

# A2DI: Apprentissage statistique

John Klein

Université de Lille - CRIStAL UMR CNRS 9189



Où en est-on dans notre problème d'apprentissage supervisé ?

- En théorie on veut minimiser  $Err_{esp}$ .
- En pratique on minimisera  $Err_{train}$  tout en s'assurant que  $Err_{train}$  ne dévie pas de  $Err_{test}$ .
- Les solutions minimisant  $Err_{train}$  au sens de la théorie de la décision pour des fonctions de perte classiques tournent autour de solutions probabilistes reposant sur la connaissance de la distribution  $p_{Y|X}$

Où en est-on dans notre problème d'apprentissage supervisé ?

- En théorie on veut minimiser  $Err_{esp}$ .
- En pratique on minimisera  $Err_{train}$  tout en s'assurant que  $Err_{train}$  ne dévie pas de  $Err_{test}$ .
- Les solutions minimisant  $Err_{train}$  au sens de la théorie de la décision pour des fonctions de perte classiques tournent autour de solutions probabilistes reposant sur la connaissance de la distribution  $p_{Y|X}$

Où en est-on dans notre problème d'apprentissage supervisé ?

- En théorie on veut minimiser  $Err_{esp}$ .
- En pratique on minimisera  $Err_{train}$  tout en s'assurant que  $Err_{train}$  ne dévie pas de  $Err_{test}$ .
- Les solutions minimisant  $Err_{train}$  au sens de la théorie de la décision pour des fonctions de perte classiques tournent autour de solutions probabilistes reposant sur la connaissance de la distribution  $p_{Y|X}$

Où en est-on dans notre problème d'apprentissage supervisé ?

- En théorie on veut minimiser  $Err_{esp}$ .
- En pratique on minimisera  $Err_{train}$  tout en s'assurant que  $Err_{train}$  ne dévie pas de  $Err_{test}$ .
- Les solutions minimisant  $Err_{train}$  au sens de la théorie de la décision pour des fonctions de perte classiques tournent autour de solutions probabilistes reposant sur la connaissance de la distribution  $p_{Y|X}$

Pour tendre vers ces solutions optimales<sup>1</sup>, il faut trouver un moyen de calculer une estimation  $\hat{p}_{Y|X}$  grâce à nos données  $\mathcal{D}$ . Dans ce chapitre, nous allons étudier différentes méthodes statistiques paramétriques pour atteindre ce but.

---

1. L'optimalité au sens de la fonction de perte choisie.

# Plan du chapitre

- 1 Modèles paramétriques : généralités
- 2 Modèles paramétriques : fit fréquentiste
- 3 Modèles paramétriques : fit Bayésien
- 4 Conclusions

Qu'est ce qu'un **modèle paramétrique** ?

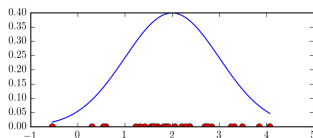
- C'est le **choix** une classe de distributions paramétrées pour  $p_{X,Y}(x,y) = g_{\theta}(x,y)$ .
- C'est aussi le **choix** d'un espace d'hypothèse  $\mathcal{H}$  (ensemble des prédicteurs parmi lesquels on doit choisir).
- En effet, le meilleur prédicteur est étroitement liée à  $p_{X,Y}$

Qu'est ce qu'un **modèle paramétrique** ? Exemples

v.a. continue

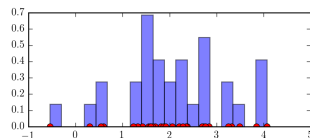
Paramétrique

loi Gaussienne, Beta, etc.



Non-Paramétrique

Histogramme



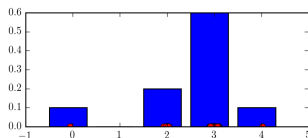


Qu'est ce qu'un **modèle paramétrique** ? Exemples

v.a. discrète

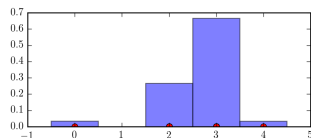
Paramétrique

loi multinomiale, catégorique



Non-Paramétrique

Histogramme

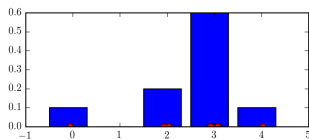


# Qu'est ce qu'un modèle paramétrique ? Exemples

v.a. discrète

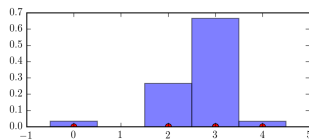
Paramétrique

loi multinomiale, catégorique



Non-Paramétrique

Histogramme



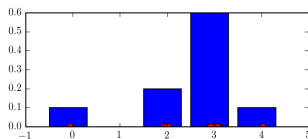
Est-ce vraiment différent ?

# Qu'est ce qu'un modèle paramétrique ? Exemples

v.a. discrète

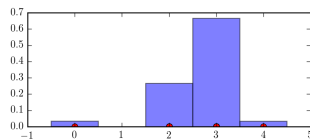
## Paramétrique

loi multinomiale, catégorique



## Non-Paramétrique

## Histogramme



Est-ce vraiment différent ?

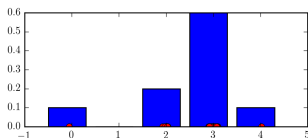
→ pas vraiment, car la distribution catégorique correspond à toutes les distributions discrètes possibles !

# Qu'est ce qu'un **modèle paramétrique** ? Exemples

v.a. discrète

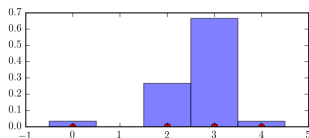
Paramétrique

loi multinomiale, catégorique



Non-Paramétrique

Histogramme



Est-ce **vraiment différent** ?

- pas vraiment, car la distribution catégorique correspond à toutes les distributions discrètes possibles !
- Aussi possible en continu, si le modèle paramétrique est très général

## Modèles paramétriques : catégories

## Modèles Génératifs

On cherche  $p_{X,Y}$

convergence plus rapide vers  
quelque chose de pertinent

## Modèles Discriminatifs

On cherche  $p_{Y|X}$

légèrement meilleur quand  
 $n$  est grand

# Plan du chapitre

- 1 Modèles paramétriques : généralités
- 2 Modèles paramétriques : fit fréquentiste**
- 3 Modèles paramétriques : fit Bayésien
- 4 Conclusions

Comment **trouver** un modèle qui colle à mes données ?

