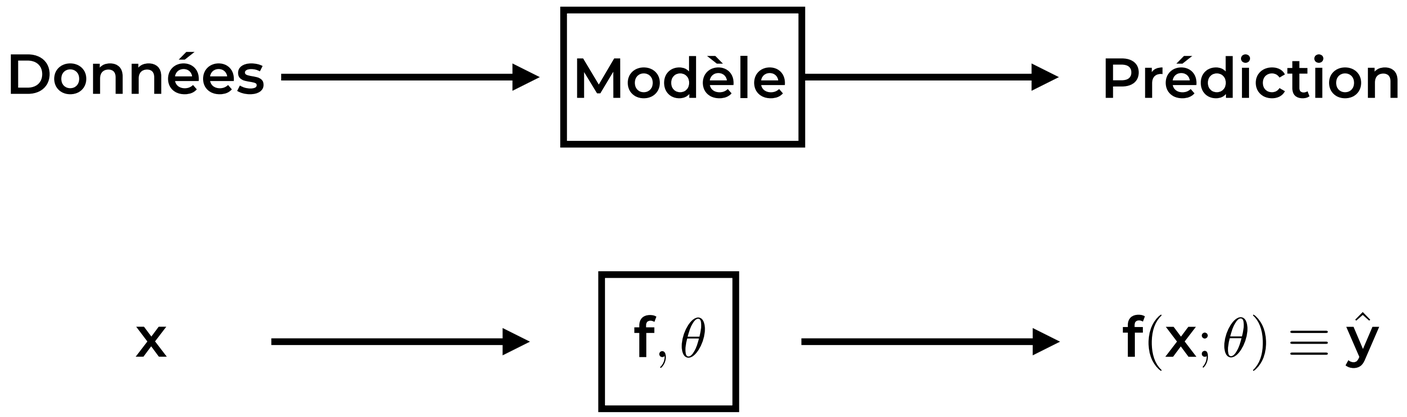
# Apprentissage Automatique I MATH60629

Sommaire jusqu’à maintenant

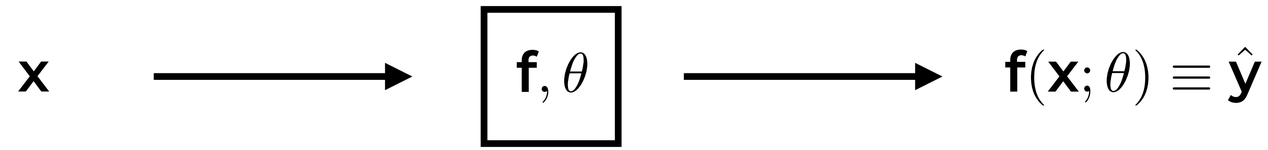
- **Bref survol de certains des concepts vus jusqu’à maintenant**

- **Emphase sur les concepts un peu plus avancés**

# Apprentissage Supervisé



# Fonction de perte (loss)

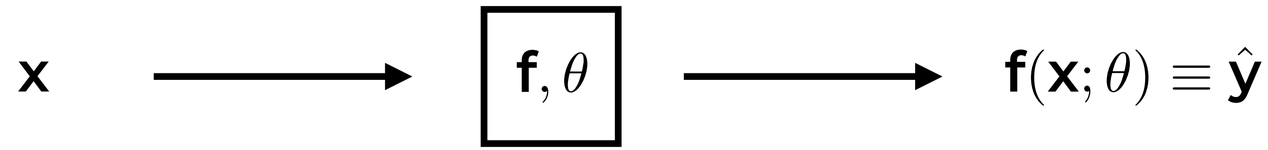



**Différentes pertes pour selon le type de y**

| | | |
|---|---|---|
| $y \in \mathcal{R}$ | Régression | $(\hat{y} - y)^2$ |
| y catégorique e.g., $\{chat, chien, oiseau\}$ | Classification | taux de classification |
| $y \in \{0, 1\}$ | Classification binaire | AUC |

# Fonction de perte (loss)

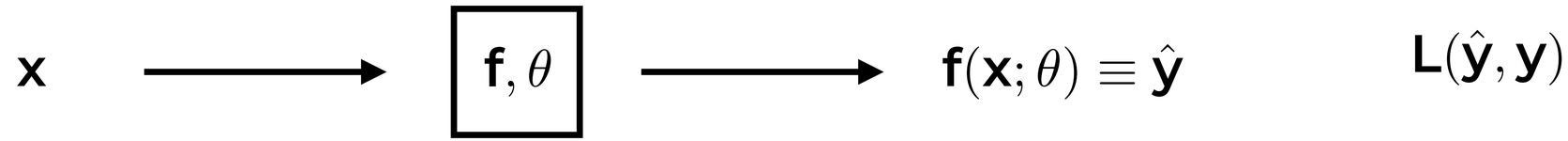
Perte




**Différentes pertes pour selon le type de y**

| | | |
|---|---|---|
| $y \in \mathcal{R}$ | Régression | $(\hat{y} - y)^2$ |
| y catégorique e.g., $\{chat, chien, oiseau\}$ | Classification | taux de classification |
| $y \in \{0, 1\}$ | Classification binaire | AUC |

