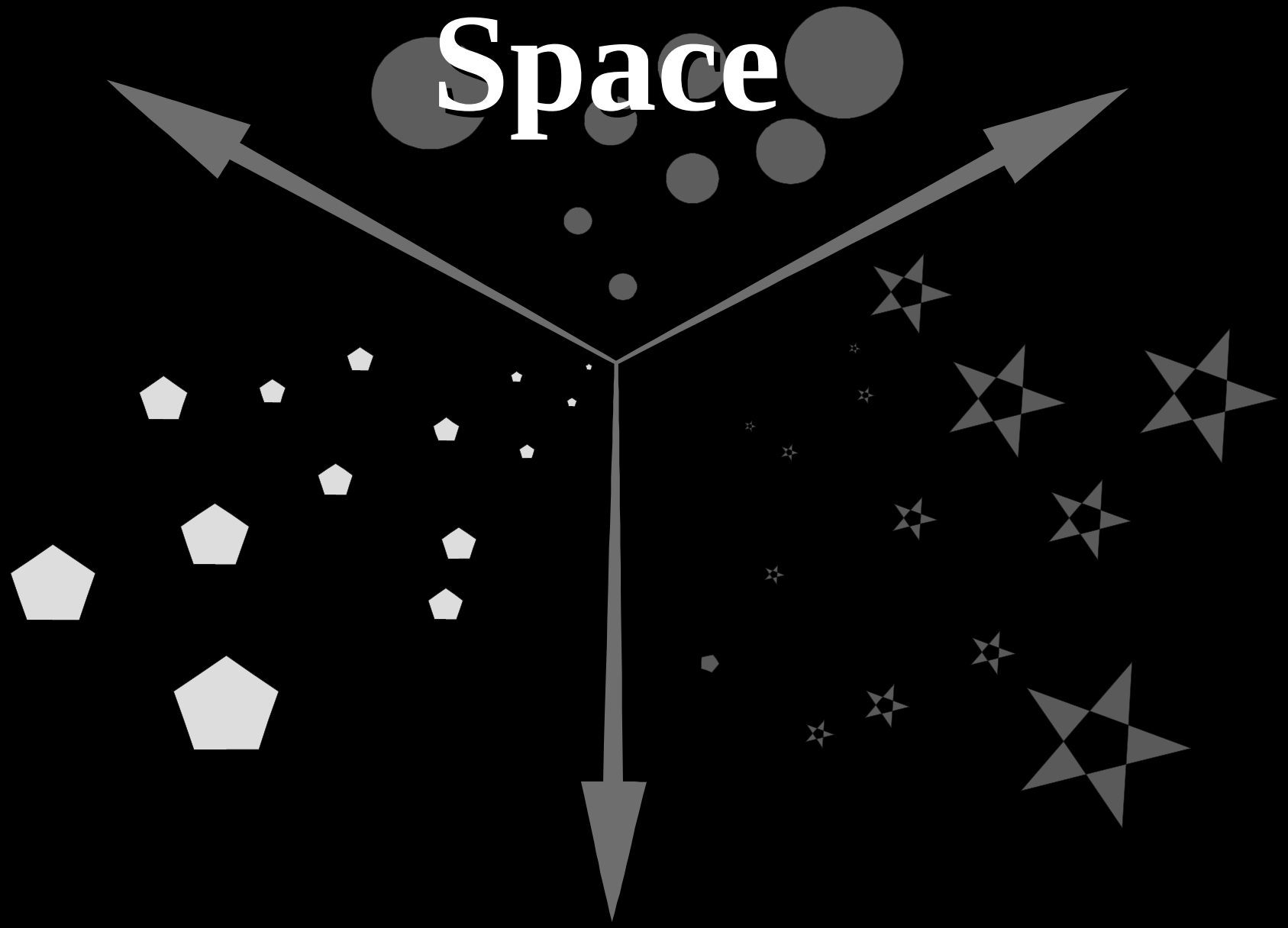
Classification in Pattern Space



Mathias Mamsch, Summer University 2002

Classification in Pattern S

Topics

1. Introduction
   1.1. Why classification ?
   1.2. Different methods for classification
2. Formal view
   2.1. Statistical modell of speech creation
   2.2. Classification functions
   2.3. The optimal decision rule
3. Learning methods
   3.1. Maximum Likelihood
   3.2. The EM-Algorithm
   3.3. Examples
4. Experiments