Approche **fréquentiste**



Comment **trouver** un modèle qui colle à mes données ?

Approche **fréquentiste**

**Rappel** : la vraie valeur de  $\theta$  est fixe et inconnue.





Comment **trouver** un modèle qui colle à mes données ?

Approche **fréquentiste**

**Rappel** : la vraie valeur de  $\theta$  est fixe et inconnue.

- Je suppose que mes données sont **i.i.d.** et  $(\mathbf{x}^{(i)}, y^{(i)})$  vient d'une v.a.  
 $(\mathbf{X}^{(i)}, Y^{(i)}) \sim f_{\theta}$ .

Comment **trouver** un modèle qui colle à mes données ?

Approche **fréquentiste**

**Rappel** : la vraie valeur de  $\theta$  est fixe et inconnue.

- Je suppose que mes données sont **i.i.d.** et  $(\mathbf{x}^{(i)}, y^{(i)})$  vient d'une v.a.  
 $(\mathbf{X}^{(i)}, Y^{(i)}) \sim f_{\theta}$ .
- La probabilité d'observer  $\mathbf{x}^{(i)}$  et  $y^{(i)}$  est donc  $f_{\theta}(\mathbf{x}^{(i)}, y^{(i)})$ .

Comment **trouver** un modèle qui colle à mes données ?

Approche **fréquentiste**

**Rappel** : la vraie valeur de  $\theta$  est fixe et inconnue.

- Je suppose que mes données sont **i.i.d.** et  $(\mathbf{x}^{(i)}, y^{(i)})$  vient d'une v.a.  
 $(\mathbf{X}^{(i)}, Y^{(i)}) \sim f_{\theta}$ .
- La probabilité d'observer  $\mathbf{x}^{(i)}$  et  $y^{(i)}$  est donc  $f_{\theta}(\mathbf{x}^{(i)}, y^{(i)})$ .
- La probabilité d'observer  $\mathcal{D}$  est donc

$$\prod_{i=1}^n f_{\theta}(\mathbf{x}^{(i)}, y^{(i)}) \quad (1)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **fréquentiste**

**Rappel** : la vraie valeur de  $\theta$  est fixe et inconnue.

- Je suppose que mes données sont **i.i.d.** et  $(\mathbf{x}^{(i)}, y^{(i)})$  vient d'une v.a.  
 $(\mathbf{X}^{(i)}, Y^{(i)}) \sim f_{\theta}$ .
- La probabilité d'observer  $\mathbf{x}^{(i)}$  et  $y^{(i)}$  est donc  $f_{\theta}(\mathbf{x}^{(i)}, y^{(i)})$ .
- La probabilité d'observer  $\mathcal{D}$  est donc

$$\prod_{i=1}^n f_{\theta}(\mathbf{x}^{(i)}, y^{(i)}) \quad (1)$$

→ C'est la fonction de **vraisemblance**  $\mathcal{L}(\theta)$ .

Comment **trouver** un modèle qui colle à mes données ?

Approche **fréquentiste**

- J'ai un bon modèle = J'ai un bon  $\theta$ ,

Comment **trouver** un modèle qui colle à mes données ?

Approche **fréquentiste**

- J'ai un bon modèle = J'ai un bon  $\theta$ ,
- J'ai un bon modèle = La proba d'observer  $\mathcal{D}$  sous ce modèle est grande,

Comment **trouver** un modèle qui colle à mes données ?

Approche **fréquentiste**

- J'ai un bon modèle = J'ai un bon  $\theta$ ,
- J'ai un bon modèle = La proba d'observer  $\mathcal{D}$  sous ce modèle est grande,
- Je choisis donc  $\theta^* = \arg \max_{\theta \in \Theta} \mathcal{L}(\theta)$  !

Comment **trouver** un modèle qui colle à mes données ?

Approche **fréquentiste**

- J'ai un bon modèle = J'ai un bon  $\theta$ ,
- J'ai un bon modèle = La proba d'observer  $\mathcal{D}$  sous ce modèle est grande,
- Je choisis donc  $\theta^* = \arg \max_{\theta \in \Theta} \mathcal{L}(\theta)$  !

Le vecteur  $\theta^*$  est appelée *maximum likelihood estimate*, souvent noté  $\hat{\theta}_{MLE}$ .



Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple de **régression linéaire**

- Je choisis un modèle **discriminatif** Gaussien :

$$p_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - (\mathbf{w}^T \cdot \mathbf{x} + b))^2}{2\sigma^2}}. \quad (2)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple de **régression linéaire**

- Je choisis un modèle **discriminatif** Gaussien :

$$p_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - (\mathbf{w}^T \cdot \mathbf{x} + b))^2}{2\sigma^2}}. \quad (2)$$

- J'ai un  $\theta =$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple de **régression linéaire**

- Je choisis un modèle **discriminatif** Gaussien :

$$p_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - (\mathbf{w}^T \cdot \mathbf{x} + b))^2}{2\sigma^2}}. \quad (2)$$

- J'ai un  $\theta =$
- J'ai donc la vraisemblance suivante :

$$\mathcal{L}(\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)} - (\mathbf{w}^T \cdot \mathbf{x}^{(i)} + b))^2}{2\sigma^2}} \quad (3)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple de **régression linéaire**

- Maximiser  $\mathcal{L}(\theta)$  = Minimiser  $\text{NLL}(\theta)$ .
- En choisissant une **fonction de perte quadratique**, montrez que la NLL s'écrit :

$$\text{NLL}(\theta) = n \log(\sigma) + \frac{n}{2\sigma^2} \text{Err}_{\text{train}} + \text{cte} \quad (4)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple de **régression linéaire**

- Minimiser  $\text{NLL}(\theta)$  est facile car c'est une fonction **convexe**.

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple de **régression linéaire**

- Minimiser  $NLL(\theta)$  est facile car c'est une fonction **convexe**.
- $\hat{\theta}_{MLE}$  est donc l'unique minimum global de  $NLL(\theta)$ ,

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple de **régression linéaire**

- Minimiser  $\text{NLL}(\theta)$  est facile car c'est une fonction **convexe**.
- $\hat{\theta}_{MLE}$  est donc l'unique minimum global de  $\text{NLL}(\theta)$ ,
- ... ou encore l'unique point tel que  $\frac{d}{d\theta}\text{NLL}(\theta) = 0$ .

