Bayesian Network

Alymzhan Toleu
alymzhan.toleu@gmail.com

Bayes decision rule

- If we know the conditional probability P(X | y) we can determine the appropriate class by using Bayes rule:

Posterior

class-conditional probability
Or Likelihood

Prior

$$P(y|X) = \frac{P(X,y)}{P(X)} = \frac{P(X|y) * P(y)}{P(X)}$$

Evidence

Recall…
X – input attributes (features)
y – the class label

But how do we determine p(X|y)?

Maximum Likelihood Estimation

- Likelihood function

$$\boldsymbol{L}(\theta) \propto P(D \mid \theta) \quad \Rightarrow \quad \hat{\theta}^{MLE} = argmax\, \boldsymbol{L}(\theta)$$

$$\Downarrow \quad \boldsymbol{log}(\boldsymbol{L}(\theta)) \propto \textstyle\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{log}\left(P\left(x^{(i)} \middle| \theta\right)\right)$$

$$\hat{\theta}^{MLE} = \operatorname{argmax} \textstyle\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{log}\left(P\left(x^{(i)} \middle| \theta\right)\right)$$

Maximize Likelihood Estimation

Naïve Bayes Classifier

- **Assumption**: attribute conditional independence
- Based on this assumption, we get

$$P(y|X) = \frac{P(X,y)}{P(X)} = \frac{P(X|y) * P(y)}{P(X)} = \frac{P(y)}{P(X)} \prod_{i=1}^{d} P(x_i \mid y)$$

where d *is the number of attributes* and x_i *refers to i-th attribute's value of x.*

- With MLE, and ignoring *P(X)* (it is sample for all classes), NB calculates

$$\text{argmax}_y\, P(y) \prod_{i=1}^{d} P(x_i \mid y)$$

Naïve Bayes Classifier (cont.)

- NB classifier's training process is to calculate the values for the prior $P(y)$ and the conditional probability $P(x_i \mid y)$.
- For example, D_y is a set of samples that belongs to the class y and $D_y \in D$, then
  - the prior probability:
    $$P(y) = \frac{|D_y|}{|D|}$$
  - for **discrete** attributes, $D_{y,\,x_i}$ is the set of samples in D_y their attribute value equal to x_i:
    $$P(x_i \mid y) = \frac{|D_{y,\,x_i}|}{|D|}$$
  - for **continuous** attributes, need to consider a distribution, e.g. the normal distribution:
    $$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Semi-naïve Bayes classifiers

Semi-naïve Bayes classifiers

- Semi-naive Bayesian learning refers to a field of Supervised Classification that seeks to enhance the classification and conditional probability estimation accuracy of naive Bayes by relaxing its **attribute independence assumption**.

Semi-naïve Bayes classifiers

- One-Dependent Estimator (ODE)
  - assume that each attribute depends on at most one other attribute.

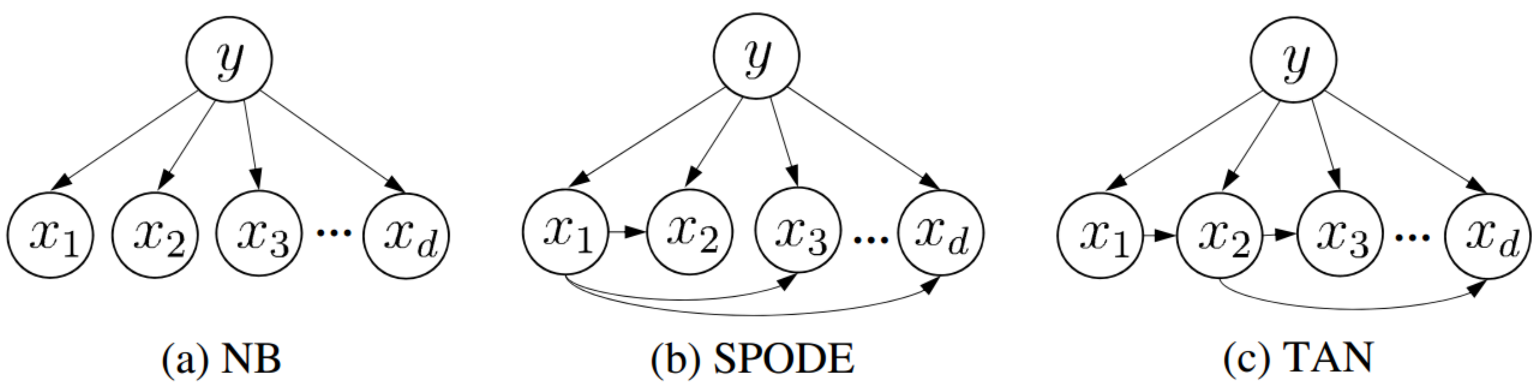
$$P(y|X) = P(y)\prod_{i=1}^{d} P(x_i \mid y, pa_i)$$

