Apprentissage automatique

Apprentissage bayésien - motivation

TYPES D'APPRENTISSAGE

Sujets: apprentissage supervisé, classification, régression

RAPPEL

- L'apprentissage supervisé est lorsqu'on a une cible à prédire
  - **classification :** la cible est un indice de classe $t \in \{1, \ldots, K\}$
    - exemple : reconnaissance de caractères
      - ✓ $\mathbf{x}$: vecteur des intensités de tous les pixels de l'image
      - ✓ t : identité du caractère
  - **régression :** la cible est un nombre réel $t \in \mathbb{R}$
    - exemple : prédiction de la valeur d'une action à la bourse
      - ✓ $\mathbf{x}$: vecteur contenant l'information sur l'activité économique de la journée
      - ✓ t : valeur d'une action à la bourse le lendemain

TYPES D'APPRENTISSAGE

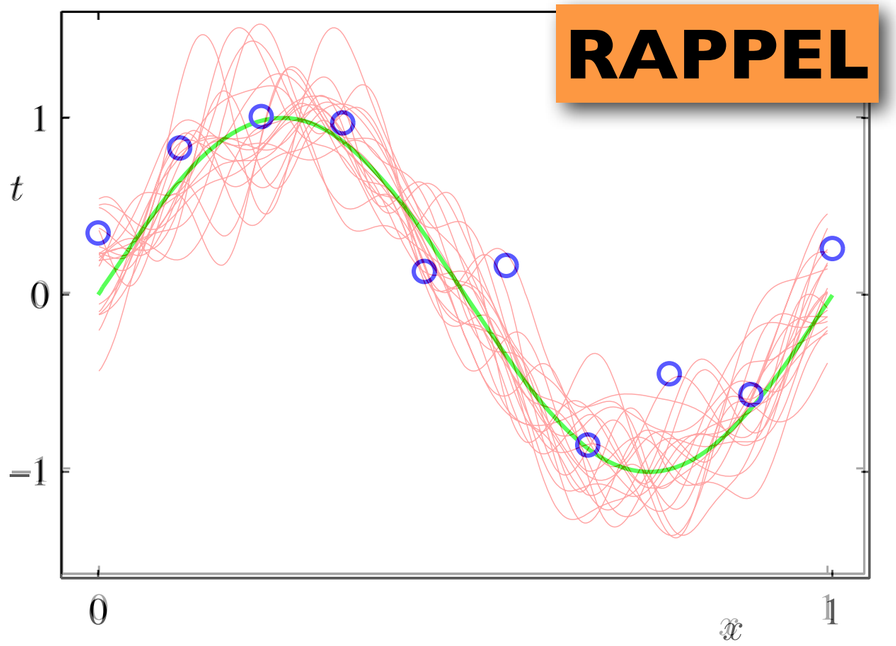
Sujets: apprentissage supervisé, classification, régression

RAPPEL

- L'apprentissage supervisé est lorsqu'on a une cible à prédire
  - **classification :** la cible est un indice de classe $t \in \{1, \ldots, K\}$
    - exemple : reconnaissance de caractères
      - ✓ $\mathbf{x}$: vecteur des intensités de tous les pixels de l'image
      - ✓ t : identité du caractère
  - **régression :** la cible est un nombre réel $t \in \mathbb{R}$
    - exemple : prédiction de la valeur d'une action à la bourse
      - ✓ $\mathbf{x}$: vecteur contenant l'information sur l'activité économique de la journée
      - ✓ t : valeur d'une action à la bourse le lendemain

EXEMPLE: RÉGRESSION

Sujets: minimisation de perte (coût, erreur)