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple de **régression linéaire**

- Minimiser  $\text{NLL}(\boldsymbol{\theta})$  est facile car c'est une fonction **convexe**.
- $\hat{\boldsymbol{\theta}}_{MLE}$  est donc l'unique minimum global de  $\text{NLL}(\boldsymbol{\theta})$ ,
- ... ou encore l'unique point tel que  $\frac{d}{d\boldsymbol{\theta}}\text{NLL}(\boldsymbol{\theta}) = 0$ .
- Concernant les paramètres  $\mathbf{w}$  et  $b$ , on déjà vu en TP que

$$\begin{pmatrix} w_1 \\ \vdots \\ w_d \\ b \end{pmatrix} = (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{X}\mathbf{y} \quad (5)$$



Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple de **régression linéaire**

- Minimiser  $\text{NLL}(\boldsymbol{\theta})$  est facile car c'est une fonction **convexe**.
- $\hat{\boldsymbol{\theta}}_{MLE}$  est donc l'unique minimum global de  $\text{NLL}(\boldsymbol{\theta})$ ,
- ... ou encore l'unique point tel que  $\frac{d}{d\boldsymbol{\theta}}\text{NLL}(\boldsymbol{\theta}) = 0$ .
- Concernant les paramètres  $\mathbf{w}$  et  $b$ , on déjà vu en TP que

$$\begin{pmatrix} w_1 \\ \vdots \\ w_d \\ b \end{pmatrix} = (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{X}\mathbf{y} \quad (5)$$

avec

$$\mathbf{X} = \begin{pmatrix} \begin{bmatrix} \mathbf{x}^{(1)} \\ 1 \end{bmatrix} & \dots & \begin{bmatrix} \mathbf{x}^{(n)} \\ 1 \end{bmatrix} \end{pmatrix} \quad \text{et} \quad \mathbf{y} = \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{pmatrix} \quad (6)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple de **régression linéaire**

- Concernant le paramètre  $\sigma$ , montrez que son estimé **MLE** est l'écart type empirique :

$$\hat{\sigma}_{MLE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y^{(i)} - \mathbf{w}^T \cdot \mathbf{x}^{(i)} - b)^2}. \quad (7)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple de **régression linéaire**

- $\hat{\sigma}_{MLE}$  ne sert à rien pour la prédiction qui est  $\mathbb{E}[Y|X] = \hat{\mathbf{w}}^T \cdot \mathbf{x}^{(i)} + \hat{b}$
- $\hat{\sigma}_{MLE}$  sert à quantifier notre confiance en la prédiction  $\mathbb{E}[Y|X]$ .

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Je choisis un modèle **génératif** et **factorisable** comme suit :

$$p_{\mathbf{X}, Y}(\mathbf{x}, c) = p_Y(c) p_{\mathbf{X}|Y=c}(\mathbf{x}), \quad (8)$$

$$= p_Y(c) \prod_{j=1}^d p_{X_j|Y=c}(x_j). \quad (9)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Je choisis un modèle **génératif** et **factorisable** comme suit :

$$p_{\mathbf{X},Y}(\mathbf{x},c) = p_Y(c) p_{\mathbf{X}|Y=c}(\mathbf{x}), \quad (8)$$

$$= p_Y(c) \prod_{j=1}^d p_{X_j|Y=c}(x_j). \quad (9)$$

- Quelle hypothèse d'**indépendance** est nécessaire pour écrire cette factorisation ?

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Je choisis un modèle **génératif** et **factorisable** comme suit :

$$p_{\mathbf{X}, Y}(\mathbf{x}, c) = p_Y(c) p_{\mathbf{X}|Y=c}(\mathbf{x}), \quad (8)$$

$$= p_Y(c) \prod_{j=1}^d p_{X_j|Y=c}(x_j). \quad (9)$$

- Quelle hypothèse d'**indépendance** est nécessaire pour écrire cette factorisation ?
- $p_{\mathbf{X}|Y=c}(\mathbf{x})$  est appelée *class conditional density*.

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Je choisis un modèle **génératif** et **factorisable** comme suit :

$$p_{\mathbf{X}, Y}(\mathbf{x}, c) = p_Y(c) p_{\mathbf{X}|Y=c}(\mathbf{x}), \quad (8)$$

$$= p_Y(c) \prod_{j=1}^d p_{X_j|Y=c}(x_j). \quad (9)$$

- Quelle hypothèse d'**indépendance** est nécessaire pour écrire cette factorisation ?
- $p_{\mathbf{X}|Y=c}(\mathbf{x})$  est appelée **class conditional density**.
- Supposons que  $X_j|Y \sim \mathcal{N}(\mu_{j,c}, \sigma_{j,c})$  et  $Y \sim \text{Cat}(\boldsymbol{\pi})$ . En notant  $\ell = \#\mathcal{C}$ , on a :

$$\boldsymbol{\theta} = \quad (10)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Etant donné mon choix de distribution, l'équation (9) s'écrit pour couple  $(\mathbf{x}^{(i)}, c^{(i)})$

$$p_{\mathbf{X}, Y}(\mathbf{x}^{(i)}, c^{(i)}) = \pi_{c^{(i)}} \prod_{j=1}^d \frac{1}{\sqrt{2\pi}\sigma_{j,c^{(i)}}} e^{-\frac{(x_j^{(i)} - \mu_{j,c^{(i)}})^2}{2\sigma_{j,c^{(i)}}^2}}. \quad (11)$$



Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Etant donné mon choix de distribution, l'équation (9) s'écrit pour couple  $(\mathbf{x}^{(i)}, c^{(i)})$

$$p_{\mathbf{X}, Y}(\mathbf{x}^{(i)}, c^{(i)}) = \pi_{c^{(i)}} \prod_{j=1}^d \frac{1}{\sqrt{2\pi}\sigma_{j,c^{(i)}}} e^{-\frac{(x_j^{(i)} - \mu_{j,c^{(i)}})^2}{2\sigma_{j,c^{(i)}}^2}}. \quad (11)$$

- La **vraisemblance** s'écrit donc :