# Fonction de perte (loss)

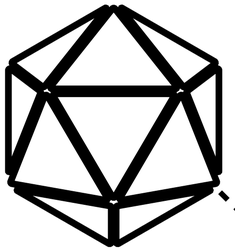



**Différentes pertes pour selon le type de y**

| | | |
|---|---|---|
| $y \in \mathcal{R}$ | Régression | $(\hat{y} - y)^2$ |
| y catégorique e.g., $\{\text{chat}, \text{chien}, \text{oiseau}\}$ | Classification | taux de classification |
| $y \in \{0, 1\}$ | Classification binaire | AUC |

# Processus d'apprentissage

Distribution sur (x,y): P(x,y)



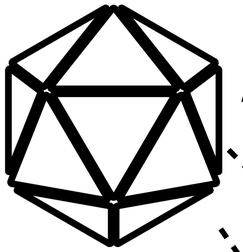
$x_{train}$ → $f, \theta$ → $\hat{y}_{train}$ → Perte $L(\hat{y}_{train}, y_{train})$

$x_{test}$ → $f, \hat{\theta}$ → $\hat{y}_{test}$ → $L(\hat{y}_{test}, y_{test})$

# Processus d'apprentissage

## En pratique

Distribution sur (x,y): P(x,y)



$x_{train}$ → $f, \theta$ → $\hat{y}_{train}$ → Perte $L(\hat{y}_{train}, y_{train})$

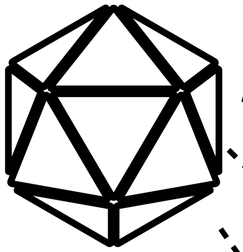
$x_{valid}$ → $f, \hat{\theta}$ → $\hat{y}_{valid}$

$x_{test}$ → $f, \hat{\theta}$ → $\hat{y}_{test}$ → Perte $L(\hat{y}_{test}, y_{test})$

# Processus d'apprentissage

## En pratique

Distribution sur (x,y): P(x,y)



$x_{train}$ → $f, \theta$ → $\hat{y}_{train}$

$x_{valid}$ → $f, \hat{\theta}$ → $\hat{y}_{valid}$

$x_{test}$ → $f, \hat{\theta}$ → $\hat{y}_{test}$

Perte

$L(\hat{y}_{train}, y_{train})$

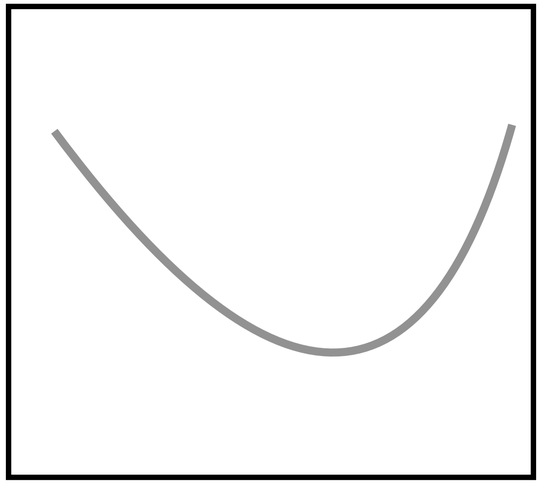
Pratique pour:
- Sélectionner les hyperparamètres
- Pour sélectionner le meilleur modèle

$L(\hat{y}_{test}, y_{test})$

# Apprentissage

- Apprendre: Changer les paramètres pour obtenir de meilleures prédictions

Loss



- En d'autres mots: changer les paramètres $\theta$ pour minimiser la perte
  - Trouver les dérivées de la perte p.r. aux paramètres: $\frac{\mathrm{d\ Loss}}{\mathrm{d}\theta}$

# Différents modèles

- f: régression linéaire, $\theta$ a une solution analytique
- f: réseaux de neurones, $\theta$ n'a pas de solution analytique. On utilise la descente de gradient

- Given a training set: $\{(\mathbf{x}_{\text{train}}, \mathbf{y}_{\text{train}})\}$

- Initialize $\hat{\theta}$ randomly

**for** t $= 1, 2, \ldots$ (epochs) **do**
  **for** i $= 1, 2, \ldots$ (datum) **do**

    - Obtain the prediction $\{\mathbf{f}(\mathbf{x}_\mathbf{i}; \hat{\theta})\}$ (Forward propagation)

    - Compute the Loss: $\mathbf{Loss}_{\mathbf{ti}} := \mathbf{L}(\mathbf{f}(\mathbf{x}_\mathbf{i}; \hat{\theta}), \mathbf{y}_\mathbf{i})$