where pa_i is the attribute on which attribute x_i depends, and it is referred to as the parent attribute of x_i.

- For each attribute x_i , if its parent attribute is known, the probability can be estimated $P(x_i \mid y, pa_i)$.
- **Problem**: how to determine the parent attribute of each attribute?

SPODE

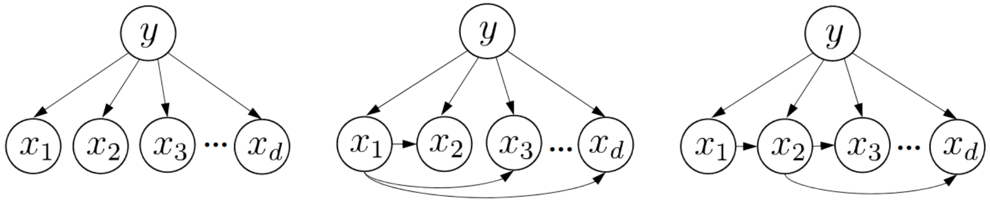
The most straightforward approach is to assume that all attributes depend on a single attribute, called the "super-parent," and then *use model selection methods such as cross-validation to determine the super-parent attribute*, forming the **Super-Parent ODE (SPODE)** method.



e. g. x_1 - is the supert-parent.

TAN

- TAN (**Tree Augmented Naïve Bayes**) [Friedman et al., 1997] based on the Maximum Weighted Spanning Tree algorithm [Chow and Liu, 1968].
- It simplifies the dependencies between attributes shown in the graph (c).



(a) NB (b) SPODE (c) TAN

Steps:

- calculate the conditional mutual information between any two attributes.

$$I(x_i, x_j \mid y) = \sum_{x_i, x_j; c \in y} P(x_i, x_j \mid c) \log \frac{P(x_i, x_j \mid c)}{P(x_i \mid c) P(x_j \mid c)}$$

- construct a complete graph using attributes as nodes, and set the weight of the edge between any two nodes as $I(x_i, x_j \mid y)$
- construct the maximum weighted spanning tree of this complete graph, select a root variable, and set the edges as directed.
- add directed edges from y to each attribute.

AODE

- AODE (**Averaged One-Dependent Estimator**) [Webb et al. 2005] is a more powerful classifier based on ensemble learning mechanism.
- It attempts to *construct SPODE* (Super-Parent One-Dependent Estimator) by taking each attribute as a super-parent, and clusters SPODEs with sufficient training data support together as the final result.

$$P(c \mid \mathbf{x}) \propto \sum_{i=1; |D_{x_i} \geq m'|}^{d} P(c, x_i) \prod_{j=1}^{d} P(x_j \mid c, x_i)$$

$$\hat{P}(x_i, c) = \frac{|D_{c,x_i}| + 1}{|D| + N_i} \qquad \hat{P}(x_j \mid c, x_i) = \frac{|D_{c,x_i,x_j}| + 1}{|D_{c,x_i}| + N_j}$$

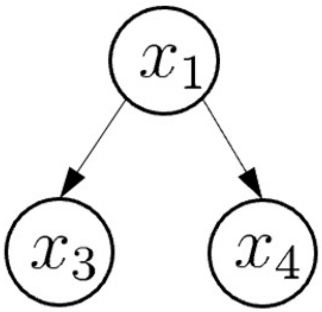
N_i represents the number of values taken on the *i-th* attribute.

m′ is a threshold value.