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^n \pi_{c^{(i)}} \prod_{j=1}^d \frac{1}{\sqrt{2\pi}\sigma_{j,c^{(i)}}} e^{-\frac{(x_j^{(i)} - \mu_{j,c^{(i)}})^2}{2\sigma_{j,c^{(i)}}^2}}. \quad (12)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Passons à la NLL :

$$\text{NLL}(\boldsymbol{\theta}) = \sum_{i=1}^n \log \pi_{c(i)} + \sum_{j=1}^d -\log \left( \sqrt{2\pi} \sigma_{j,c(i)} \right) - \frac{\left( x_j^{(i)} - \mu_{j,c(i)} \right)^2}{2\sigma_{j,c(i)}^2}.$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Passons à la NLL :

$$\text{NLL}(\boldsymbol{\theta}) = \sum_{i=1}^n \log \pi_{c^{(i)}} + \sum_{j=1}^d -\log \left( \sqrt{2\pi} \sigma_{j,c^{(i)}} \right) - \frac{\left( x_j^{(i)} - \mu_{j,c^{(i)}} \right)^2}{2\sigma_{j,c^{(i)}}^2}.$$

- Cette NLL est une fonction **convexe** de  **$\boldsymbol{\theta}$**

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Passons à la NLL :

$$\text{NLL}(\boldsymbol{\theta}) = \sum_{i=1}^n \log \pi_{c(i)} + \sum_{j=1}^d -\log \left( \sqrt{2\pi} \sigma_{j,c(i)} \right) - \frac{\left( x_j^{(i)} - \mu_{j,c(i)} \right)^2}{2\sigma_{j,c(i)}^2}.$$

- Cette NLL est une fonction **convexe** de  $\boldsymbol{\theta}$
- Intéressons nous d'abord aux  $\pi_{c_k}$ . On doit résoudre :

$$\frac{\partial}{\partial \pi_{c_k}} \text{NLL}(\boldsymbol{\theta}) = 0,$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Passons à la NLL :

$$\text{NLL}(\boldsymbol{\theta}) = \sum_{i=1}^n \log \pi_{c^{(i)}} + \sum_{j=1}^d -\log \left( \sqrt{2\pi} \sigma_{j,c^{(i)}} \right) - \frac{\left( x_j^{(i)} - \mu_{j,c^{(i)}} \right)^2}{2\sigma_{j,c^{(i)}}^2}.$$

- Cette NLL est une fonction **convexe** de  $\boldsymbol{\theta}$
- Intéressons nous d'abord aux  $\pi_{c_k}$ . On doit résoudre :

$$\frac{\partial}{\partial \pi_{c_k}} \text{NLL}(\boldsymbol{\theta}) = 0,$$

sous la **contrainte**  $\sum_{k=1}^{\ell} \pi_{c_k} = 1$ .

Comment **trouver** un modèle qui colle à mes données ?



**Pause Optim !**

- Comment minimiser une fonction objectif sous une **contrainte** d'égalité ?
- Astuce du **Lagrangien** :
  - On rajoute (temporairement) un nouveau paramètre  $\lambda$ .

Comment **trouver** un modèle qui colle à mes données ?



**Pause Optim !**

- Comment minimiser une fonction objectif sous une **contrainte** d'égalité ?
- Astuce du **Lagrangien** :
  - On rajoute (temporairement) un nouveau paramètre  $\lambda$ .
  - On modifie la fonction objectif en y incorporant la **contrainte** d'égalité, par exemple :

$$J(\boldsymbol{\theta}, \lambda) = \text{NLL}(\boldsymbol{\theta}) + \lambda \left( 1 - \sum_{k=1}^{\ell} \pi_{c_k} \right).$$

Comment **trouver** un modèle qui colle à mes données ?



**Pause Optim !**

- Comment minimiser une fonction objectif sous une **contrainte** d'égalité ?
- Astuce du **Lagrangien** :
  - On rajoute (temporairement) un nouveau paramètre  $\lambda$ .
  - On modifie la fonction objectif en y incorporant la **contrainte** d'égalité, par exemple :

$$J(\boldsymbol{\theta}, \lambda) = \text{NLL}(\boldsymbol{\theta}) + \lambda \left( 1 - \sum_{k=1}^{\ell} \pi_{c_k} \right).$$

- La fonction objectif modifiée  $J$  reste **convexe** et on a :

$$\frac{\partial}{\partial \lambda} J(\boldsymbol{\theta}, \lambda) = 0 \quad \Leftrightarrow \quad \sum_{k=1}^{\ell} \pi_{c_k} = 1.$$



Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Nous cherchons donc à présent :

$$\frac{\partial}{\partial \pi_{c_k}} J(\boldsymbol{\theta}) = 0, \quad (13)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Nous cherchons donc à présent :

$$\frac{\partial}{\partial \pi_{c_k}} J(\boldsymbol{\theta}) = 0,$$

$$\Leftrightarrow \frac{\partial}{\partial \pi_{c_k}} \text{NLL}(\boldsymbol{\theta}) - \lambda = 0, \quad (13)$$

$$(14)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Nous cherchons donc à présent :

$$\frac{\partial}{\partial \pi_{c_k}} J(\boldsymbol{\theta}) = 0,$$

$$\Leftrightarrow \frac{\partial}{\partial \pi_{c_k}} \text{NLL}(\boldsymbol{\theta}) - \lambda = 0, \quad (13)$$

$$\Leftrightarrow \frac{\partial}{\partial \pi_{c_k}} \sum_{i=1}^n \log \pi_{c(i)} = \lambda, \quad (14)$$

$$(15)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Nous cherchons donc à présent :

$$\frac{\partial}{\partial \pi_{c_k}} J(\boldsymbol{\theta}) = 0,$$

$$\Leftrightarrow \frac{\partial}{\partial \pi_{c_k}} \text{NLL}(\boldsymbol{\theta}) - \lambda = 0, \quad (13)$$

$$\Leftrightarrow \frac{\partial}{\partial \pi_{c_k}} \sum_{i=1}^n \log \pi_{c(i)} = \lambda, \quad (14)$$

$$\Leftrightarrow n_k \times \frac{\partial}{\partial \pi_{c_k}} \log \pi_{c_k} = \lambda, \quad (15)$$

$$(16)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Nous cherchons donc à présent :

$$\frac{\partial}{\partial \pi_{c_k}} J(\boldsymbol{\theta}) = 0,$$

$$\Leftrightarrow \frac{\partial}{\partial \pi_{c_k}} \text{NLL}(\boldsymbol{\theta}) - \lambda = 0, \quad (13)$$

$$\Leftrightarrow \frac{\partial}{\partial \pi_{c_k}} \sum_{i=1}^n \log \pi_{c(i)} = \lambda, \quad (14)$$

$$\Leftrightarrow n_k \times \frac{\partial}{\partial \pi_{c_k}} \log \pi_{c_k} = \lambda, \quad (15)$$

$$\Leftrightarrow \frac{n_k}{\pi_{c_k}} = \lambda. \quad (16)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Cette relation est vraie pour tout  $k$ , d'où

$$\sum_{k=1}^{\ell} n_k = \sum_{k=1}^{\ell} \lambda \pi_{c_k}, \quad (17)$$

(18)

