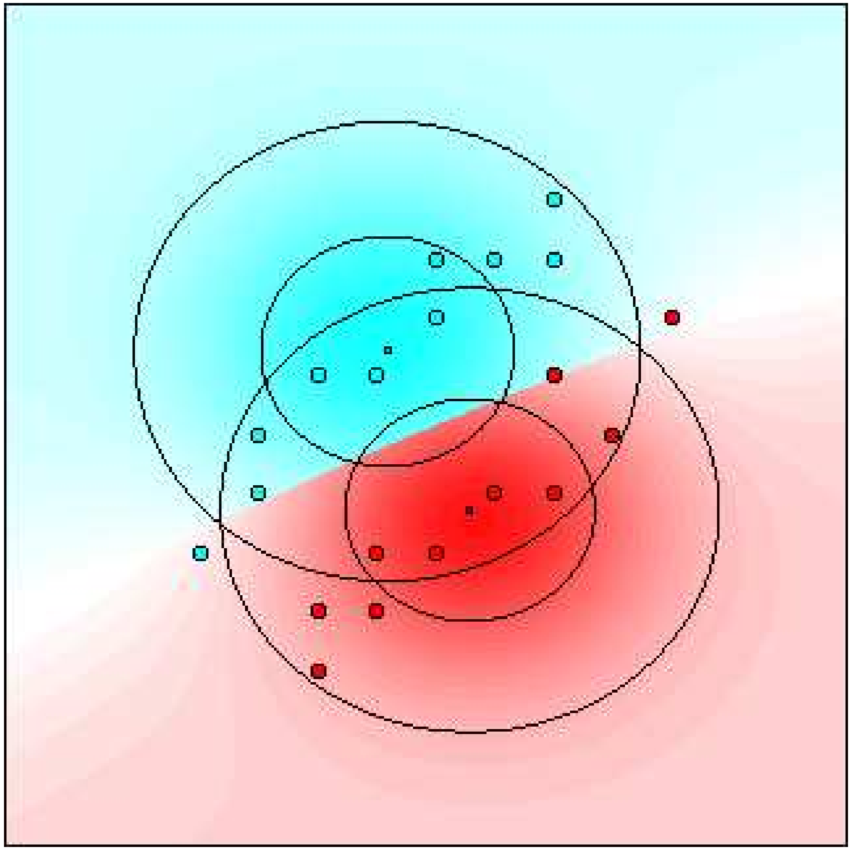
- 100 data points, 2 classes
- Small squares: mean values
- Inner ellipses: one standard deviation
- Outer ellipses: two standard deviations
- Classes overlap: classification is not perfect