Bayesian Network

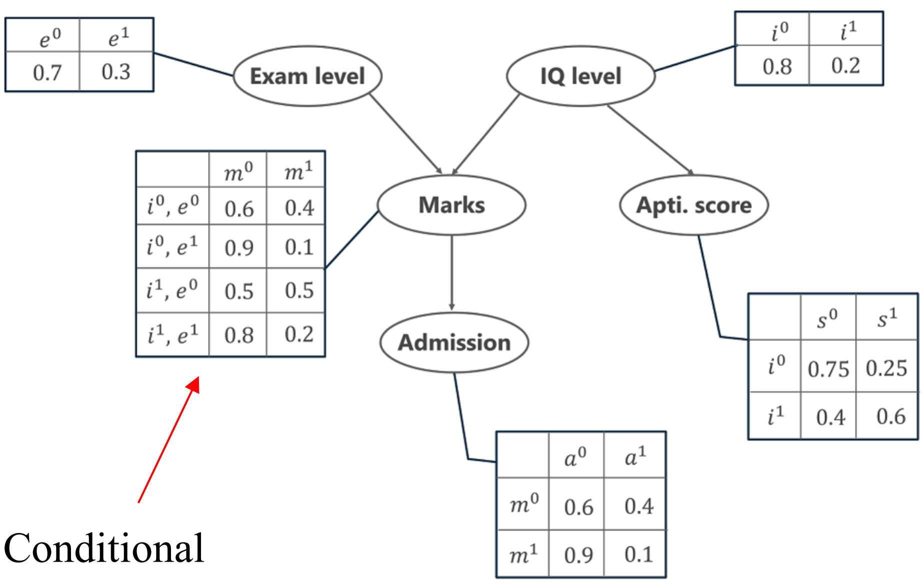
Bayesian network

- **Bayesian network**, also known as a belief network, uses a directed acyclic graph (DAG) to describe the dependency relationship between attributes, and uses conditional probability tables (CPT) to describe the joint probability distribution of attributes.
- Example:



Bayesian network: Example

- Let's creating a Bayesian Network that will model the marks (m) of a student on his examination.



e^0	e^1
0.7	0.3

i^0	i^1
0.8	0.2

	m^0	m^1
i^0, e^0	0.6	0.4
i^0, e^1	0.9	0.1
i^1, e^0	0.5	0.5
i^1, e^1	0.8	0.2

	s^0	s^1
i^0	0.75	0.25
i^1	0.4	0.6

	a^0	a^1
m^0	0.6	0.4
m^1	0.9	0.1

Conditional probability

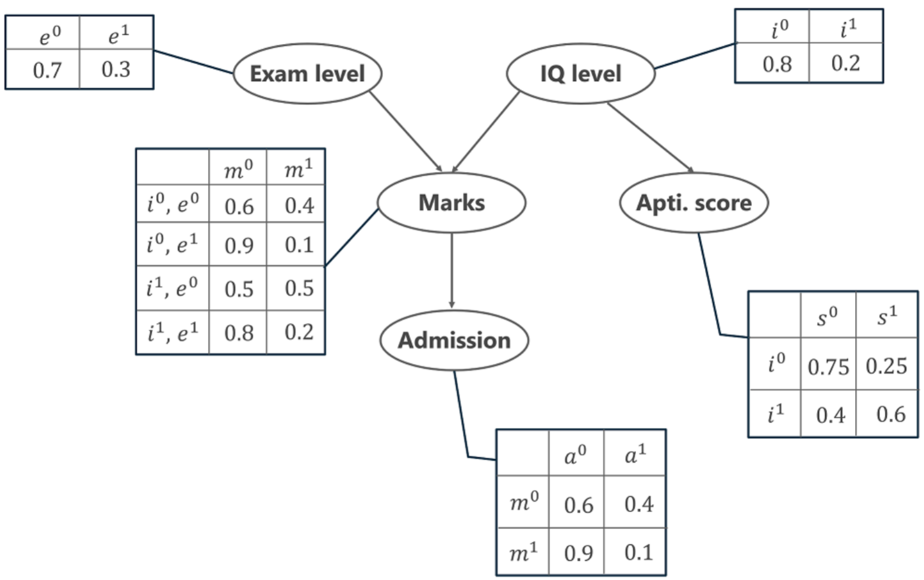
Marks (m) depend on:
Exam level (e): (difficult, easy)
IQ of the student (i): (high, low)

Marks will predict whether or not he/she will get **admitted (a)** to a university.
IQ will also predict the **aptitude score (s)** of the student.