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Cette relation est vraie pour tout  $k$ , d'où

$$\sum_{k=1}^{\ell} n_k = \sum_{k=1}^{\ell} \lambda \pi_{c_k}, \quad (17)$$

$$\Leftrightarrow n = \lambda \sum_{k=1}^{\ell} \pi_{c_k}, \quad (18)$$

$$(19)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Cette relation est vraie pour tout  $k$ , d'où

$$\sum_{k=1}^{\ell} n_k = \sum_{k=1}^{\ell} \lambda \pi_{c_k}, \quad (17)$$

$$\Leftrightarrow n = \lambda \sum_{k=1}^{\ell} \pi_{c_k}, \quad (18)$$

$$\Leftrightarrow n = \lambda. \quad (19)$$



Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Cette relation est vraie pour tout  $k$ , d'où

$$\sum_{k=1}^{\ell} n_k = \sum_{k=1}^{\ell} \lambda \pi_{c_k}, \quad (17)$$

$$\Leftrightarrow n = \lambda \sum_{k=1}^{\ell} \pi_{c_k}, \quad (18)$$

$$\Leftrightarrow n = \lambda. \quad (19)$$

- Au final, il vient :

$$\hat{\pi}_{c_k, MLE} = \frac{n_k}{n}. \quad (20)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Passons à présent aux paramètres suivants :  $\mu_{j,c_k}$

$$\frac{\partial}{\partial \mu_{j,c_k}} \text{NLL}(\boldsymbol{\theta}) = 0, \quad (21)$$

$$(22)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Passons à présent aux paramètres suivants :  $\mu_{j,c_k}$

$$\frac{\partial}{\partial \mu_{j,c_k}} \text{NLL}(\boldsymbol{\theta}) = 0, \quad (21)$$

$$\Leftrightarrow \sum_{i=1}^n \sum_{j=1}^d -\frac{1}{2\sigma_{j,c(i)}^2} \frac{\partial}{\partial \mu_{j,c_k}} \left( x_j^{(i)} - \mu_{j,c(i)} \right)^2 = 0, \quad (22)$$

$$(23)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Passons à présent aux paramètres suivants :  $\mu_{j,c_k}$

$$\frac{\partial}{\partial \mu_{j,c_k}} \text{NLL}(\boldsymbol{\theta}) = 0, \quad (21)$$

$$\Leftrightarrow \sum_{i=1}^n \sum_{j=1}^d -\frac{1}{2\sigma_{j,c^{(i)}}^2} \frac{\partial}{\partial \mu_{j,c_k}} \left( x_j^{(i)} - \mu_{j,c^{(i)}} \right)^2 = 0, \quad (22)$$

$$\Leftrightarrow \sum_{r=1}^{n_k} \sum_{j=1}^d -\frac{1}{2\sigma_{j,c_k}^2} \frac{\partial}{\partial \mu_{j,c_k}} \left( x_j^{(r)} - \mu_{j,c_k} \right)^2 = 0, \quad (23)$$

$$(24)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Passons à présent aux paramètres suivants :  $\mu_{j,c_k}$

$$\frac{\partial}{\partial \mu_{j,c_k}} \text{NLL}(\boldsymbol{\theta}) = 0, \quad (21)$$

$$\Leftrightarrow \sum_{i=1}^n \sum_{j=1}^d -\frac{1}{2\sigma_{j,c^{(i)}}^2} \frac{\partial}{\partial \mu_{j,c_k}} \left( x_j^{(i)} - \mu_{j,c^{(i)}} \right)^2 = 0, \quad (22)$$

$$\Leftrightarrow \sum_{r=1}^{n_k} \sum_{j=1}^d -\frac{1}{2\sigma_{j,c_k}^2} \frac{\partial}{\partial \mu_{j,c_k}} \left( x_j^{(r)} - \mu_{j,c_k} \right)^2 = 0, \quad (23)$$

$$\Leftrightarrow -\frac{1}{2\sigma_{j,c_k}^2} \sum_{r=1}^{n_k} \frac{\partial}{\partial \mu_{j,c_k}} \left( x_j^{(r)} - \mu_{j,c_k} \right)^2 = 0. \quad (24)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Calcul des  $\mu_{j,c_k}$  (suite) :

$$\Leftrightarrow -\frac{1}{2\sigma_{j,c_k}^2} \sum_{r=1}^{n_k} 2 \times -1 \times \left( x_j^{(r)} - \mu_{j,c_k} \right) = 0, \quad (25)$$

(26)

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Calcul des  $\mu_{j,c_k}$  (suite) :

$$\Leftrightarrow -\frac{1}{2\sigma_{j,c_k}^2} \sum_{r=1}^{n_k} 2 \times -1 \times \left( x_j^{(r)} - \mu_{j,c_k} \right) = 0, \quad (25)$$

$$\Leftrightarrow \sum_{r=1}^{n_k} \left( x_j^{(r)} - \mu_{j,c_k} \right) = 0, \quad (26)$$

$$(27)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Calcul des  $\mu_{j,c_k}$  (suite) :

$$\Leftrightarrow -\frac{1}{2\sigma_{j,c_k}^2} \sum_{r=1}^{n_k} 2 \times -1 \times \left( x_j^{(r)} - \mu_{j,c_k} \right) = 0, \quad (25)$$

$$\Leftrightarrow \sum_{r=1}^{n_k} \left( x_j^{(r)} - \mu_{j,c_k} \right) = 0, \quad (26)$$

$$\Leftrightarrow \mu_{j,c_k} = \frac{1}{n_k} \sum_{r=1}^{n_k} x_j^{(r)} \quad (27)$$



Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Calcul des  $\mu_{j,c_k}$  (suite) :

$$\Leftrightarrow -\frac{1}{2\sigma_{j,c_k}^2} \sum_{r=1}^{n_k} 2 \times -1 \times (x_j^{(r)} - \mu_{j,c_k}) = 0, \quad (25)$$

$$\Leftrightarrow \sum_{r=1}^{n_k} (x_j^{(r)} - \mu_{j,c_k}) = 0, \quad (26)$$

$$\Leftrightarrow \mu_{j,c_k} = \frac{1}{n_k} \sum_{r=1}^{n_k} x_j^{(r)} \quad (27)$$

En conclusion,  $\hat{\mu}_{j,c_k,MLE}$  est la moyenne empirique des exemples de la classe  $c_k$  pour la dimension  $j$ .

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Il ne reste plus que les paramètres suivants :  $\sigma_{j,c_k}$
- On montrerait de même que  $\hat{\sigma}_{j,c_k,MLE}^2$  est la variance empirique des exemples de la classe  $c_k$  pour la dimension  $j$  :

$$\hat{\sigma}_{j,c_k,MLE}^2 = \frac{1}{n_k} \sum_{r=1}^{n_k} \left( x_j^{(r)} - \hat{\mu}_{j,c_k,MLE} \right)^2 \quad (28)$$

- → Bon exercice à faire chez soi ..

Comment **trouver** un modèle qui colle à mes données ?

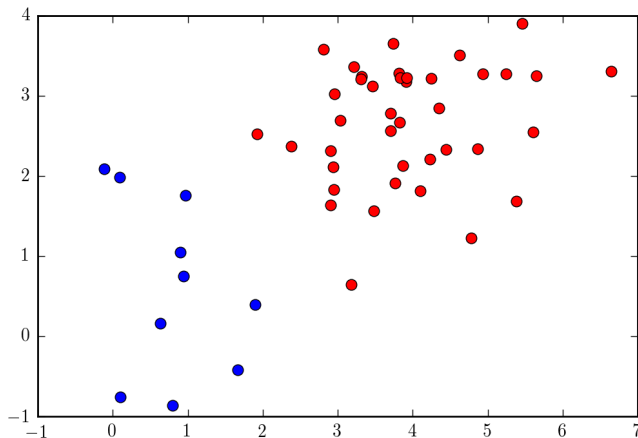
Approche **MLE** : exemple du **classifieur naïf bayésien**

- Le **classifieur naïf bayésien** est un des modèles **génératifs** les plus utilisés.
- Il est simple à **fitter** et présente peu de risque d'**overfitting** du fait des hypothèses simplificatrices limitant le nombre de paramètres.

Comment **trouver** un modèle qui colle à mes données ?

Approche **MLE** : exemple du **classifieur naïf bayésien**

- Résumé des opérations :



# Plan du chapitre

- 1 Modèles paramétriques : généralités
- 2 Modèles paramétriques : fit fréquentiste
- 3 Modèles paramétriques : fit Bayésien**
- 4 Conclusions

Comment **trouver** un modèle qui colle à mes données ?

Approche **Bayésienne**



Comment **trouver** un modèle qui colle à mes données ?

Approche **Bayésienne**

**Rappel** : le vecteur  $\theta$  est **aléatoire** et j'ai une connaissance *a priori* concernant ses valeurs probables.



Comment **trouver** un modèle qui colle à mes données ?

Approche **Bayésienne**

**Rappel** : le vecteur  $\theta$  est **aléatoire** et j'ai une connaissance *a priori* concernant ses valeurs probables.

- Concernant mes données, je fais les mêmes hypothèses que dans le cas **fréquentiste** : données **i.i.d.** et  $(\mathbf{x}^{(i)}, y^{(i)})$  vient d'une v.a.  
 $(\mathbf{X}^{(i)}, Y^{(i)}) \sim f_{\theta}$ .



Comment **trouver** un modèle qui colle à mes données ?

Approche **Bayésienne**

**Rappel** : le vecteur  $\theta$  est **aléatoire** et j'ai une connaissance **a priori** concernant ses valeurs probables.

- Concernant mes données, je fais les mêmes hypothèses que dans le cas **fréquentiste** : données **i.i.d.** et  $(\mathbf{x}^{(i)}, y^{(i)})$  vient d'une v.a.  
 $(\mathbf{X}^{(i)}, Y^{(i)}) \sim f_{\theta}$ .
- C'est encore la fonction de **vraisemblance**  $\mathcal{L}(\theta)$  qui résume l'information venant de mes données.

Comment **trouver** un modèle qui colle à mes données ?

Approche **Bayésienne**

**Rappel** : le vecteur  $\theta$  est **aléatoire** et j'ai une connaissance **a priori** concernant ses valeurs probables.

- Concernant mes données, je fais les mêmes hypothèses que dans le cas **fréquentiste** : données **i.i.d.** et  $(\mathbf{x}^{(i)}, y^{(i)})$  vient d'une v.a.  
 $(\mathbf{X}^{(i)}, Y^{(i)}) \sim f_{\theta}$ .
- C'est encore la fonction de **vraisemblance**  $\mathcal{L}(\theta)$  qui résume l'information venant de mes données.
- Je dispose en plus d'un **prior**  $p_{\theta}$ .
- Je cherche à présent la **posterior** sur  $\theta$  :

$$p_{\theta|\text{données}} = \frac{p_{\text{données}|\theta} \times p_{\theta}}{p_{\text{données}}}. \quad (29)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **Bayésienne**

**Rappel** : le vecteur  $\theta$  est **aléatoire** et j'ai une connaissance **a priori** concernant ses valeurs probable.

- Je cherche à présent la **posterior** sur  $\theta$  :

$$p_{\theta|\text{données}} = \frac{p_{\text{données}|\theta} \times p_{\theta}}{p_{\text{données}}}, \quad (30)$$

(31)

Comment **trouver** un modèle qui colle à mes données ?

Approche **Bayésienne**

**Rappel** : le vecteur  $\theta$  est **aléatoire** et j'ai une connaissance **a priori** concernant ses valeurs probable.

