

# Machine Learning in High Dimension

## IA317

### Naive Bayes

Thomas Bonald

2023 – 2024



# Program

1. Nearest neighbors
2. Locally sensitive hashing
3. Dimension reduction
4. Ensemble methods
5. **Naive Bayes**  
→ A statistical approach to **classification**
6. Sparse regression
7. Anomaly detection

## A statistical approach to classification

**Ideal:** Given some data sample  $x \in \mathbb{R}^d$ , predict the probability of its label  $y$  using **Bayes' Theorem**:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)}$$

# A statistical approach to classification

**Ideal:** Given some data sample  $x \in \mathbb{R}^d$ , predict the probability of its label  $y$  using **Bayes' Theorem**:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)}$$

## Algorithm

The decision is then based on the **maximum a posteriori** (MAP):

$$\hat{y} = \arg \max_y p(y|x)$$

**Note:** Data considered as **i.i.d. samples**

## Why “naive” Bayes?

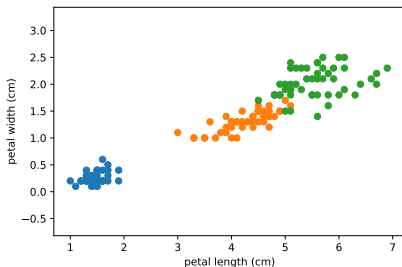
Because of the assumption of **conditional independence** across features:

$$p(x|y) = \prod_{j=1}^d p(x_j|y)$$

# Why “naive” Bayes?

Because of the assumption of **conditional independence** across features:

$$p(x|y) = \prod_{j=1}^d p(x_j|y)$$



Example: 2 dimensions of the **Iris** dataset

# Prior distribution

By **Bayes' Theorem**:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)} \propto \underbrace{p(y)}_{\text{prior}} \underbrace{p(x|y)}_{\text{model}}$$

# Prior distribution

By **Bayes' Theorem**:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)} \propto \underbrace{p(y)}_{\text{prior}} \underbrace{p(x|y)}_{\text{model}}$$

The **prior distribution** can be either...

- ▶ fitted to data

$p(y)$  = fraction of samples with label  $y$

- ▶ or given by some pre-defined distribution (e.g., uniform)



# Outline

1. **Statistical model**

Gaussian, Bernoulli, Multinomial, Categorical

2. A linear classifier

3. Laplace smoothing

4. Missing values

## Principle

Let  $y \in \{1, \dots, L\}$  be some label

### Statistical model

For **each** feature  $j$ , we fit a statistical model with parameter  $\theta_j$  so that

$$p(x_j|y) = p_{\theta_j}(x_j)$$

**Note:** The parameter  $\theta_j$  depends on the label  $y$

# Principle

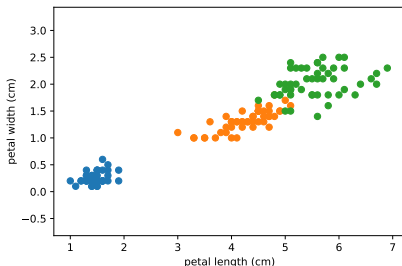
Let  $y \in \{1, \dots, L\}$  be some label

## Statistical model

For **each** feature  $j$ , we fit a statistical model with parameter  $\theta_j$  so that

$$p(x_j|y) = p_{\theta_j}(x_j)$$

**Note:** The parameter  $\theta_j$  depends on the label  $y$



Example: 2 dimensions of the **Iris** dataset

# Principle

Let  $y \in \{1, \dots, L\}$  be some label

## Statistical model

For **each** feature  $j$ , we fit a statistical model with parameter  $\theta_j$  so that

$$p(x_j|y) = p_{\theta_j}(x_j)$$

**Note:** The parameter  $\theta_j$  depends on the label  $y$

By the **conditional independence** assumption, we get:

$$p(x|y) = \prod_{j=1}^d p_{\theta_j}(x_j)$$

# 1. Gaussian model

For **numerical** data  $x \in \mathbb{R}^d$

Let  $j \in \{1, \dots, d\}$  be some feature.

We use the simple notation  $x \equiv x_j$  and  $\theta \equiv \theta_j$ .

