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 minisera 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 [1], 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.

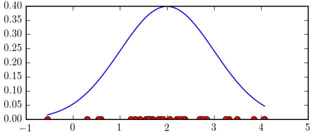
Plan du chapitre

Qu'est ce qu'un modèle paramétrique ?

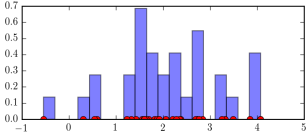
- C'est le choix d'une classe de distributions paramétrées pour $p_{X,Y}(x, y) = g_{\boldsymbol{\theta}}(x, y)$.
- C'est aussi le choix d'un espace d'hypothèse $\mathcal{H}$ (ensemble des prédicteurs parmi lesquels on doit choisir).
- En effet, le meilleur prédicteur est étroitement liée à $p_{X,Y}$

Qu'est ce qu'un modèle paramétrique ? Exemples

	Paramétrique	Non-Paramétrique
v.a. continue	loi Gaussienne, Beta, ..	Tout Histogramme, ..



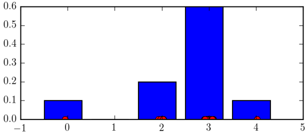
$\text{len}\left(\boldsymbol{\theta}\right) < \infty$

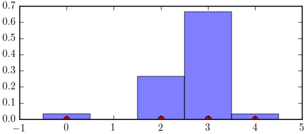


$\text{len}\left(\boldsymbol{\theta}\right) = \infty$

Qu'est ce qu'un modèle paramétrique ? Exemples

	Paramétrique	(Non-Paramétrique)
v.a. discrète	loi multinomiale, catégorique	Tout Histrogramme





→ pas de réelle différence, car la distribution catégorique correspond à toutes les distributions discrètes possibles !

Modèles paramétriques : catégories

Modèles Génératifs	Modèles Discriminatifs
On cherche $p_{X,Y}$	On cherche $p_{Y\|X}$
convergence plus rapide vers quelque chose de pertinent	légèrement meilleur quand n est grand

Plan du chapitre

Comment trouver un modèle qui colle à mes données ?
Approche fréquentiste
$\boxed{\text{Rappel}}$: la vraie valeur de $\boldsymbol{\theta}$ est fixe et inconnue.

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

$$\prod_{i=1}^{n} f_{\boldsymbol{\theta}}\left(\mathbf{x}^{(i)}, y^{(i)}\right) \tag{1}$$

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

Comment trouver un modèle qui colle à mes données ?
Approche fréquentiste

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

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

Comment trouver un modèle qui colle à mes données ?
Approche MLE : exemple de régression linéaire

- Je choisis un modèle discriminatif Gaussien :

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

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

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

Comment trouver un modèle qui colle à mes données ?
Approche MLE : exemple de régression linéaire

- Maximiser $\mathcal{L}(\boldsymbol{\theta}) =$ Minimiser $\mathrm{NLL}(\boldsymbol{\theta})$.

- En choisissant une fonction de perte quadratique, montrez que la NLL s'écrit :

$$\mathrm{NLL}(\boldsymbol{\theta}) = n \log(\sigma) + \frac{n}{2\sigma^2} Err_{\mathrm{train}} + cte \tag{4}$$

Comment trouver un modèle qui colle à mes données ?
Approche MLE : exemple de régression linéaire

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

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

avec

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

Comment trouver un modèle qui colle à mes données ?
Approche MLE : exemple de régression linéaire

- Concernant le paramètre σ, montrez que son estimé MLE est l'écart type empirique :

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

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 $\mathbb{E}[Y|X] = \hat{\mathbf{w}}^T.\mathbf{x}^{(i)} + \hat{b}$

- $\hat{\sigma}_{MLE}$ sert à prédire un niveau de confiance en le fait que Y soit proche de 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 :

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

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

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

$$\boldsymbol{\theta} = \tag{10}$$

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 $\left(\mathbf{x}^{(i)}, c^{(i)}\right)$

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

- La vraisemblance s'écrit donc :