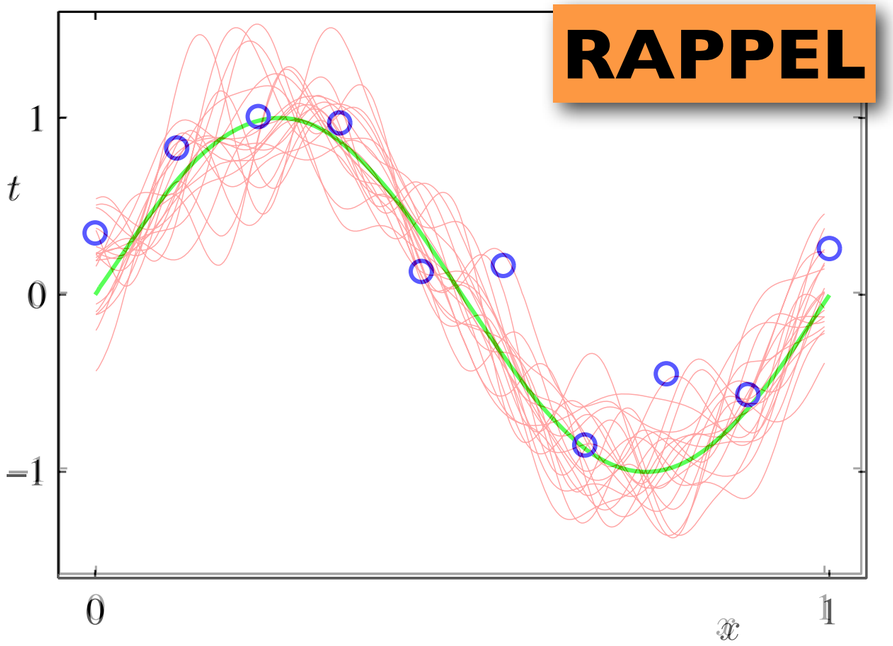
- Comme trouver $\mathbf{w}$? **(problème d'apprentissage)**
  - on cherche le $\mathbf{w}^\star$ qui minimise la somme de notre perte / erreur / coût sur l'ensemble d'entraînement

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

  - le « $\frac{1}{2}$ » n'est pas important (mais simplifiera certains calculs)
- Un algorithme d'apprentissage résoudrait ce problème
  - à partir des données, il va retourner $\mathbf{w}^\star$

EXEMPLE: RÉGRESSION

Sujets: minimisation de perte (coût, erreur)



- Comme trouver $\mathbf{w}$? **(problème d'apprentissage)**
  - on cherche le $\mathbf{w}^\star$ qui minimise la somme de notre perte / erreur / coût sur l'ensemble d'entraînement

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

  - le « $\frac{1}{2}$ » n'est pas important (mais simplifiera certains calculs)
- Un algorithme d'apprentissage résoudrait ce problème
  - à partir des données, il va retourner $\mathbf{w}^\star$

EXEMPLE: RÉGRESSION

Sujets: régularisation

RAPPEL

- Comment utiliser un grand M avec peu de données
  - par exemple, si on connait le «vrai» M
- **Régularisation** : on réduit la capacité autrement
  - exemple : on pénalise la somme du carré des paramètres (i.e. la norme Euclidienne au carré)

contrôle la capacité

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

  - où $\|\mathbf{w}\|^2 \equiv \mathbf{w}^{\mathrm{T}}\mathbf{w} = w_0^2 + w_1^2 + \ldots + w_M^2$

EXEMPLE: RÉGRESSION

Sujets: loi a priori et loi a posteriori

RAPPEL

- $p(\mathbf{w}|\alpha)$ exprime notre croyance a priori sur la valeur de $\mathbf{w}$
  - c'est une **loi a priori** (*prior*)
- Lorsqu'on observe des données, on peut mettre à jour notre croyance

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

  - c'est la **loi a posteriori** (*posterior*)

EXEMPLE: RÉGRESSION

Sujets: loi a priori et loi a posteriori

RAPPEL

- $p(\mathbf{w}|\alpha)$ exprime notre croyance a priori sur la valeur de $\mathbf{w}$
  - c'est une **loi a priori** (*prior*)
- Lorsqu'on observe des données, on peut mettre à jour notre croyance

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

  - c'est la **loi a posteriori** (*posterior*)

EXEMPLE: RÉGRESSION

Sujets: loi a priori et loi a posteriori

RAPPEL

- $p(\mathbf{w}|\alpha)$ exprime notre croyance a priori sur la valeur de $\mathbf{w}$

  - c'est une **loi a priori** (*prior*)

- Lorsqu'on observe des données, on peut mettre à jour notre croyance

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

  - c'est la **loi a posteriori** (*posterior*)

EXEMPLE: RÉGRESSION

Sujets: maximum a posteriori

RAPPEL

- On pourrait choisir le modèle $\mathbf{w}$ qui est le plus (log-)probable selon nos croyances a posteriori $p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta)$
  - on appelle ça la solution **maximum a posteriori**

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

- Équivalent à la perte régularisée si $\lambda = \alpha/\beta$

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

EXEMPLE: RÉGRESSION

Sujets: maximum a posteriori

RAPPEL

- On pourrait choisir le modèle $\mathbf{w}$ qui est le plus (log-)probable selon nos croyances a posteriori $p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta)$
  - on appelle ça la solution **maximum a posteriori**