Classification in Patterns

1. Introduction

1.1 Why classification ?

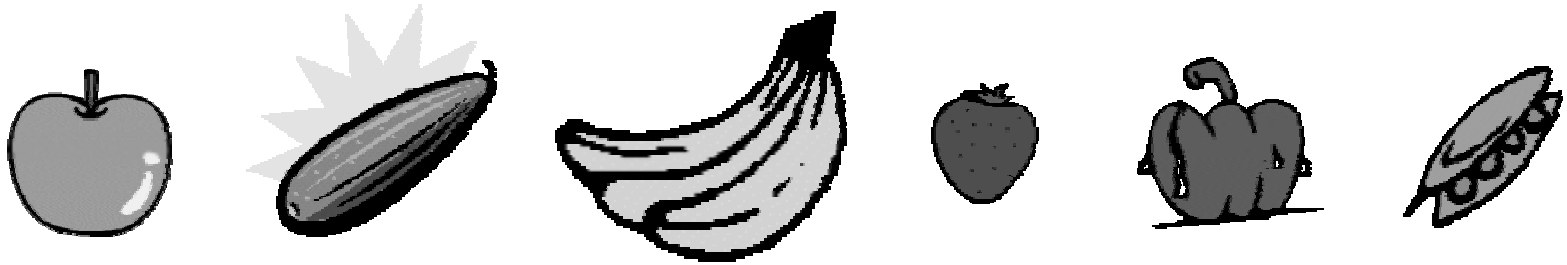
Classification in Patterns

1. Introduction

1.1 Why classification ?

- What is classification ?

Fruit or vegetable ?

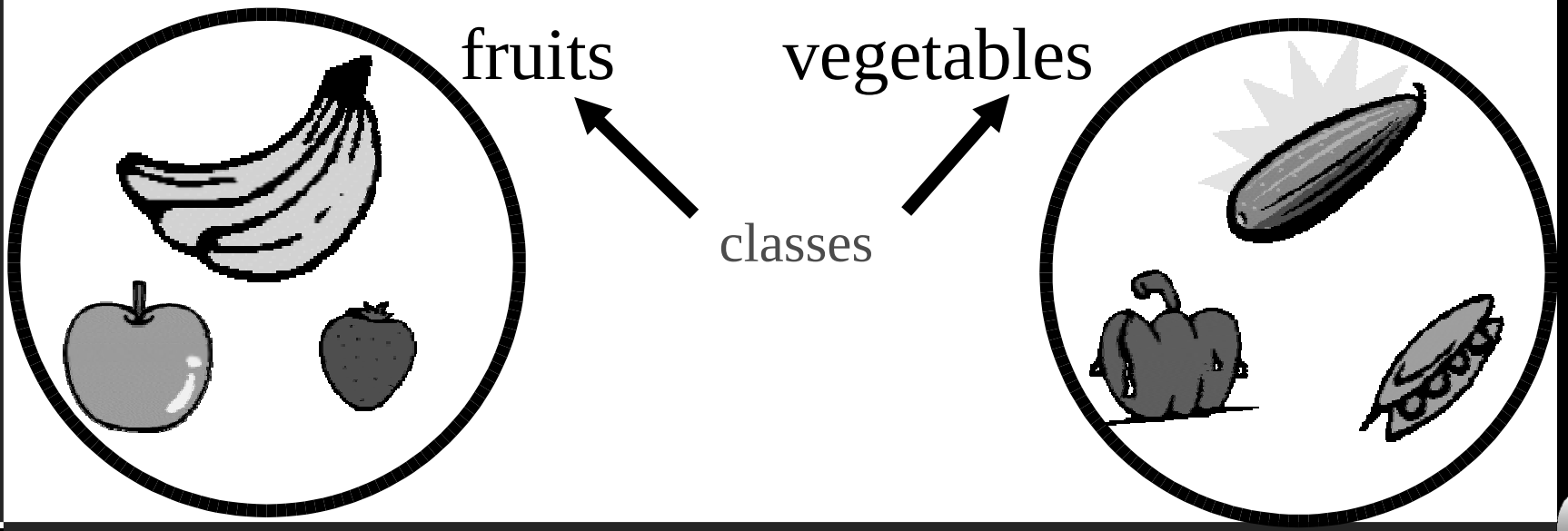


Classification in Pattern S

1. Introduction

1.1 Why classification ?

- What is classification ?