How to calculate the joint probability distribution of these 5 variables?
P(a, m, i, e, s) = ?

Bayesian network: Example

$$P(a, m, i, e, s) = P(a|m) * P(m|\, i, e) * P(i) * P(e) * P(s|i)$$



e^0	e^1
0.7	0.3

i^0	i^1
0.8	0.2

	m^0	m^1
i^0, e^0	0.6	0.4
i^0, e^1	0.9	0.1
i^1, e^0	0.5	0.5
i^1, e^1	0.8	0.2

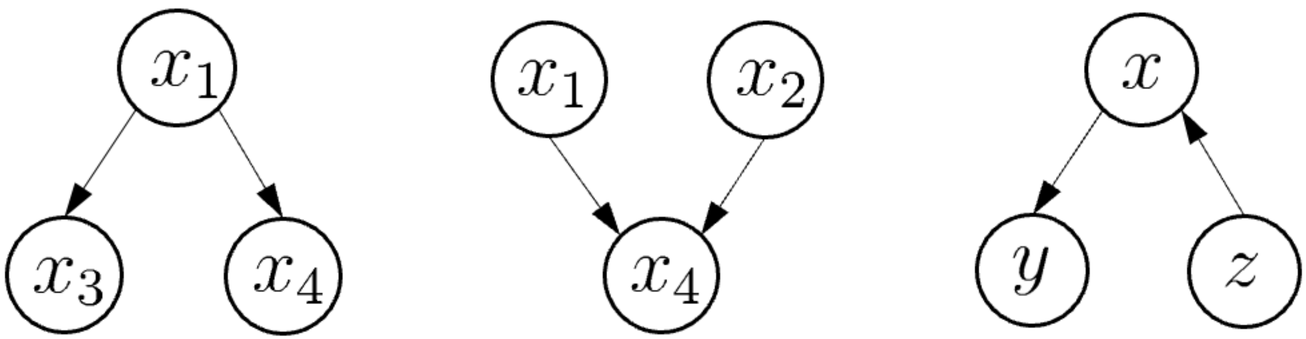
	s^0	s^1
i^0	0.75	0.25
i^1	0.4	0.6

	a^0	a^1
m^0	0.6	0.4
m^1	0.9	0.1

- p(a|m) represents the conditional probability of a student getting an admission based on his marks.
- p(m|i,e) represents the conditional probability of the student's marks, given his IQ level and exam level.
- p(i) denotes the probability of his IQ level (high or low)
- p(e) denotes the probability of the exam level (difficult or easy)
- p(s | i) denotes the conditional probability of his aptitude scores, given his IQ level

Bayesian network: Structure

- The typical dependency relationship between three variables in a Bayesian network:



Gaussian Naïve Bayes Classifiers

Gaussian Naïve Bayes

- A Gaussian naïve is based on continuous variable that are assumed to have a Gaussian (normal) distribution.

Prior $P(y)$:

$$P(y) = \frac{|D_y|}{|D|}$$

Conditional probability $P(x_i|\, y)$:

$$P(x_i|\, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Posterior $P(y|X)$:

$$P(y|X) = P(y) \prod_{i=1}^{d} P(x_i \mid y)$$

Example with Iris Data Set

Iris Data Set

Download: Data Folder, Data Set Description

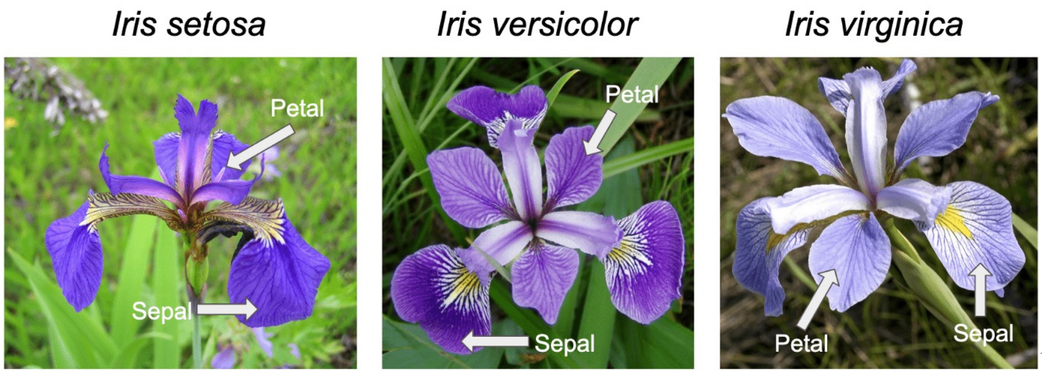
Abstract: Famous database; from Fisher, 1936

Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	5169206

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:

-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica



Example with Iris Data Set

- Data (150 samples)

	sepal length	sepal width	petal length	petal width	label
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	2
146	6.3	2.5	5.0	1.9	2
147	6.5	3.0	5.2	2.0	2
148	6.2	3.4	5.4	2.3	2
149	5.9	3.0	5.1	1.8	2

150 rows × 5 columns

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:

-- Iris Setosa (0)
-- Iris Versicolour (1)
-- Iris Virginica (2)

Gaussian Naïve Bayes

Example of 3 Samples

	sepal length	sepal width	petal length	petal width	label
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0

Iris Setosa(0)

$$P(y=0)=\frac{|D_y|}{|D|}=\frac{3}{6}$$

	sepal length	sepal width	petal length	petal width	label
50	7.0	3.2	4.7	1.4	1
51	6.4	3.2	4.5	1.5	1
52	6.9	3.1	4.9	1.5	1

Iris Versicolour (1)

$$P(y=1)=\frac{|D_y|}{|D|}=\frac{3}{6}$$

Gaussian Naïve Bayes: Calculate Prior

	sepal length	sepal width	petal length	petal width	label
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0

Iris Setosa(0)

	sepal length	sepal width	petal length	petal width	label
50	7.0	3.2	4.7	1.4	1
51	6.4	3.2	4.5	1.5	1
52	6.9	3.1	4.9	1.5	1

Iris Versicolour (1)

Gaussian Naïve Bayes (cont.)

	sepal length	sepal width	petal length	petal width	label
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0

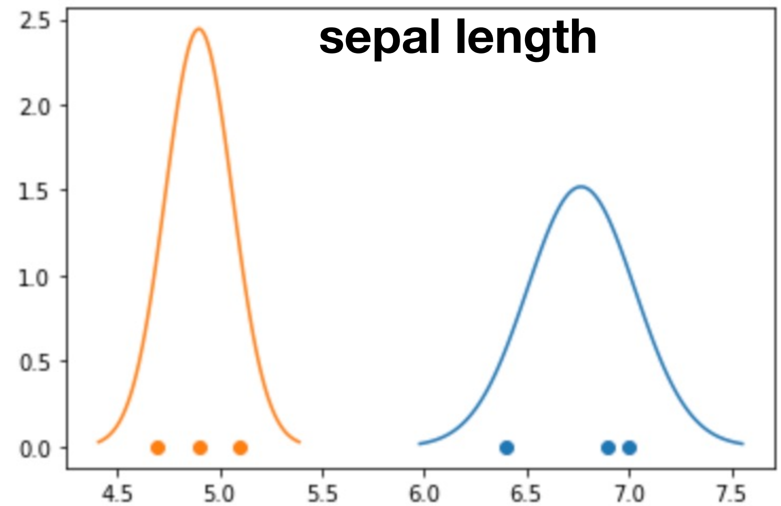
(6.76, 0.262)

	sepal length	sepal width	petal length	petal width	label
50	7.0	3.2	4.7	1.4	1
51	6.4	3.2	4.5	1.5	1
52	6.9	3.1	4.9	1.5	1

Gaussian Naïve Bayes (cont.)