- Je cherche à présent la **posterior** sur  $\theta$  :

$$p_{\theta|\text{données}} = \frac{p_{\text{données}|\theta} \times p_{\theta}}{p_{\text{données}}}, \quad (30)$$

$$p_{\theta|\mathcal{D}} = \frac{\mathcal{L}(\theta) \times p_{\theta}}{p_{\mathcal{D}}}, \quad (31)$$

$$(32)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **Bayésienne**

**Rappel** : le vecteur  $\theta$  est **aléatoire** et j'ai une connaissance **a priori** concernant ses valeurs probable.

- Je cherche à présent la **posterior** sur  $\theta$  :

$$p_{\theta|\text{données}} = \frac{p_{\text{données}|\theta} \times p_{\theta}}{p_{\text{données}}}, \quad (30)$$

$$p_{\theta|\mathcal{D}} = \frac{\mathcal{L}(\theta) \times p_{\theta}}{p_{\mathcal{D}}}, \quad (31)$$

$$p_{\theta|\mathcal{D}} = \frac{\mathcal{L}(\theta) \times p_{\theta}}{\int \mathcal{L}(\theta) \times p_{\theta} d\theta} \quad (32)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **Bayésienne** : même procédure que dans le cas fréquentiste

- J'ai un bon modèle = J'ai un bon  $\theta$ ,

Comment **trouver** un modèle qui colle à mes données ?

Approche **Bayésienne** : même procédure que dans le cas fréquentiste

- J'ai un bon modèle = J'ai un bon  $\theta$ ,
- J'ai un bon modèle = La proba de  $\theta$  ayant observé  $\mathcal{D}$  sous ce modèle est grande,

Comment **trouver** un modèle qui colle à mes données ?

Approche **Bayésienne** : même procédure que dans le cas fréquentiste

- J'ai un bon modèle = J'ai un bon  $\theta$ ,
- J'ai un bon modèle = La proba de  $\theta$  ayant observé  $\mathcal{D}$  sous ce modèle est grande,
- Je choisis donc  $\theta^* = \arg \max_{\theta \in \Theta} p_{\theta|\mathcal{D}} !$



Comment **trouver** un modèle qui colle à mes données ?

Approche **Bayésienne** : même procédure que dans le cas fréquentiste

- J'ai un bon modèle = J'ai un bon  $\theta$ ,
- J'ai un bon modèle = La proba de  $\theta$  ayant observé  $\mathcal{D}$  sous ce modèle est grande,
- Je choisis donc  $\theta^* = \arg \max_{\theta \in \Theta} p_{\theta|\mathcal{D}} !$

Le vecteur  $\theta^*$  est appelée *maximum a posteriori estimate*, souvent noté  $\hat{\theta}_{MAP}$ .

Comment **trouver** un modèle qui colle à mes données ?

Approche **Bayésienne**

- Si mon seul but est de maximiser la **posterior**, je peux me contenter d'étudier

$$\tilde{p}_{\theta|\mathcal{D}} = \mathcal{L}(\theta) \times p_{\theta} \propto p_{\theta|\mathcal{D}} \quad (33)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **Bayésienne**

- Si mon seul but est de maximiser la **posterior**, je peux me contenter d'étudier

$$\tilde{p}_{\theta|\mathcal{D}} = \mathcal{L}(\theta) \times p_{\theta} \propto p_{\theta|\mathcal{D}} \quad (33)$$

- $\tilde{p}_{\theta|\mathcal{D}}$  n'est pas une probabilité car elle n'est pas normalisée.

Comment **trouver** un modèle qui colle à mes données ?

Approche **Bayésienne**

- Si mon seul but est de maximiser la **posterior**, je peux me contenter d'étudier

$$\tilde{p}_{\theta|\mathcal{D}} = \mathcal{L}(\theta) \times p_{\theta} \propto p_{\theta|\mathcal{D}} \quad (33)$$

- $\tilde{p}_{\theta|\mathcal{D}}$  n'est pas une probabilité car elle n'est pas normalisée.
- Si j'ai besoin d'avoir de savoir à quel point je peux avoir confiance en  $\hat{\theta}_{MAP}$ , il faut nécessairement avoir la distribution normalisée.

Comment **trouver** un modèle qui colle à mes données ?

Approche **MAP** : exemple de **régression linéaire** le retour..

- Je choisis un modèle **discriminatif** Gaussien :

$$p_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - (\mathbf{w}^T \cdot \mathbf{x} + b))^2}{2\sigma^2}}. \quad (34)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MAP** : exemple de **régression linéaire** le retour..

- Je choisis un modèle **discriminatif** Gaussien :

$$p_{Y|X=\mathbf{x}}(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - (\mathbf{w}^T \cdot \mathbf{x} + b))^2}{2\sigma^2}}. \quad (34)$$

- J'ai (toujours) un  $\boldsymbol{\theta} = [\mathbf{w} \ b \ \sigma]^T$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MAP** : exemple de **régression linéaire** le retour..

- Je choisis un modèle **discriminatif** Gaussien :

$$p_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - (\mathbf{w}^T \cdot \mathbf{x} + b))^2}{2\sigma^2}}. \quad (34)$$

- J'ai (toujours) un  $\boldsymbol{\theta} = [\mathbf{w} \ b \ \sigma]^T$
- J'ai la vraisemblance suivante :

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)} - (\mathbf{w}^T \cdot \mathbf{x}^{(i)} + b))^2}{2\sigma^2}} \quad (35)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MAP** : exemple de **régression linéaire** le retour..

- Supposons à présent que j'ai un **prior** gaussien sur uniquement sur les paramètres **w** et **b**.

$$p_{\mathbf{w},b}(\mathbf{w}, b) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{\begin{bmatrix} \mathbf{w} & b \end{bmatrix} \cdot \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}}{2\sigma_0^2}}. \quad (36)$$



Comment **trouver** un modèle qui colle à mes données ?

Approche **MAP** : exemple de **régression linéaire** le retour..

- Supposons à présent que j'ai un **prior** gaussien sur uniquement sur les paramètres **w** et **b**.

$$p_{\mathbf{w},b}(\mathbf{w}, b) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{\begin{bmatrix} \mathbf{w} & b \end{bmatrix} \cdot \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}}{2\sigma_0^2}}. \quad (36)$$

- Cela signifie qu'au départ, je considère que des valeurs proches de 0 pour **w** et **b** sont plus probables.