    - Find the derivative of the loss: $\frac{\mathbf{d}\ \mathbf{Loss}_{\mathbf{ti}}}{\mathbf{d}\ \hat{\theta}}$
    - Get new parameters: $\hat{\theta}' = \hat{\theta} - \alpha \frac{\mathbf{d}\ \mathbf{Loss}_{\mathbf{ti}}}{\mathbf{d}\ \hat{\theta}}$
    - If $||\hat{\theta}' - \hat{\theta}||_2^2 < \epsilon$ then stop
    - Update parameters: $\hat{\theta} = \hat{\theta}'$
  **end for**
**end for**

- Given a training set: $\{(x_{train}, y_{train})\}$

- Initialize $\hat{\theta}$ randomly

Stochastic Gradient Descent

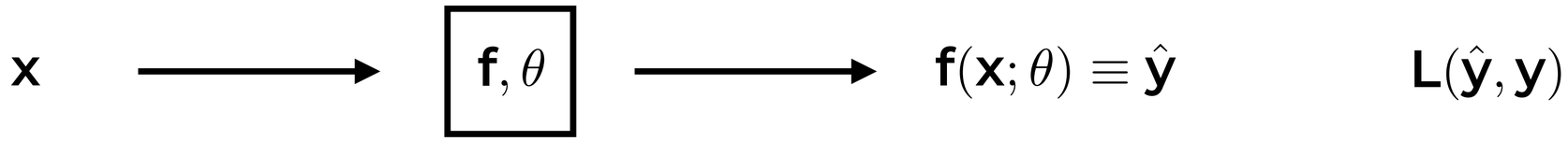
**for** t $= 1, 2, \ldots$ (epochs) **do**
  **for** i $= 1, 2, \ldots$ (datum) **do**

    - Obtain the prediction $\{f(x_i; \hat{\theta})\}$ (Forward propagation)

    - Compute the Loss: $Loss_{ti} := L(f(x_i; \hat{\theta}), y_i)$

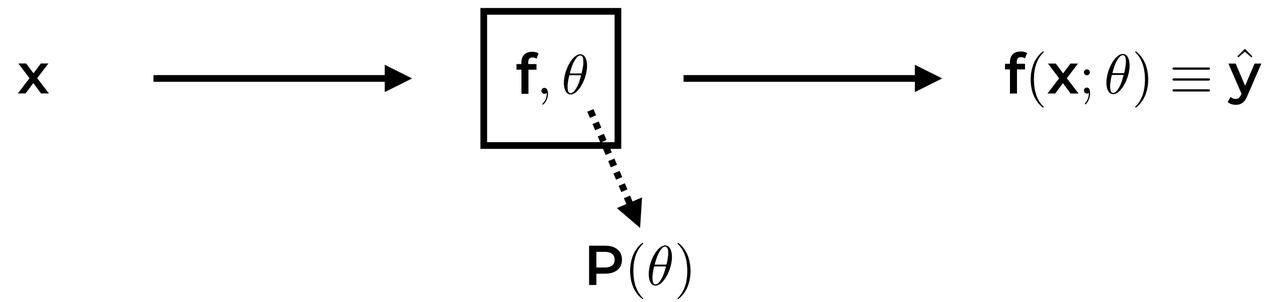
    - Find the derivative of the loss: $\frac{d\ Loss_{ti}}{d\ \hat{\theta}}$
    - Get new parameters: $\hat{\theta}' = \hat{\theta} - \alpha \frac{d\ Loss_{ti}}{d\ \hat{\theta}}$
    - If $||\hat{\theta}' - \hat{\theta}||_2^2 < \epsilon$ then stop
    - Update parameters: $\hat{\theta} = \hat{\theta}'$
  **end for**
**end for**

# Modèles probabilistes




**Likelihood**
$\mathbf{P}(\mathbf{data} \mid \mathbf{params})$
$\mathbf{P}(\mathbf{x}, \mathbf{y} \mid \theta)$

# Modèles probabilistes




**Perte**

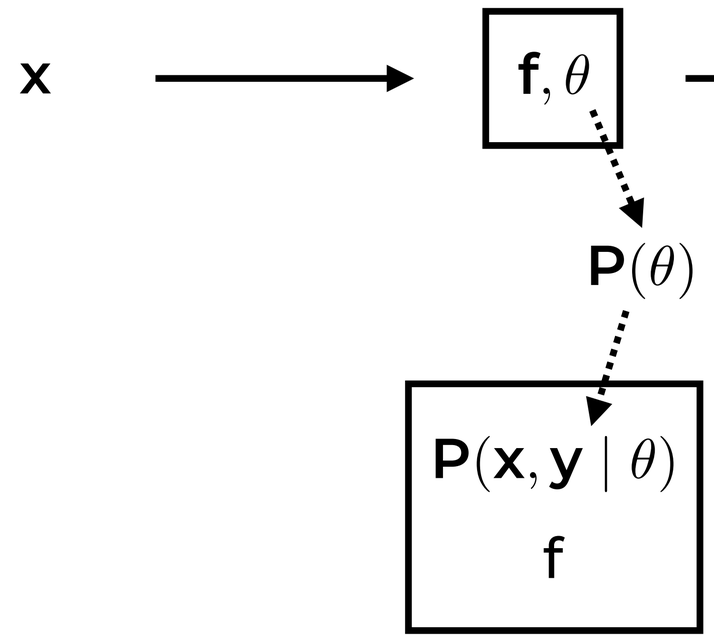
$\mathbf{L}(\hat{\mathbf{y}}, \mathbf{y})$