	sepal length	sepal width	petal length	petal width	label
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0

(6.76, 0.262)



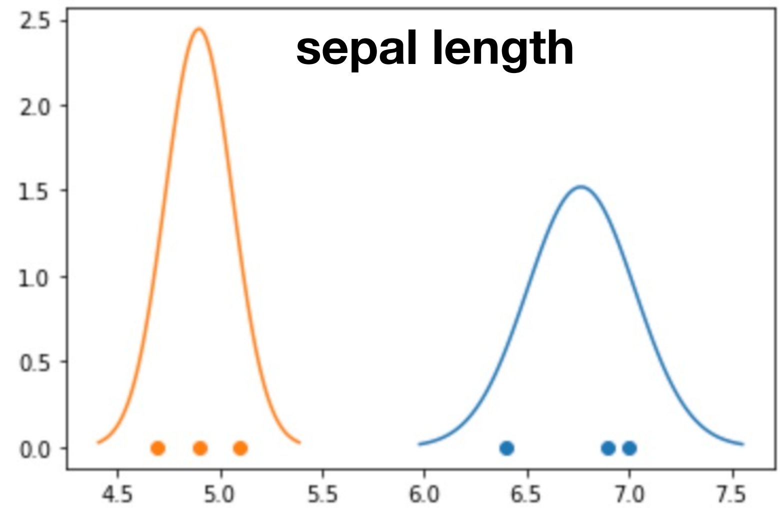
	sepal length	sepal width	petal length	petal width	label
50	7.0	3.2	4.7	1.4	1
51	6.4	3.2	4.5	1.5	1
52	6.9	3.1	4.9	1.5	1

(3.16, 0.047)

Gaussian Naïve Bayes (cont.)

	sepal length	sepal width	petal length	petal width	label
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0

(6.76, 0.262) (4.89, 0.163)



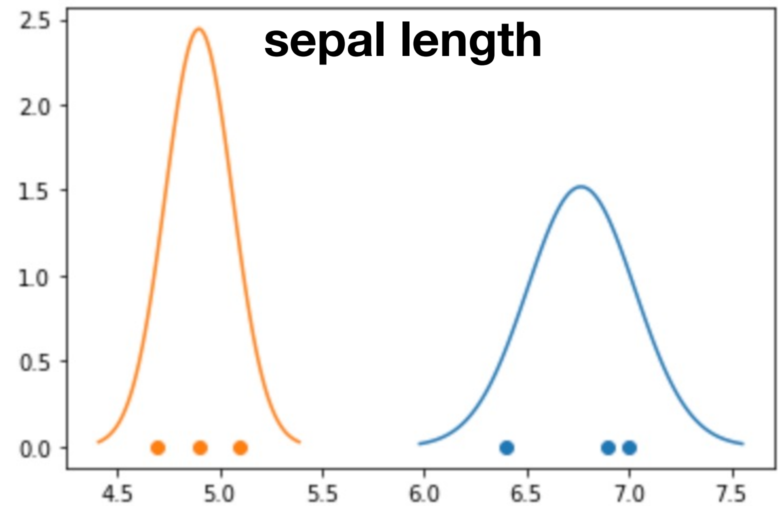
	sepal length	sepal width	petal length	petal width	label
50	7.0	3.2	4.7	1.4	1
51	6.4	3.2	4.5	1.5	1
52	6.9	3.1	4.9	1.5	1

(3.16, 0.047)

Gaussian Naïve Bayes (cont.)

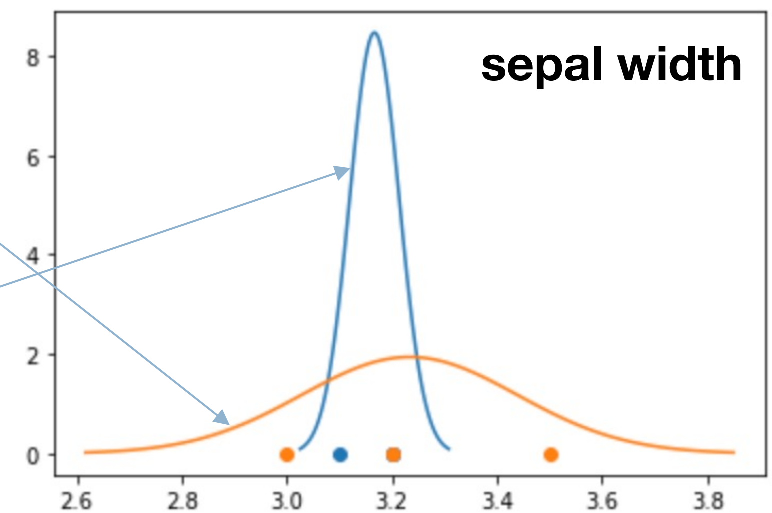
	sepal length	sepal width	petal length	petal width	label
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0