## Statistical model

Probability density function:

$$p_{\theta}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

with  $\theta = (\mu, \sigma^2)$

# 1. Gaussian model

For **numerical** data  $x \in \mathbb{R}^d$

Let  $j \in \{1, \dots, d\}$  be some feature.

We use the simple notation  $x \equiv x_j$  and  $\theta \equiv \theta_j$ .

## Statistical model

Probability density function:

$$p_{\theta}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

with  $\theta = (\mu, \sigma^2)$

## Maximum likelihood

$\mu$  and  $\sigma^2$  given by the **empirical mean** and **variance**  
(for each label  $y$  and each feature  $j$ )

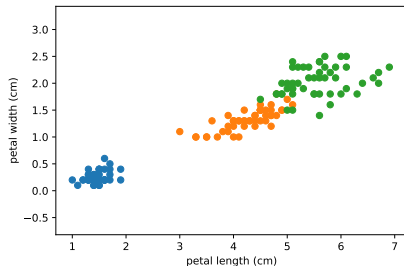
## Example: Iris

$$X \in \mathbb{R}^{n \times d}$$

$n = 150$  samples

$d = 4$

3 classes



Example: 2 dimensions of the **Iris** dataset

## Example: MNIST

$$X \in \{0, \dots, 255\}^{n \times d}$$

$n = 10,000$  samples

$$d = 28 \times 28 = 784$$



Samples



Means



Generated samples



## 2. Bernoulli model

For **binary** data  $x \in \{0, 1\}^d$

Let  $j \in \{1, \dots, d\}$  be some feature.

We use the simple notation  $x \equiv x_j$  and  $\theta \equiv \theta_j$ .

### Statistical model

Probability distribution:

$$p_{\theta}(x) = \theta^x (1 - \theta)^{1-x} \quad \theta \in [0, 1]$$

## 2. Bernoulli model

For **binary** data  $x \in \{0, 1\}^d$

Let  $j \in \{1, \dots, d\}$  be some feature.

We use the simple notation  $x \equiv x_j$  and  $\theta \equiv \theta_j$ .

### Statistical model

Probability distribution:

$$p_{\theta}(x) = \theta^x (1 - \theta)^{1-x} \quad \theta \in [0, 1]$$

### Maximum likelihood

$\theta$  is the **empirical mean** (fraction of 1s)  
(for each label  $y$  and each feature  $j$ )

## Example: MNIST

$$X \in \{0, \dots, 255\}^{n \times d}$$

$n = 10,000$  samples

$$d = 28 \times 28 = 784$$



2	8	1	1	1	5	6	7	9	8
2	9	5	3	1	3	8	2	6	1
2	8	1	5	1	3	5	6	8	7
4	8	8	3	1	1	3	3	3	1
7	9	0	7	1	6	2	3	1	3
7	9	2	7	3	0	1	9	1	1
6	6	5	1	4	6	8	8	9	6
0	0	1	6	4	9	9	7	1	0
2	4	1	3	1	7	0	7	4	7
3	2	4	4	1	0	2	2	3	1

Samples



0	1	2	3	4
5	6	7	8	9

Parameters



0	1	2	3	4
5	6	7	8	9

Generated samples

### 3. Multinomial model

For **count** data  $x \in \mathbb{N}^d$

#### Statistical model

Probability distribution:

$$p_{\theta}(x) = \binom{|x|}{x_1, \dots, x_d} \prod_{j=1}^d \theta_j^{x_j}$$

with  $\theta \geq 0$ ,  $\sum_{j=1}^d \theta_j = 1$ .

### 3. Multinomial model

For **count** data  $x \in \mathbb{N}^d$

#### Statistical model

Probability distribution:

$$p_{\theta}(x) = \binom{|x|}{x_1, \dots, x_d} \prod_{j=1}^d \theta_j^{x_j}$$

with  $\theta \geq 0$ ,  $\sum_{j=1}^d \theta_j = 1$ .