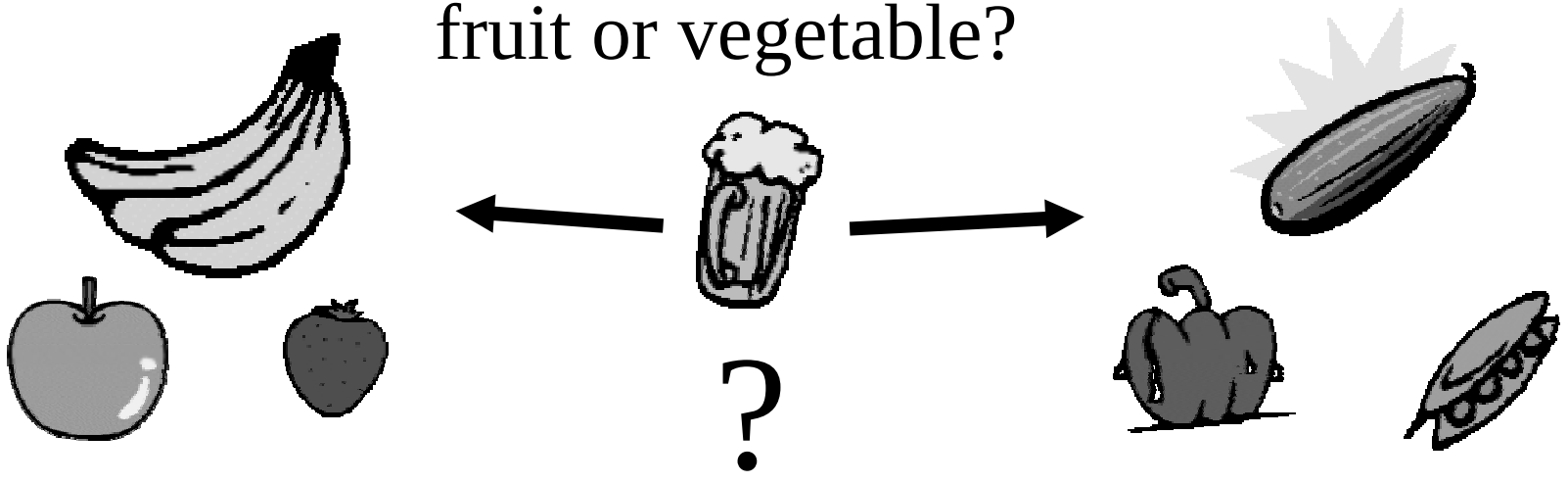
Classification in Pattern S

1. Introduction

1.1 Why classification ?

- What is classification ?

fruit or vegetable?