(6.76, 0.262) (4.89, 0.163)



	sepal length	sepal width	petal length	petal width	label
50	7.0	3.2	4.7	1.4	1
51	6.4	3.2	4.5	1.5	1
52	6.9	3.1	4.9	1.5	1

(3.16, 0.047) (3.23, 0.205)



Gaussian Naïve Bayes (cont.)

We get a new sample:

- Sepal length = 6.3cm
- Sepal width = 3cm

$$P(x_i \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Prior:

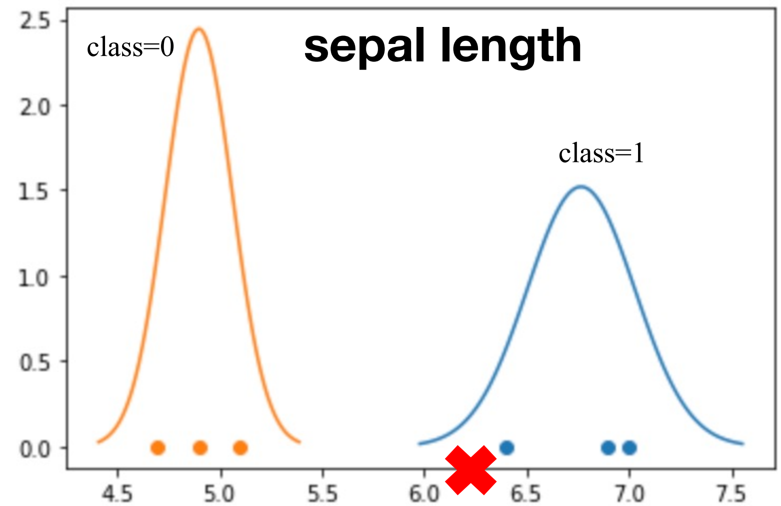
P(new sample = 0) = 0.5

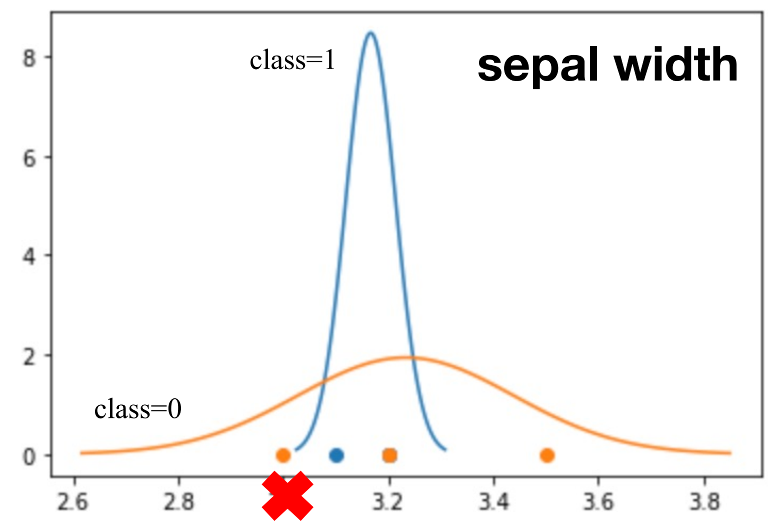
P(new sample = 1) = 0.5

Likelihood:

P(new sample | y = 0) = log(0.04 *6.24e-51)
= -118.819

P(new sample | y = 1) = log(7.18e-116 * 1.397)
= -264.794





- Thank you!