#### Maximum likelihood

$\theta_1, \dots, \theta_d$  is the **empirical distribution** of counts over features (for each label  $y$ )

## Example: MNIST

$$X \in \{0, \dots, 255\}^{n \times d}$$

$n = 10,000$  samples

$$d = 28 \times 28 = 784$$



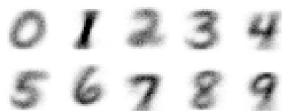
2 8 1 1 1 5 6 7 9 8  
2 9 5 3 1 3 8 2 6 1  
2 8 1 5 1 3 5 6 8 7  
4 8 8 3 1 1 3 3 3 1  
7 9 0 7 1 6 2 3 1 3  
7 9 2 7 3 0 1 9 1 1  
6 6 5 1 4 6 8 8 9 6  
0 0 1 6 4 9 9 7 1 0  
2 4 1 3 1 7 0 7 4 7  
3 2 4 4 1 0 2 2 3 1

Samples



0 1 2 3 4  
5 6 7 8 9

Parameters



0 1 2 3 4  
5 6 7 8 9

Generated samples

## First property: Conditional independence

Although the features are **not** conditionally independent in the multinomial model, everything works **as if** they were:

$$\begin{aligned}\forall x \in \mathbb{N}^d, \quad p(y|x) &\propto \underbrace{p(y)}_{\text{prior}} \underbrace{p(x|y)}_{\text{multinomial}} \\ &\propto p(y) \binom{|x|}{x_1, \dots, x_d} \prod_{j=1}^d \theta_j^{x_j} \\ &\propto p(y) \prod_{j=1}^d \theta_j^{x_j}\end{aligned}$$

The parameters  $\theta_1, \dots, \theta_d$  depend on the label  $y$  and sum to 1.

## Second property: Positive features

For **positive features**, the multinomial model is equivalent to the **exponential model** with dependent parameters:

$$\begin{aligned}\forall x \in \mathbb{R}_+^d, \quad p(y|x) &\propto \underbrace{p(y)}_{\text{prior}} \underbrace{p(x|y)}_{\text{multinomial}} \\ &\propto p(y) \prod_{j=1}^d \theta_j^{x_j} \\ &\propto p(y) \prod_{j=1}^d e^{x_j \log \theta_j}\end{aligned}$$

The parameters  $\theta_1, \dots, \theta_d$  depend on the label  $y$  and sum to 1.



## 4. Categorical model

For **categorical** data  $x$ .

Let  $j \in \{1, \dots, d\}$  be some feature.

We use the simple notation  $x \equiv x_j$  and  $\theta \equiv \theta_j$ .

### Statistical model

Probability distribution over  $K$  categories:

$$p_{\theta} = (\theta_1, \dots, \theta_K)$$

with  $\theta \geq 0$ ,  $\sum_{k=1}^K \theta_k = 1$ .

## 4. Categorical model

For **categorical** data  $x$ .

Let  $j \in \{1, \dots, d\}$  be some feature.

We use the simple notation  $x \equiv x_j$  and  $\theta \equiv \theta_j$ .

### Statistical model

Probability distribution over  $K$  categories:

$$p_\theta = (\theta_1, \dots, \theta_K)$$

with  $\theta \geq 0$ ,  $\sum_{k=1}^K \theta_k = 1$ .

### Maximum likelihood

$\theta$  is the **empirical distribution** over the categories  
(for each label  $y$  and each feature  $j$ )

# Categorical vs. Multinomial

The categorical model:

$$p(x = k) = \theta_k$$

is equivalent to the **multinomial model** after **one-hot encoding**:

$$p(x^{\text{binary}}) = \theta_k$$

with

$$x^{\text{binary}} = (0, \dots, 0, \underbrace{1}_k, 0, \dots, 0)$$

# Overview

Model	Binary	Count	Positive	Numerical	Categorical
Gaussian	(✓)	(✓)	(✓)	✓	✗
Bernoulli	✓	✗	✗	✗	✗
Multinomial	(✓)	✓	✓	✗	✗
Categorical	(✓)	✗	✗	✗	✓

✓ Yes

✗ No

(✓) Applicable

# Outline

1. Statistical model  
Gaussian, Bernoulli, Multinomial, Categorical
2. **A linear classifier**
3. Laplace smoothing
4. Missing values