**Likelihood**

$\mathbf{P}(\mathbf{data} \mid \mathbf{params})$

$\mathbf{P}(\mathbf{x}, \mathbf{y} \mid \theta)$

# Modèles probabilistes



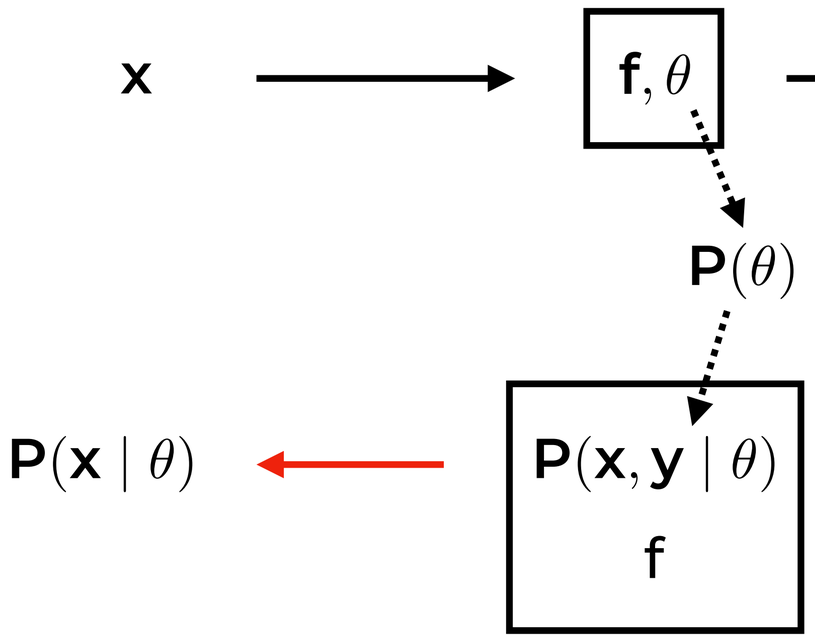

**Perte**

$\mathbf{L}(\hat{\mathbf{y}}, \mathbf{y})$

$\mathbf{f}(\mathbf{x}; \theta) \equiv \hat{\mathbf{y}}$