Naive Bayes Classifier

Naive Bayes Classifiers: Simple Example 3

- 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



Naive Bayes Classifier

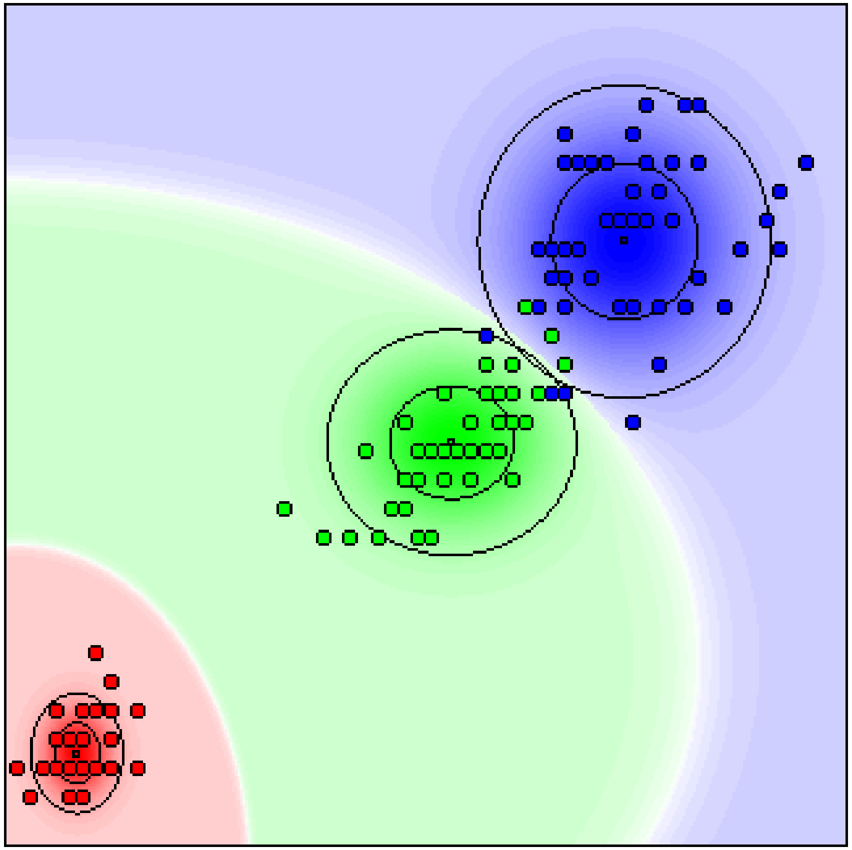
- 150 data points, 3 classes

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)



Naive 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.

$$
\begin{aligned}
& f(A_1 = a_1, \ldots, A_m = a_m \mid C = c_i) \\
&= \frac{1}{\sqrt{(2\pi)^m |\mathbf{\Sigma}_i|}} \exp\left(-\frac{1}{2}(\vec{a} - \vec{\mu}_i)^\top \mathbf{\Sigma}_i^{-1} (\vec{a} - \vec{\mu}_i)\right)
\end{aligned}
$$

  $\vec{\mu}_i$: mean value vector for class c_i

  $\mathbf{\Sigma}_i$: covariance matrix for class c_i

- Intuitively: Each class has a bell-shaped probability density.
- Naive Bayes classifiers: Covariance matrices are diagonal matrices.
  (Details about this relation are given below.)

Estimation of Probabilities:

- Estimate of mean value vector

$$\hat{\vec{\mu}}_i = \frac{1}{\#(C = c_i)} \sum_{j=1}^{\#(C=c_i)} \vec{a}(j)$$

- Estimate of covariance matrix

$$\widehat{\mathbf{\Sigma}}_i = \frac{1}{\xi} \sum_{j=1}^{\#(C=c_i)} \left(\vec{a}(j) - \hat{\vec{\mu}}_i\right) \left(\vec{a}(j) - \hat{\vec{\mu}}_i\right)^\top$$

$\xi = \#(C = c_i)$: Maximum likelihood estimation
$\xi = \#(C = c_i) - 1$: Unbiased estimation

$\vec{x}^\top$ denotes the transpose of the vector $\vec{x}$.

$\vec{x}\vec{x}^\top$ is the so-called **outer product** or **matrix product** of $\vec{x}$ with itself.

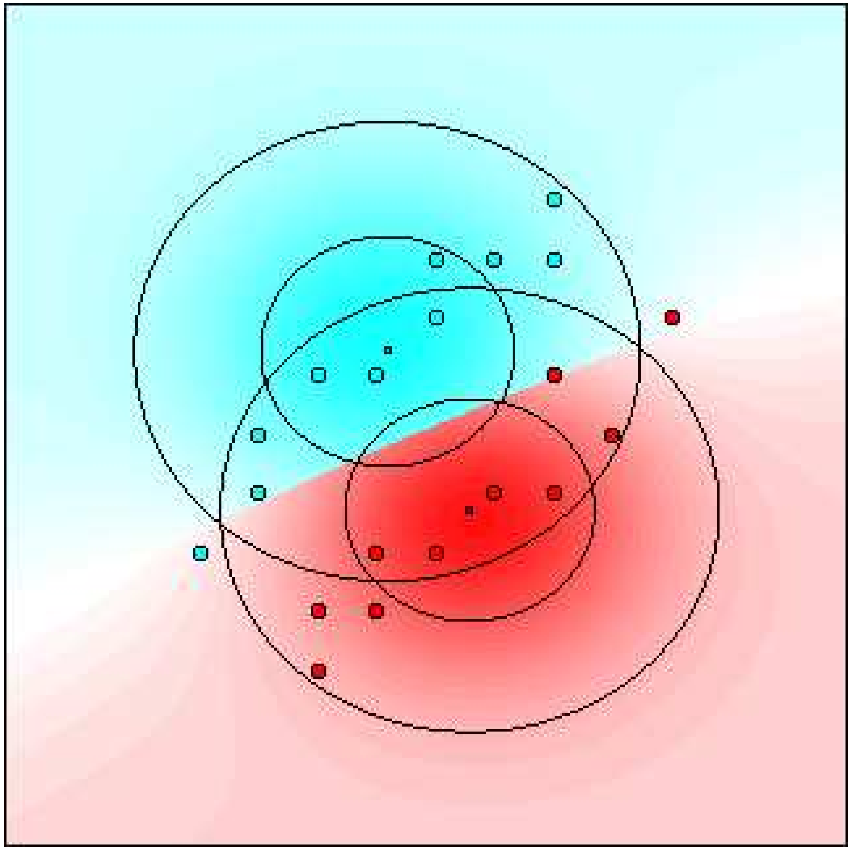
Comparison of Naive and Full Bayes Classifiers