## Observation

Consider the log-likelihood:

$$\begin{aligned}\log p(y|x) &= \log \frac{p(y)p(x|y)}{p(x)} \\ &= \log p(y) + \log p(x|y) - \underbrace{\log p(x)}_{\text{constant}} \\ &= \log p(y) + \sum_{j=1}^d \log p(x_j|y) + c\end{aligned}$$

## Prediction

$$\hat{y} = \arg \max_y \log p(y|x)$$

# 1. Gaussian model

The log-likelihood is:

$$\begin{aligned}\log p(y|x) &= \log p(y) - \sum_{j=1}^d \frac{(x_j - \mu_j)^2}{2\sigma_j^2} - \sum_{j=1}^d \log \sqrt{2\pi\sigma_j^2} + c \\ &= w^T \phi(x) + b \quad \text{with} \quad \phi(x) = (x, x^2)\end{aligned}$$

where  $w, b$  depend on the label  $y$

# 1. Gaussian model

The log-likelihood is:

$$\begin{aligned}\log p(y|x) &= \log p(y) - \sum_{j=1}^d \frac{(x_j - \mu_j)^2}{2\sigma_j^2} - \sum_{j=1}^d \log \sqrt{2\pi\sigma_j^2} + c \\ &= w^T \phi(x) + b \quad \text{with} \quad \phi(x) = (x, x^2)\end{aligned}$$

where  $w, b$  depend on the label  $y$

## Property

Gaussian Naive Bayes is a **linear classifier** in the feature space  $(x, x^2)$

**Example:** For binary labels  $y \in \{+, -\}$ :

$$\hat{y} = \text{sign}(w^T \phi(x) + b)$$

with  $w = w_+ - w_-$ ,  $b = b_+ - b_-$



## 2. Bernoulli model

The log-likelihood is:

$$\begin{aligned}\log p(y|x) &= \log p(y) + \sum_{j=1}^d (x_j \log \theta_j + (1 - x_j) \log(1 - \theta_j)) + c \\ &= w^T x + b\end{aligned}$$

where  $w, b$  depend on the label  $y$

## 2. Bernoulli model

The log-likelihood is:

$$\begin{aligned}\log p(y|x) &= \log p(y) + \sum_{j=1}^d (x_j \log \theta_j + (1 - x_j) \log(1 - \theta_j)) + c \\ &= w^T x + b\end{aligned}$$

where  $w, b$  depend on the label  $y$

### Property

Bernoulli Naive Bayes is a **linear classifier**

**Example:** For binary labels  $y \in \{+, -\}$ :

$$\hat{y} = \text{sign}(w^T x + b)$$

with  $w = w_+ - w_-$ ,  $b = b_+ - b_-$

### 3. Multinomial model

The log-likelihood is:

$$\begin{aligned}\log p(y|x) &= \log p(y) + \sum_{j=1}^d x_j \log \theta_j + c \\ &= w^T x + b\end{aligned}$$

where  $w, b$  depend on the label  $y$

### 3. Multinomial model

The log-likelihood is:

$$\begin{aligned}\log p(y|x) &= \log p(y) + \sum_{j=1}^d x_j \log \theta_j + c \\ &= w^T x + b\end{aligned}$$

where  $w, b$  depend on the label  $y$

#### Property

Multinomial Naive Bayes is a **linear classifier**

**Example:** For binary labels  $y \in \{+, -\}$ :

$$\hat{y} = \text{sign}(w^T x + b)$$

with  $w = w_+ - w_-$ ,  $b = b_+ - b_-$

## 4. Categorical model

We use the equivalence with the **Multinomial** model

### Property

Categorical Naive Bayes is a **linear classifier** after one-hot encoding of the categories

**Example:** For binary labels  $y \in \{+, -\}$ :

$$\hat{y} = \text{sign}(w^T x^{\text{binary}} + b)$$

with  $w = w_+ - w_-$ ,  $b = b_+ - b_-$

# Outline

1. Statistical model  
Gaussian, Bernoulli, Multinomial, Categorical
2. A linear classifier
3. **Laplace smoothing**
4. Missing values

# The sunrise problem

What is the **probability** that the sun will rise tomorrow?



# Laplace's answer

A **Bayesian** approach!