Comment **trouver** un modèle qui colle à mes données ?

Approche **MAP** : exemple de **régression linéaire** le retour..

- Supposons à présent que j'ai un **prior** gaussien sur uniquement sur les paramètres **w** et **b**.

$$p_{\mathbf{w},b}(\mathbf{w}, b) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{\begin{bmatrix} \mathbf{w} & b \end{bmatrix} \cdot \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}}{2\sigma_0^2}}. \quad (36)$$

- Cela signifie qu'au départ, je considère que des valeurs proches de 0 pour **w** et **b** sont plus probables.
- Concernant le paramètre du bruit  $\sigma$ , je vais me contenter du MLE.

Comment **trouver** un modèle qui colle à mes données ?

Approche **MAP** : exemple de **régression linéaire** le retour..

- Supposons à présent que j'ai un **prior** gaussien sur uniquement sur les paramètres **w** et **b**.

$$p_{\mathbf{w},b}(\mathbf{w}, b) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{\begin{bmatrix} \mathbf{w} & b \end{bmatrix} \cdot \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}}{2\sigma_0^2}}. \quad (36)$$

- Cela signifie qu'au départ, je considère que des valeurs proches de 0 pour **w** et **b** sont plus probables.
- Concernant le paramètre du bruit  $\sigma$ , je vais me contenter du MLE.
- On a donc la **posterior** non normalisée suivante :

$$\tilde{p}_{\theta|\mathcal{D}}(\mathbf{w}, b, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)} - (\mathbf{w}^T \cdot \mathbf{x}^{(i)} + b))^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{\begin{bmatrix} \mathbf{w} & b \end{bmatrix} \cdot \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}}{2\sigma_0^2}} \quad (37)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MAP** : exemple de **régression linéaire** le retour..

- Si on utilise l'opération permettant de passer à la NLL, on obtient la fonction objectif  $J$  suivante :

$$J(\mathbf{w}, b, \sigma) = \text{NLL}(\mathbf{w}, b, \sigma) - \log(\sqrt{2\pi}\sigma_0) - \frac{\begin{bmatrix} \mathbf{w} & b \end{bmatrix} \cdot \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}}{2\sigma_0^2}. \quad (38)$$

Comment **trouver** un modèle qui colle à mes données ?

Approche **MAP** : exemple de **régression linéaire** le retour..

- Si on utilise l'opération permettant de passer à la NLL, on obtient la fonction objectif  $J$  suivante :

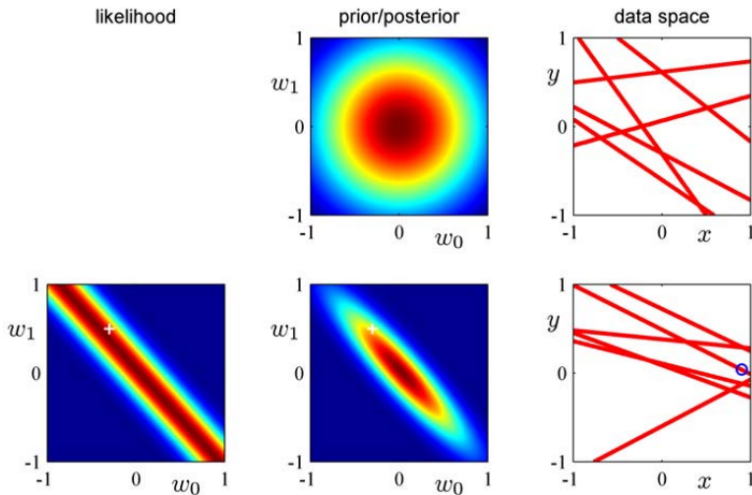
$$J(\mathbf{w}, b, \sigma) = \text{NLL}(\mathbf{w}, b, \sigma) - \log(\sqrt{2\pi}\sigma_0) - \frac{\begin{bmatrix} \mathbf{w} & b \end{bmatrix} \cdot \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}}{2\sigma_0^2}. \quad (38)$$

- Le terme  $\log(\sqrt{2\pi}\sigma_0)$  est constant par rapport aux paramètres recherchés. Il peut donc être éliminé de la fonction  $J$ .
- Le terme  $\frac{\begin{bmatrix} \mathbf{w} & b \end{bmatrix} \cdot \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}}{2\sigma_0^2}$  s'appelle terme de **régularisation**. Il limite les degrés de liberté du modèle. Il permet de lutter contre l'**overfitting**.

Comment **trouver** un modèle qui colle à mes données?

Approche **MAP** : exemple de **régression linéaire** le retour..

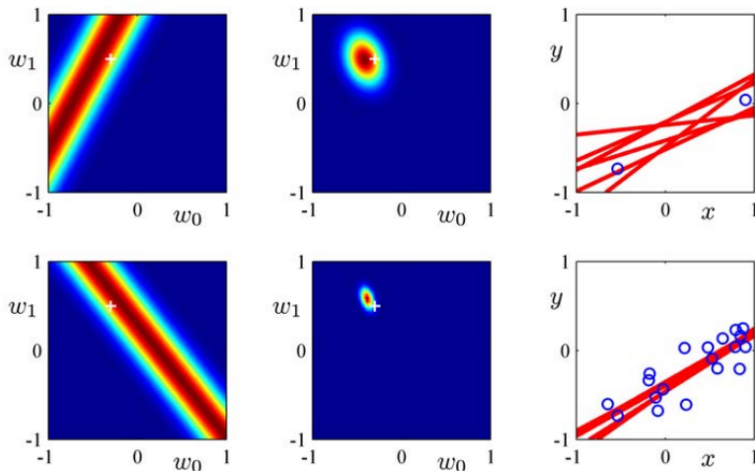
Illustration de l'influence du **prior**



Comment **trouver** un modèle qui colle à mes données ?

Approche **MAP** : exemple de **régression linéaire** le retour..

Illustration de l'influence du **prior**



## Messages importants du chapitre :

- En se dotant d'un **modèle paramétrique**, on parvient à estimer  $p_{X,Y|\theta}$  à partir des données.
- Cette distribution est utilisée dans les classifieurs/régresseurs **optimaux** au sens des **pertes classiques**.
- Le **prior** est dominé par la **likelihood** quand  $n$  croît.