**Likelihood**
$\mathbf{P}(\text{data} \mid \text{params})$
$\mathbf{P}(\mathbf{x}, \mathbf{y} \mid \theta)$

# Modèles probabilistes



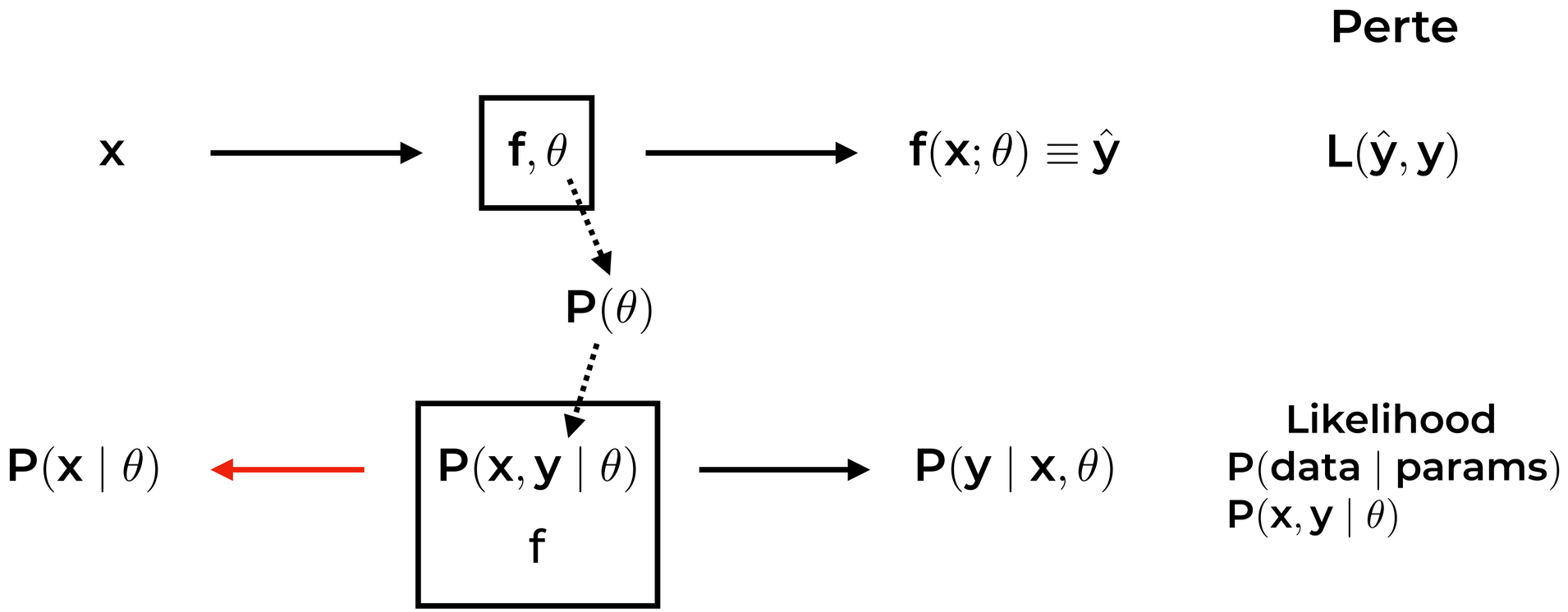

**Perte**

$\mathbf{L}(\hat{\mathbf{y}}, \mathbf{y})$

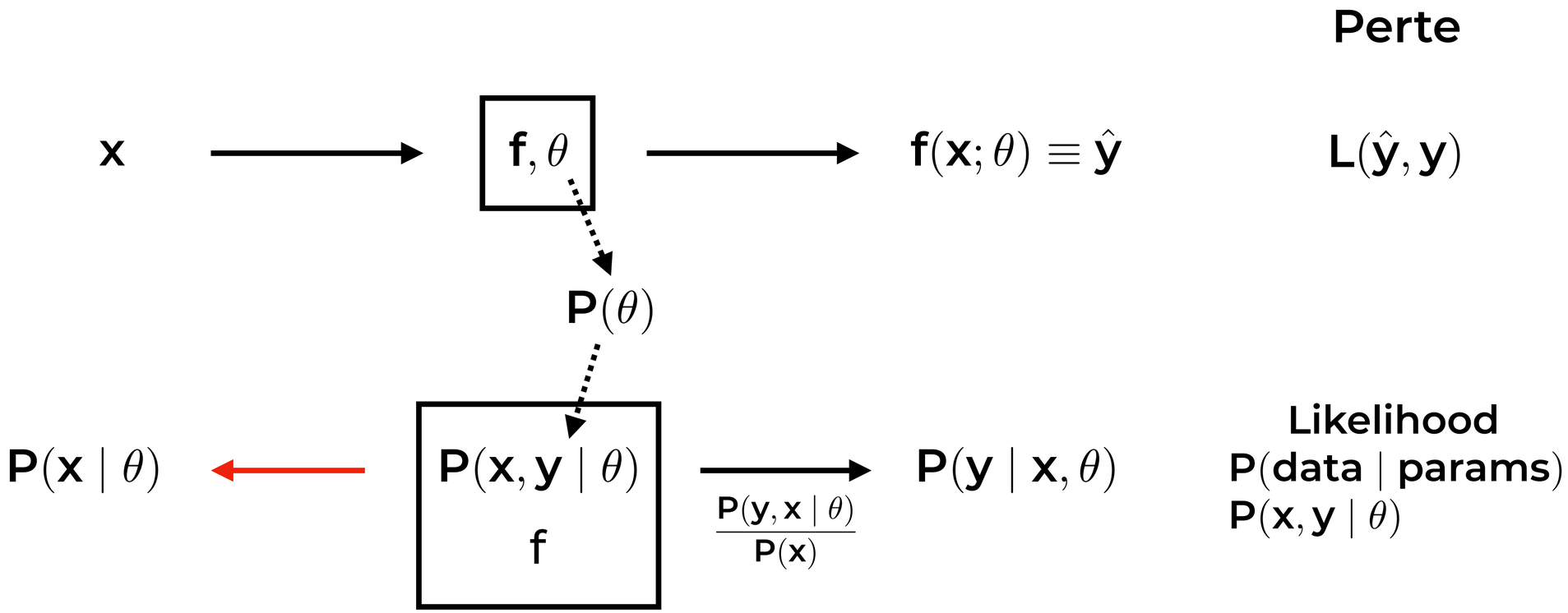
$\mathbf{f}(\mathbf{x}; \theta) \equiv \hat{\mathbf{y}}$