- ▶ Let  $\theta$  be the **unknown** probability that the sun will rise
- ▶ Assume a **uniform prior** distribution<sup>1</sup> on  $\theta$

---

<sup>1</sup>When the probability of a single event is unknown we may suppose it equal to any value from zero to unity.

Pierre-Simon, Marquis de Laplace, A philosophical Essay on Probabilities, 1814.



# Laplace's answer

A **Bayesian** approach!

- ▶ Let  $\theta$  be the **unknown** probability that the sun will rise
- ▶ Assume a **uniform prior** distribution on  $\theta$

## Bayes' estimator

Given  $n$  observations that the sun rose,

$$\hat{\theta} = \frac{n+1}{n+2}$$

**Note:** As if there were 2 (fake) observations<sup>2</sup>, one where the sun rose, another where it didn't!

---

<sup>2</sup>Thus we find that an event having occurred successively any number of times, the probability that it will happen again the next time is equal to this number increased by unity divided by the same number, increased by two units. Pierre-Simon, Marquis de Laplace, A philosophical Essay on Probabilities, 1814.

## The problem of unseen values

- ▶ Consider the **Gaussian model** and assume  $\mu_1 = 0$ ,  $\sigma_1 = 0$  for each label  $y$ . What if  $x_1 = 1$ ?

## The problem of unseen values

- Consider the **Gaussian model** and assume  $\mu_1 = 0$ ,  $\sigma_1 = 0$  for each label  $y$ . What if  $x_1 = 1$ ?

For each label  $y$ ,

$$p(y|x) \propto p(y) \underbrace{p(x|y)}_{=0} = 0$$

## The problem of unseen values

- ▶ Consider the **Gaussian model** and assume  $\mu_1 = 0$ ,  $\sigma_1 = 0$  for each label  $y$ . What if  $x_1 = 1$ ?

For each label  $y$ ,

$$p(y|x) \propto p(y) \underbrace{p(x|y)}_{=0} = 0$$

- ▶ Consider the **Bernoulli model** and assume  $\theta_1 = 1$  for each label  $y$ . What if  $x_1 = 0$ ?

## The problem of unseen values

- Consider the **Gaussian model** and assume  $\mu_1 = 0$ ,  $\sigma_1 = 0$  for each label  $y$ . What if  $x_1 = 1$ ?

For each label  $y$ ,

$$p(y|x) \propto p(y) \underbrace{p(x|y)}_{=0} = 0$$

- Consider the **Bernoulli model** and assume  $\theta_1 = 1$  for each label  $y$ . What if  $x_1 = 0$ ?

For each label  $y$ ,

$$p(y|x) \propto p(y) \underbrace{p(x|y)}_{=0} = 0$$

# 1. Gaussian model

For **numerical** feature  $x \in \mathbb{R}$

## Statistical model

Probability density function:

$$p_{\theta}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

with  $\theta = (\mu, \sigma^2)$

# 1. Gaussian model

For **numerical** feature  $x \in \mathbb{R}$

## Statistical model

Probability density function:

$$p_{\theta}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

with  $\theta = (\mu, \sigma^2)$

## Laplace smoothing

$$\sigma^2 \leftarrow \frac{\text{total square error} + 1}{\text{nb of samples} + 1}$$

**Note:**  $\sigma^2 > 0$

## Exercise

What are the parameters of the **Gaussian model** with Laplace smoothing?

$x_1$	$x_2$	$y$
0	1	0
1	1	0
0	1	0
1	0	1



## 2. Bernoulli model

For **binary** feature  $x \in \{0, 1\}$

### Statistical model

Probability distribution:

$$p_{\theta}(x) = \theta^x (1 - \theta)^{1-x} \quad \theta \in [0, 1]$$

## 2. Bernoulli model

For **binary** feature  $x \in \{0, 1\}$

### Statistical model

Probability distribution:

$$p_{\theta}(x) = \theta^x (1 - \theta)^{1-x} \quad \theta \in [0, 1]$$

### Laplace smoothing

$$\theta \leftarrow \frac{\text{nb of 1s} + 1}{\text{nb of samples} + 2}$$

**Note:** We always get  $\theta \in (0, 1)$

## Exercise

What are the parameters of the **Bernoulli model** with Laplace smoothing?