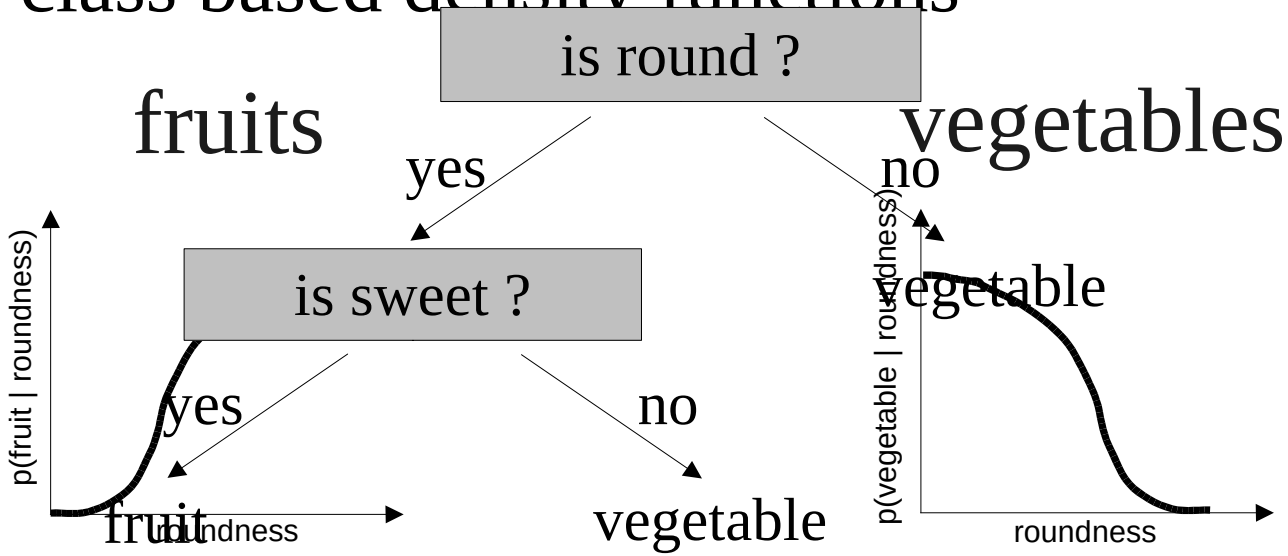


Classification in Patterns

1. Introduction

1.2 Different methods for classification

- decision trees
- artificial neural networks
- → class based density functions



Classification in Pattern S

2. Formal view

Situation for a typical classification problem:

- you have some data / objects you want to classify
- you have some preclassified data (optional)

Then you need to:

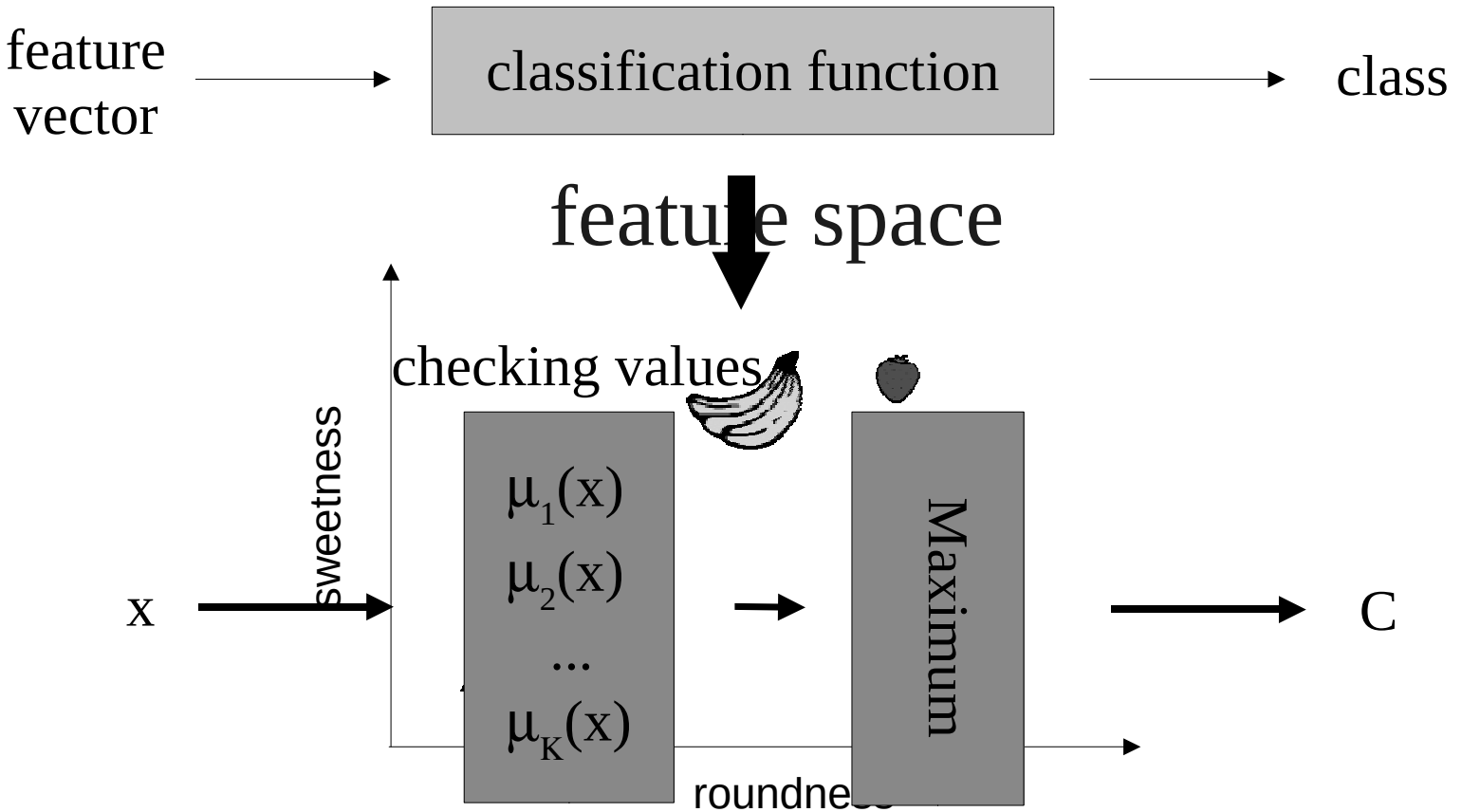
- define some features (eg. roundness, sweetness, ...)
- do a feature-extraction for your data