Naive Bayes classifiers for metric/numeric data are equivalent to full Bayes classifiers with diagonal covariance matrices:

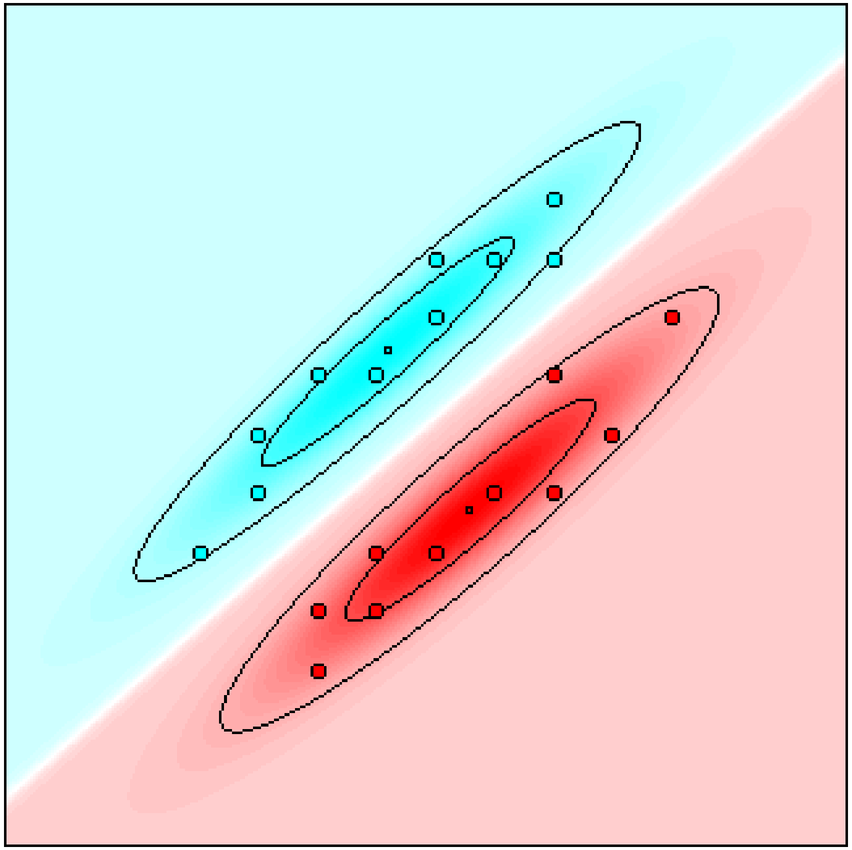
$$
\begin{aligned}
& f(A_1 = a_1, \ldots, A_m = a_m \mid C = c_i) \\
&= \frac{1}{\sqrt{(2\pi)^m |\mathbf{\Sigma}_i|}} \cdot \exp\left(-\frac{1}{2}(\vec{a} - \vec{\mu}_i)^\top \mathbf{\Sigma}_i^{-1} (\vec{a} - \vec{\mu}_i)\right) \\
&= \frac{1}{\sqrt{(2\pi)^m \prod_{k=1}^m \sigma_{i,k}^2}} \cdot \exp\left(-\frac{1}{2}(\vec{a} - \vec{\mu}_i)^\top \operatorname{diag}\left(\sigma_{i,1}^{-2}, \ldots, \sigma_{i,m}^{-2}\right) (\vec{a} - \vec{\mu}_i)\right) \\
&= \frac{1}{\prod_{k=1}^m \sqrt{2\pi\sigma_{i,k}^2}} \cdot \exp\left(-\frac{1}{2} \sum_{k=1}^{m} \frac{(a_k - \mu_{i,k})^2}{\sigma_{i,k}^2}\right) \\
&= \prod_{k=1}^{m} \frac{1}{\sqrt{2\pi\sigma_{i,k}^2}} \cdot \exp\left(-\frac{(a_k - \mu_{i,k})^2}{2\sigma_{i,k}^2}\right) \quad \widehat{=} \quad \prod_{k=1}^{m} f(A_k = a_k \mid C = c_i),
\end{aligned}
$$