$x_1$	$x_2$	$y$
0	1	0
1	1	0
0	1	0
1	0	1

### 3. Multinomial model

For **count** data  $x \in \mathbb{N}^d$

#### Multinomial model

Probability distribution:

$$p_{\theta}(x) = \binom{|x|}{x_1, \dots, x_d} \prod_{j=1}^d \theta_j^{x_j} \quad \sum_{j=1}^d \theta_j = 1$$

### 3. Multinomial model

For **count** data  $x \in \mathbb{N}^d$

#### Multinomial model

Probability distribution:

$$p_{\theta}(x) = \binom{|x|}{x_1, \dots, x_d} \prod_{j=1}^d \theta_j^{x_j} \quad \sum_{j=1}^d \theta_j = 1$$

#### Laplace smoothing

$$\theta_j \leftarrow \frac{\text{count of } j + 1}{\text{total count} + d}$$

**Note:** We get  $\theta_j > 0$  for all  $j$

## Exercise

What are the parameters of the **Multinomial model** with Laplace smoothing?

$x_1$	$x_2$	$x_3$	$y$
0	3	1	0
1	2	0	0
0	1	0	0
1	0	2	1

## 4. Categorical model

For **categorical** feature  $x \in \{1, \dots, K\}$

### Statistical model

Probability distribution:

$$p_{\theta} = (\theta_1, \dots, \theta_K) \quad \text{with} \quad \sum_k \theta_k = 1$$

## 4. Categorical model

For **categorical** feature  $x \in \{1, \dots, K\}$

### Statistical model

Probability distribution:

$$p_{\theta} = (\theta_1, \dots, \theta_K) \quad \text{with} \quad \sum_k \theta_k = 1$$

### Laplace smoothing

$$\theta_k \leftarrow \frac{\text{nb in category } k + 1}{\text{nb of samples} + K}$$

**Warning:** In general, **not** equivalent to Laplace smoothing of the multinomial model after one-hot encoding!



## Exercise

What are the parameters of the **Categorical model** with Laplace smoothing?

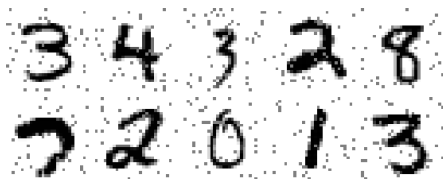
$x_1$	$x_2$	$y$
a	B	0
a	B	0
b	A	0
c	B	1

# Outline

1. Statistical model  
Gaussian, Bernoulli, Multinomial, Categorical
2. A linear classifier
3. Laplace smoothing
4. **Missing values**

# Motivation

What if some values are **missing**?  
(typically, coded as NaN)



Example: MNIST samples with missing pixels

# Prediction

Given a sample  $x \in \mathbb{R}^d$ , predict its label  $y$  using **Bayes' Theorem**:

$$p(y|x) \propto \underbrace{p(y)}_{\text{prior}} \underbrace{p(x|y)}_{\text{model}}$$

and **conditional independence**:

$$p(x|y) = \prod_{j: x_j \text{ not missing}} p_{\theta_j}(x_j)$$

# Fit

Fit the statistical model of each feature on **non-missing** values

## Gaussian model

$\theta \leftarrow$  mean and (smoothed) variance of **non-missing** values

## Bernoulli model

$$\theta \leftarrow \frac{\text{nb of 1s} + 1}{\text{nb of } \mathbf{non-missing} \text{ samples} + 2}$$

## Multinomial model

$$\theta_j \leftarrow \frac{\text{counts of } j + 1}{\text{counts of } \mathbf{non-missing} \text{ samples} + d}$$

## Exercise

What are the parameters of the **Bernoulli model** with Laplace smoothing?

$x_1$	$x_2$	$y$
0	NaN	0
1	NaN	0
0	1	0
NaN	0	1

# Summary

## Naive Bayes

- ▶ A statistical approach to **classification**  
Relies on **conditional independence**  
Efficient in **high dimension**
- ▶ A **linear classifier**  
Like logistic regression, but with **explicit** weights
- ▶ Importance of **smoothing**
- ▶ Natural extension to **missing values**