Result

	apple	strawberry	banana	cucumber	peas	paprica
roundness:	0.9	0.7	0.5	0.3	0.25	0.7
sweetness:	0.6	0.8	0.7	0.25	0.3	0.2

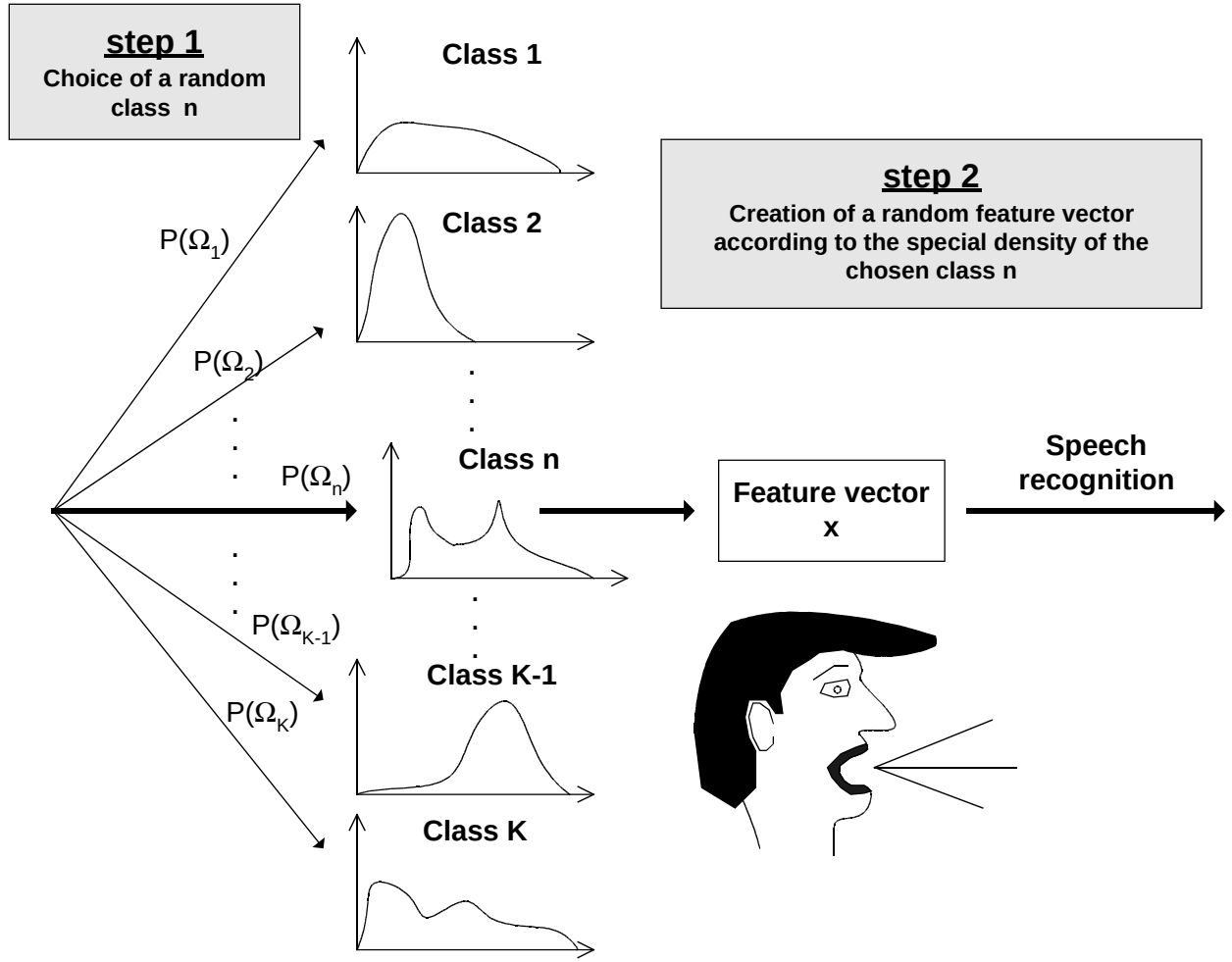
Classification in Pattern S

2. Formal view



Classification in Pattern S

2.1. Statistical modell of speech creation



Classification in Pattern S

2.1. Statistical modell of speech creation

$$P(\Omega_i), i=1 \ldots K \quad \text{and} \quad \sum_{i=1}^{K} P(\Omega_i)=1$$