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

Comment trouver un modèle qui colle à mes données ?
Approche MLE : exemple du classifieur naïf bayésien

- Passons à la NLL :

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

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

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

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 λ.
  - On modifie la fonction objectif en y incorporant la contrainte d'égalité, par exemple :

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

  - 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 :

$$
\begin{aligned}
\frac{\partial}{\partial \pi_{c_k}} J(\boldsymbol{\theta}) &= 0, \\
\Leftrightarrow \frac{\partial}{\partial \pi_{c_k}} \mathrm{NLL}(\boldsymbol{\theta}) + \lambda &= 0, && (13) \\
\Leftrightarrow \frac{\partial}{\partial \pi_{c_k}} \sum_{i=1}^{n} \log \pi_{c(i)} &= \lambda, && (14) \\
\Leftrightarrow n_k \times \frac{\partial}{\partial \pi_{c_k}} \log \pi_{c_k} &= \lambda, && (15) \\
\Leftrightarrow \frac{n_k}{\pi_{c_k}} &= \lambda. && (16)
\end{aligned}
$$

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

$$\begin{aligned} \sum_{k=1}^{\ell} n_k &= \sum_{k=1}^{\ell} \lambda \pi_{c_k}, && (17) \\ \Leftrightarrow n &= \lambda \sum_{k=1}^{\ell} \pi_{c_k}, && (18) \\ \Leftrightarrow n &= \lambda. && (19) \end{aligned}$$

- Au final, il vient :

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

Comment trouver un modèle qui colle à mes données ?
Approche MLE : exemple du classifieur naïf bayésien

- Passons à présent aux paramètres suivants : μ_{j,c_k}

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

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

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

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

Comment trouver un modèle qui colle à mes données ?
Approche MLE : exemple du classifieur naïf bayésien

- Calcul des μ_{j,c_k} (suite) :

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

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

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

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 : σ_{j,c_k}
- On montrerait de même que $\hat{\sigma}^2_{j,c_k,MLE}$ est la variance empirique des exemples de la classe c_k pour la dimension j :

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

- $\rightarrow$ Bon exercice à faire chez soi ..

Comment trouver un modèle qui colle à mes données ?
Approche MLE : exemple du classifieur naïf bayésien

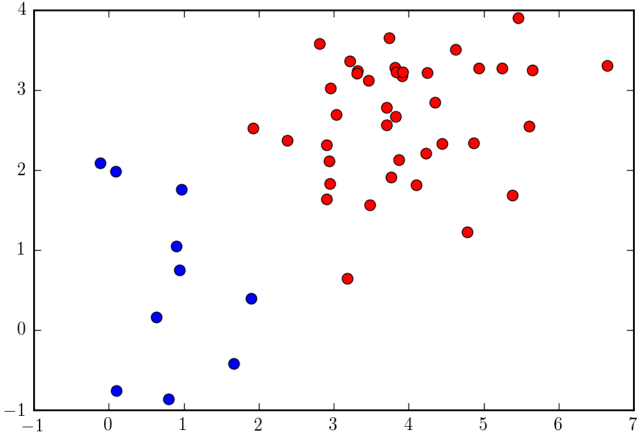
- Le classifieur naïf bayésien est un des modèles génératifs les plus utilisés.

- 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

Comment trouver un modèle qui colle à mes données ?

Approche Bayésienne

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

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

$$p_{\boldsymbol{\theta}|\text{données}} = \frac{p_{\text{données}|\boldsymbol{\theta}} \times p_{\boldsymbol{\theta}}}{p_{\text{données}}}. \tag{29}$$

Comment trouver un modèle qui colle à mes données ?