**Likelihood**
$\mathbf{P}(\mathbf{data} \mid \mathbf{params})$
$\mathbf{P}(\mathbf{x}, \mathbf{y} \mid \theta)$

# Modèles probabilistes



# Modèles probabilistes



# Exemple

**Données: 952 1064 965 1037 871 1029 1138 (problème non-supervisé)**

**Modèle:** $\mathbf{P}(\mathbf{x} \mid \theta) := \mathcal{N}(\boldsymbol{\mu}, \mathbf{1})$

# Exemple

**Données: 952 1064 965 1037 871 1029 1138 (problème non-supervisé)**

**Modèle:** $\mathbf{P}(\mathbf{x} \mid \theta) := \mathcal{N}(\boldsymbol{\mu}, \mathbf{1})$

# Exemple

**Données: 952 1064 965 1037 871 1029 1138 (problème non-supervisé)**

**Modèle:** $\mathbf{P}(\mathbf{x} \mid \theta) := \mathcal{N}(\boldsymbol{\mu}, \mathbf{1})$

Likelihood d'une donnée:

# Exemple

**Données: 952 1064 965 1037 871 1029 1138 (problème non-supervisé)**

**Modèle:** $\mathbf{P}(\mathbf{x} \mid \theta) := \mathcal{N}(\boldsymbol{\mu}, \mathbf{1})$

Likelihood d'une donnée:

$$\textbf{Likelihood}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{1}) = \frac{\mathbf{1}}{\sqrt{\mathbf{2\pi}}} \exp - \frac{(\mathbf{x} - \boldsymbol{\mu})^{\mathbf{2}}}{\mathbf{2}}$$

# Exemple

**Données: 952 1064 965 1037 871 1029 1138 (problème non-supervisé)**

**Modèle:** $\mathbf{P}(\mathbf{x} \mid \theta) := \mathcal{N}(\boldsymbol{\mu}, \mathbf{1})$

Likelihood d'une donnée:

$$\textbf{Likelihood}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{1}) = \frac{\mathbf{1}}{\sqrt{\mathbf{2\pi}}} \exp - \frac{(\mathbf{x} - \boldsymbol{\mu})^2}{\mathbf{2}}$$

**Log-Likelihood**

$$= \log \frac{\mathbf{1}}{\sqrt{\mathbf{2\pi}}} \exp - \frac{(\mathbf{x} - \boldsymbol{\mu})^2}{\mathbf{2}}$$

$$= \mathbf{log1} - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{log2\pi} - \frac{(\mathbf{x} - \boldsymbol{\mu})^2}{\mathbf{2}}$$

# Exemple

**Données:** 952 1064 965 1037 871 1029 1138 (problème non-supervisé)

**Modèle:** $\mathbf{P}(\mathbf{x} \mid \theta) := \mathcal{N}(\boldsymbol{\mu}, \mathbf{1})$

Likelihood d'une donnée:

$$\textbf{Likelihood}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{1}) = \frac{\mathbf{1}}{\sqrt{\mathbf{2\pi}}} \exp -\frac{(\mathbf{x} - \boldsymbol{\mu})^2}{\mathbf{2}}$$

**Log-Likelihood**

$$= \log \frac{\mathbf{1}}{\sqrt{\mathbf{2\pi}}} \exp -\frac{(\mathbf{x} - \boldsymbol{\mu})^2}{\mathbf{2}}$$

$$= \mathbf{log1} - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{log2\pi} - \frac{(\mathbf{x} - \boldsymbol{\mu})^2}{\mathbf{2}}$$

Quelle valeur de $\boldsymbol{\mu}$ la maximise?

# Exemple

**Données:** 952 1064 965 1037 871 1029 1138 (problème non-supervisé)

**Modèle:** $\mathbf{P}(\mathbf{x} \mid \theta) := \mathcal{N}(\boldsymbol{\mu}, \mathbf{1})$

Likelihood d'une donnée:

$$\textbf{Likelihood}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{1}) = \frac{\mathbf{1}}{\sqrt{\mathbf{2\pi}}} \exp -\frac{(\mathbf{x}-\boldsymbol{\mu})^{\mathbf{2}}}{\mathbf{2}}$$

**Log-Likelihood**

$$= \log \frac{\mathbf{1}}{\sqrt{\mathbf{2\pi}}} \exp -\frac{(\mathbf{x}-\boldsymbol{\mu})^{\mathbf{2}}}{\mathbf{2}}$$

$$= \mathbf{log1} - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{log2\pi} - \frac{(\mathbf{x}-\boldsymbol{\mu})^{\mathbf{2}}}{\mathbf{2}}$$

Quelle valeur de $\boldsymbol{\mu}$ la maximise?

$$\frac{\mathbf{d\ Log\text{-}Likelihood}}{\mathbf{d}\ \boldsymbol{\mu}}$$

$$= \frac{\mathbf{d}\ \frac{(\mathbf{x}-\boldsymbol{\mu})^{\mathbf{2}}}{\mathbf{2}}}{\mathbf{d}\ \boldsymbol{\mu}}$$

$$= (\mathbf{x}-\boldsymbol{\mu})$$

**set to 0**

$$\boldsymbol{\mu} = \mathbf{x}$$

# Exemple

**Données:** 952 1064 965 1037 871 1029 1138 (problème non-supervisé)

**Modèle:** $\mathbf{P}(\mathbf{x} \mid \theta) := \mathcal{N}(\boldsymbol{\mu}, \mathbf{1})$

Likelihood d'une donnée:

$$\textbf{Likelihood}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{1}) = \frac{\mathbf{1}}{\sqrt{\mathbf{2\pi}}} \exp -\frac{(\mathbf{x}-\boldsymbol{\mu})^2}{\mathbf{2}}$$