$$P(\boldsymbol{x} \mid \Omega_i) \quad \text{and} \quad \int_{\mathbb{R}^D} P(\boldsymbol{x} \mid \Omega_i)\, \boldsymbol{dx}=1$$

but we get a temporal stream of feature vectors:

$$\underset{t=1}{\begin{bmatrix} \boldsymbol{x_1} \end{bmatrix}}, \underset{t=2}{\begin{bmatrix} \boldsymbol{x_2} \end{bmatrix}}, \underset{t=3}{\begin{bmatrix} \boldsymbol{x_3} \end{bmatrix}}, \cdots \qquad P(\boldsymbol{x})=\int_{-\infty}^{\infty} P(\boldsymbol{x}, \Omega)\, d\,\Omega$$

How to approximate P(x | Ω) and P(Ω) ?
→ learning problem (later in this lecture)

Classification in Patterns

2.2. Classification functions

Intuitive: Maximum Likelihood (ML)

→ take the class which has the highest probability to produce **x**

$$\Omega_i \quad \text{with} \quad \max_i P(\boldsymbol{x} \mid \Omega_i)$$

less Intuitive: Maximum a-posteriori (MAP)

→ take the class that is likeliest to produce **x**

$$\Omega_i \quad \text{with} \quad \max_i P(\Omega_i \mid \boldsymbol{x})$$

$$P(\Omega_i \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x} \mid \Omega_i) \cdot \boldsymbol{P}(\Omega_i)}{P(\boldsymbol{x})}$$

constant

$P(x \mid \Omega_1) = 0.6$ $\quad P(x \mid \Omega_2) = 0.3$

$P(\Omega_1) = 0.01$ $\quad P(\Omega_2) = 0.99$

Classification in Pattern S

2.2. Classification functions

Empirical Risk Minimization

$\mathbf{r} = \{r_{ik}\}$

$\rightarrow$ risk of classifying a vector of class i in class j

Expected risk:

$$R(\boldsymbol{x}) = \sum_{i=1}^{K} P(\Omega_i) \cdot P(\boldsymbol{x} \mid \Omega_i) \boldsymbol{R}(\boldsymbol{x} \mid \Omega_i)$$

$$\boldsymbol{R}(\boldsymbol{x} \mid \Omega_j) = \sum_{j=1}^{K} r_{i,j} \cdot \delta(\Omega_j \mid \boldsymbol{x})$$

$$R(\boldsymbol{x}) = \sum_{j}^{K} \underbrace{\delta(\Omega_j \mid \boldsymbol{x})}_{\delta_j} \underbrace{\sum_{i=1}^{K} r_{i,j} P(\Omega_i) \cdot P(\boldsymbol{x} \mid \Omega_i)}_{\mu_j}$$

$R(x \mid \Omega_i)$ – expected risk for a vector **x** of class i

$\delta(\Omega_j \mid x)$ – probability for decision for for class j given the vector **x**