Approche Bayésienne

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

- Je cherche à présent la posterior sur $\boldsymbol{\theta}$:

$$p_{\boldsymbol{\theta}|\text{données}} = \frac{p_{\text{données}|\boldsymbol{\theta}} \times p_{\boldsymbol{\theta}}}{p_{\text{données}}}, \tag{30}$$

$$p_{\boldsymbol{\theta}|\mathcal{D}} = \frac{\mathcal{L}\left(\boldsymbol{\theta}\right) \times p_{\boldsymbol{\theta}}}{p_{\mathcal{D}}}, \tag{31}$$

$$p_{\boldsymbol{\theta}|\mathcal{D}} = \frac{\mathcal{L}\left(\boldsymbol{\theta}\right) \times p_{\boldsymbol{\theta}}}{\int \mathcal{L}\left(\boldsymbol{\theta}\right) \times p_{\boldsymbol{\theta}} d\boldsymbol{\theta}} \tag{32}$$

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

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

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

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

- $\tilde{p}_{\boldsymbol{\theta}|\mathcal{D}}$ 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 $\hat{\boldsymbol{\theta}}_{MAP}$, 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 :

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

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

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

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 $\mathbf{w}$ et b.

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

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

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

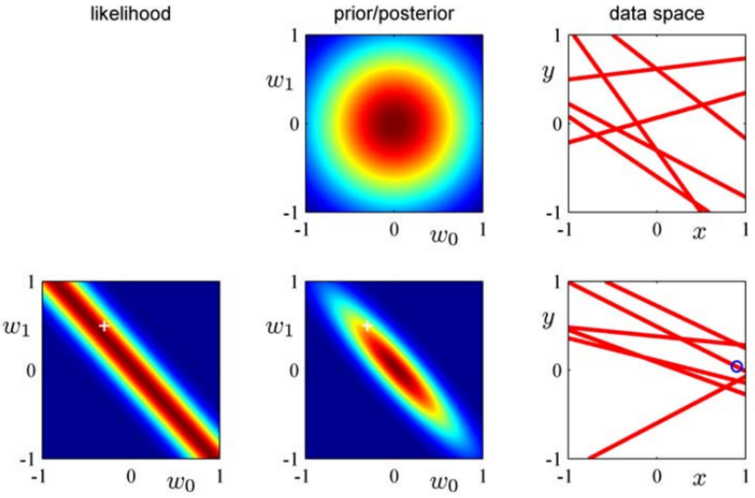
Comment trouver un modèle qui colle à mes données ?
Approche MAP : exemple de régression linéaire le retour..

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

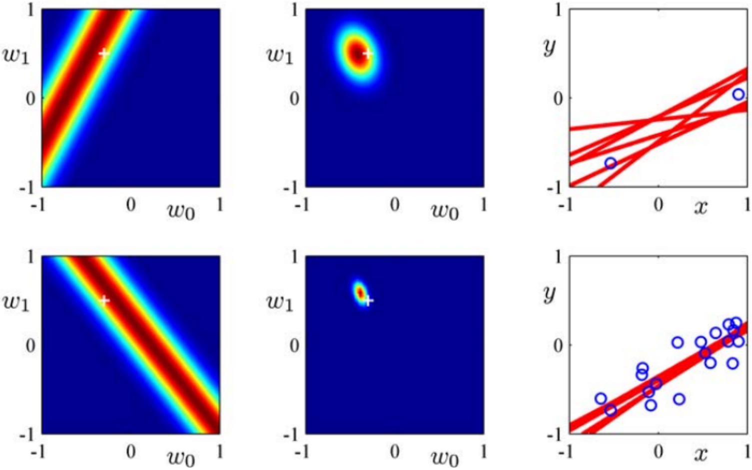
$$J\left(\mathbf{w}, b, \sigma\right) = \mathrm{NLL}\left(\mathbf{w}, b, \sigma\right) + \log\left(\sqrt{2\pi}\sigma_0\right) + \frac{\begin{bmatrix}\mathbf{w} & b\end{bmatrix} \cdot \begin{bmatrix}\mathbf{w} \\ b\end{bmatrix}}{2\sigma_0^2}. \quad (38)$$

- Le terme $\log\left(\sqrt{2\pi}\sigma_0\right)$ 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 $p_{X,Y|\boldsymbol{\theta}}$ 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.