**Log-Likelihood**

$$= \log \frac{\mathbf{1}}{\sqrt{\mathbf{2\pi}}} \exp -\frac{(\mathbf{x}-\boldsymbol{\mu})^2}{\mathbf{2}}$$

$$= \mathbf{log1} - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{log2\pi} - \frac{(\mathbf{x}-\boldsymbol{\mu})^2}{\mathbf{2}}$$

Quelle valeur de $\boldsymbol{\mu}$ la maximise?

$$\frac{\mathbf{d}\ \textbf{Log-Likelihood}}{\mathbf{d}\ \boldsymbol{\mu}}$$

$$= \frac{\mathbf{d}\ \frac{(\mathbf{x}-\boldsymbol{\mu})^2}{\mathbf{2}}}{\mathbf{d}\ \boldsymbol{\mu}}$$

$$= (\mathbf{x}-\boldsymbol{\mu})$$

**set to 0**

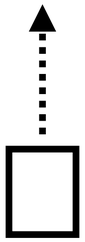
$\boldsymbol{\mu} = \mathbf{x}$ = 952

# Données: (x,y)
# Naive Bayes

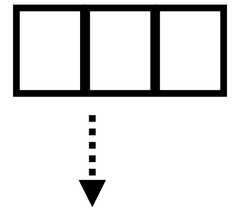
**Modele** : $\mathbf{P}(\mathbf{x}, \mathbf{y} \mid \theta) = \mathbf{P}(\mathbf{x} \mid \mathbf{y}, \theta)\mathbf{P}(\mathbf{y} \mid \theta)$

# Données: (x,y)
# Naive Bayes

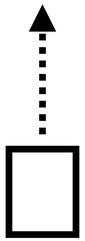
**Modele** : $\mathbf{P}(\mathbf{x}, \mathbf{y} \mid \theta) = \mathbf{P}(\mathbf{x} \mid \mathbf{y}, \theta)\mathbf{P}(\mathbf{y} \mid \theta)$



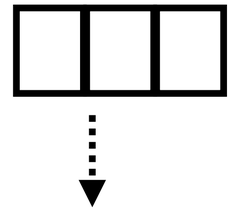
# Données: (x,y)
# Naive Bayes

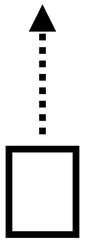


Modele : $\mathbf{P}(\mathbf{x}, \mathbf{y} \mid \theta) = \mathbf{P}(\mathbf{x} \mid \mathbf{y}, \theta)\mathbf{P}(\mathbf{y} \mid \theta)$



## Données: (x,y)
## Naive Bayes



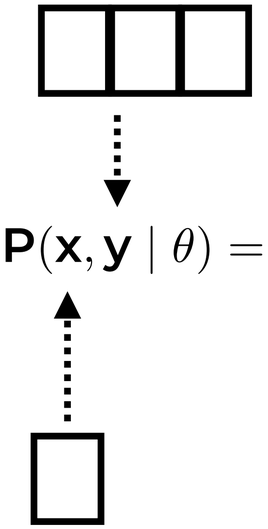


Modele : $\mathbf{P}(\mathbf{x}, \mathbf{y} \mid \theta) = \mathbf{P}(\mathbf{x} \mid \mathbf{y}, \theta)\mathbf{P}(\mathbf{y} \mid \theta)$



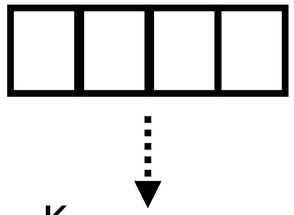
## Données: x
## Gaussian Mixture Models

# Données: (x,y)
# Naive Bayes



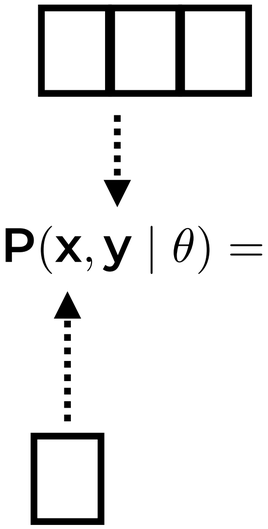
Modele : $\mathbf{P}(\mathbf{x}, \mathbf{y} \mid \theta) = \mathbf{P}(\mathbf{x} \mid \mathbf{y}, \theta)\mathbf{P}(\mathbf{y} \mid \theta)$