where $f(A_k = a_k \mid C = c_i)$ are the density functions used by a naive Bayes classifier.

Comparison of Naive and Full Bayes Classifiers



Naive Bayes Classifier



Full Bayes Classifier

Full Bayes Classifiers: Iris Data

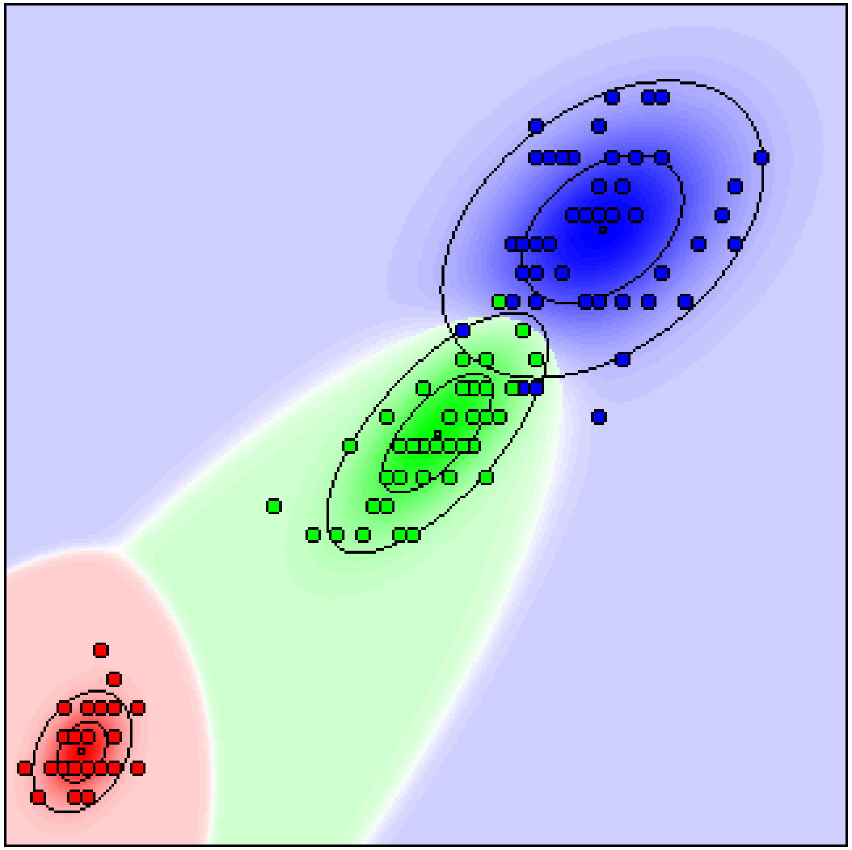
- 150 data points, 3 classes

  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 Bayes Classifiers

- **Probabilistic Classification**: Assign the most probable class.
- **Bayes' Rule**: "Invert" the conditional class probabilities.
- **Naive Bayes Classifiers**
  - Simplifying Assumption:
    Attributes are conditionally independent given the class.
  - Can handle nominal/symbolic as well as metric/numeric attributes.
- **Full Bayes Classifiers**
  - Simplifying Assumption:
    Each class can be described by a multivariate normal distribution.
  - Can handle only metric/numeric attributes.