Classification in Pattern S

2.3. The optimal decision rule

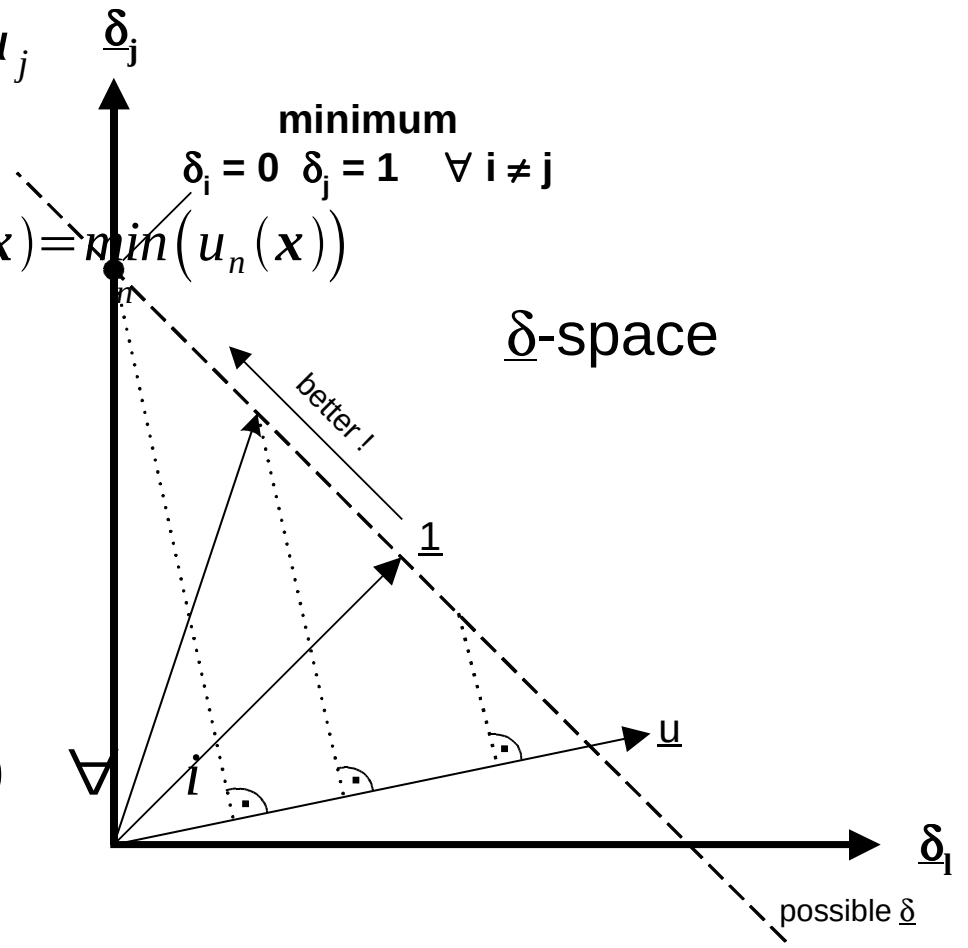
$$\min_{\delta_j} R(\boldsymbol{x}) = \sum_{j}^{K} \delta_j \cdot \mu_j$$

$$\delta^* = \begin{cases} \mathbf{1} & \text{falls } u_j(\boldsymbol{x}) = min\left(u_n(\boldsymbol{x})\right) \\ \mathbf{0} & \text{sonst} \end{cases}$$

Hard decision !

Soft decision:

$$\sum_{i=1}^{K} \delta_i = 1 \quad , \quad \delta_i \geq 0 \quad \forall \; i$$



Classification in Pattern S

3. Learning methods

Supervised Training	Unsupervised Training
→ we have got preclassified trainingsdata	→ we have data, which is not preclassified
→ we know about the number of classes	
→ we want to find the position of the classes in the feature space	→ we want to find clusters in the feature space

Classification in Pattern S

3. Learning methods

3.1. Maximum likelihood

We have n training – examples:

$X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$

We have got a model function for the distribution:

$P(\boldsymbol{x} \mid \phi)$ $\phi -$ parameter vector

$$\max_{\phi} P(\phi \mid \boldsymbol{X}) = \max_{\phi} \sum_{i=1}^{N} \log P(\boldsymbol{x_i} \mid \phi) + \log P(\phi) \quad \rightarrow \quad \hat{\phi} = \begin{bmatrix} \hat{\mu} \\ \hat{\sigma} \end{bmatrix}$$

$$P(\boldsymbol{X} \mid \phi) = \prod_{i=1}^{N} P(\boldsymbol{x_i} \mid \phi)$$

advantage : easy, just differentiate L(ϕ) !
disadvantage : does not work most of the times !

Classification in Pattern S

3. Learning methods

3.2. EM algorithm (expectation maximization)

We observe data $X = [x_1, x_2, \ldots, x_N]$, which is <u>incomplete</u> !

eg. first step of the statistical model for speech creation

$\rightarrow$ we don't know the classes Ω_i for each vector x_i

$X \rightarrow$ observed $U \rightarrow$ unknown data $P(x \mid \phi) \rightarrow$ model

$Z = [X,U] \rightarrow$ complete data

$$\max_{\phi} \log P(\boldsymbol{Z} \mid \phi) \qquad \rightarrow P(Z \mid \phi) = P(U \mid X, \phi)\, P(X \mid \phi)$$

$$\max_{\phi} E_U\left[\log P(\boldsymbol{Z} \mid \hat{\phi}) \mid X, \phi\right]$$

random variable !