# Données: x
# Gaussian Mixture Models

K=4 

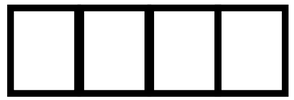
Modele : $\mathbf{P}(\mathbf{x} \mid \theta) = \sum_{\mathbf{k}=1}^{\mathbf{K}} \mathbf{P}(\theta_{\mathbf{k}}) \underbrace{\mathbf{P}(\mathbf{x} \mid \theta_{\mathbf{k}})}_{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{k}}, \Sigma_{\mathbf{k}})}$ **(K groupes)**

## Données: (x,y)
## Naive Bayes



Modele : $\mathbf{P}(\mathbf{x}, \mathbf{y} \mid \theta) = \mathbf{P}(\mathbf{x} \mid \mathbf{y}, \theta)\mathbf{P}(\mathbf{y} \mid \theta)$

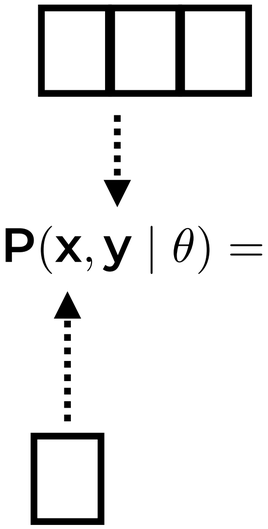
## Données: x
## Gaussian Mixture Models

K=4 

Modele : $\mathbf{P}(\mathbf{x} \mid \theta) = \sum_{\mathbf{k}=1}^{\mathbf{K}} \mathbf{P}(\theta_{\mathbf{k}}) \underbrace{\mathbf{P}(\mathbf{x} \mid \theta_{\mathbf{k}})}_{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{k}}, \Sigma_{\mathbf{k}})}$ **(K groupes)**

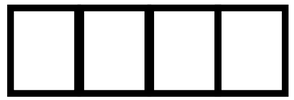
**Max. likelihood (MLE)** : $\hat{\theta}_{\mathbf{MLE}} = \arg\max_{\theta} \mathbf{P}(\mathbf{x} \mid \theta)$

## Données: (x,y)
## Naive Bayes



Modele : $\mathbf{P}(\mathbf{x}, \mathbf{y} \mid \theta) = \mathbf{P}(\mathbf{x} \mid \mathbf{y}, \theta)\mathbf{P}(\mathbf{y} \mid \theta)$

## Données: x
## Gaussian Mixture Models

K=4 

Modele : $\mathbf{P}(\mathbf{x} \mid \theta) = \sum_{\mathbf{k}=1}^{\mathbf{K}} \mathbf{P}(\theta_{\mathbf{k}}) \underbrace{\mathbf{P}(\mathbf{x} \mid \theta_{\mathbf{k}})}_{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{k}}, \Sigma_{\mathbf{k}})}$ **(K groupes)**

**Max. likelihood (MLE)** : $\hat{\theta}_{\mathbf{MLE}} = \arg\max_{\theta} \mathbf{P}(\mathbf{x} \mid \theta)$

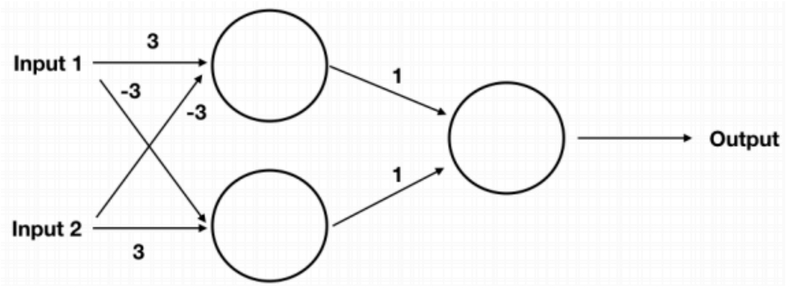
**Max. a posteriori (MAP)** : $\hat{\theta}_{\mathbf{MAP}} = \arg\max_{\theta} \mathbf{P}(\theta \mid \mathbf{x}, \gamma) = \arg\max_{\theta} \dfrac{\mathbf{P}(\mathbf{x} \mid \theta)\mathbf{P}(\theta \mid \gamma)}{\mathbf{P}(\mathbf{x})}$

# MLPs / RNNs / CNNs

- MLPs: les couches sont complètement connectées à la prochaine couche
- RNNs: entrées à chaque couche
  - Application standard: modélisation des séries temporelles
- CNNs: remplacent les multiplications de matrices par des convolutions (sparse connections, weight sharing) + et du pooling
  - Application standard: reconnaissance d’objets dans des images

# MLPs

(e) (4 points) Consider the neural network below. We have estimated its parameters (shown next to their corresponding arrows).




The activation function of each unit in the network is a simple thresholding function:

$$\text{threshold}(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 & \text{if } x > 0. \end{cases} \tag{1}$$

For each of these four sets of inputs write down the network's output (i.e., its prediction) in the "Output" column of the table below.

| Input 1 | Input 2 | Output |
|---|---|---|
| 0 | 0 | |
| 1 | 1 | |
| 0 | 1 | |
| 1 | 0 | |