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

- Équivalent à la perte régularisée si $\lambda = \alpha/\beta$

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

APPRENTISSAGE BAYÉSIEN

Sujets: apprentissage bayésien

- Pourquoi choisir un seul $\mathbf{w}$?
  - peut-être que d'autres valeurs de $\mathbf{w}$ ont aussi une probabilité a posteriori élevée
- Cette observation motive l'**apprentissage bayésien**
  - on va tenir compte de notre incertitude sur la bonne valeur de $\mathbf{w}$
  - le résultat sera un **ensemble** de modèles, où chaque modèle aura un poids $p(\mathbf{w}|\text{«données»})$

APPRENTISSAGE BAYÉSIEN

Sujets: apprentissage bayésien

- L'apprentissage bayésien ne sera pas interprétable comme la minimisation d'une somme de pertes (régularisée)
  - c'est une approche fondamentalement différente à la minimisation de perte

- L'apprentissage bayésien aura tendance à être moins affecté par le sur-apprentissage
  - c'est dû au fait qu'on ne se commet pas à une seule valeur du modèle (un seul $\mathbf{w}$)
  - on tient compte de la variance par rapport à $\mathbf{w}$

APPRENTISSAGE BAYÉSIEN

Sujets: apprentissage bayésien

- En résumé, l'apprentissage bayésien c'est
  - définir $p(\text{«modèle»})$
  - définir $p(\text{«données»}|\text{«modèle»})$
  - calculer et manipuler la loi a posteriori

$$p(\text{«modèle»}|\text{«données»}) \propto p(\text{«données»}|\text{«modèle»})\ p(\text{«modèle»})$$

afin de faire des prédictions sur de nouvelles données

Apprentissage automatique

Apprentissage bayésien - régression linéaire bayésienne

APPRENTISSAGE BAYÉSIEN

Sujets: apprentissage bayésien

- En résumé, l'apprentissage bayésien c'est
  - définir $p(\text{«modèle»})$
  - définir $p(\text{«données»}|\text{«modèle»})$
  - calculer et manipuler la loi a posteriori

$$p(\text{«modèle»}|\text{«données»}) \propto p(\text{«données»}|\text{«modèle»})\ p(\text{«modèle»})$$

afin de faire des prédictions sur de nouvelles données

APPRENTISSAGE BAYÉSIEN

Sujets: apprentissage bayésien

- En résumé, l'apprentissage bayésien c'est
  - définir $p(\text{«modèle»})$
  - définir $p(\text{«données»}|\text{«modèle»})$
  - calculer et manipuler la loi a posteriori

$$p(\text{«modèle»}|\text{«données»}) \propto p(\text{«données»}|\text{«modèle»})\ p(\text{«modèle»})$$

afin de faire des prédictions sur de nouvelles données

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

- En résumé, l'apprentissage bayésien c'est
  - définir $p(\text{«modèle»})$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$= \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

- En résumé, l'apprentissage bayésien c'est
  - définir $p(\text{«modèle»})$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

à quel point w s'éloigne de 0 (hyper-paramètre)

$$= \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

- En résumé, l'apprentissage bayésien c'est
  - définir $p(\text{«données»}|\text{«modèle»})$

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

- En résumé, l'apprentissage bayésien c'est
  - définir $p(\text{«données»}|\text{«modèle»})$

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

à quel point t_n s'éloigne de $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)$
(hyper-paramètre)

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

- En résumé, l'apprentissage bayésien c'est
  - définir $p(\text{«données»}|\text{«modèle»})$

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

à quel point t_n s'éloigne de $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)$
(hyper-paramètre)

ou de façon équivalente :

$$\mathbf{t} = \boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\epsilon}$$

où $\boldsymbol{\epsilon}$ est gaussien, de moyenne 0 et matrice de covariance $\beta^{-1}\mathbf{I}$

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

- En résumé, l'apprentissage bayésien c'est
  - calculer la loi a posteriori $p(\text{«modèle»} | \text{«données»})$

On remarque que $p(\mathbf{w}, \text{«données»})$ est gaussien, puisque

$$\begin{pmatrix} \mathbf{w} \\ \mathbf{t} \end{pmatrix} = \begin{pmatrix} \mathbf{w} \\ \mathbf{\Phi w} + \boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{\Phi} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ \boldsymbol{\epsilon} \end{pmatrix}$$

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

- En résumé, l'apprentissage bayésien c'est
  - calculer la loi a posteriori $p(\text{«modèle»} | \text{«données»})$

On remarque que $p(\mathbf{w}, \text{«données»})$ est gaussien, puisque

$$\begin{pmatrix} \mathbf{w} \\ \mathbf{t} \end{pmatrix} = \begin{pmatrix} \mathbf{w} \\ \mathbf{\Phi w} + \boldsymbol{\epsilon} \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{\Phi} & \mathbf{I} \end{pmatrix}}_{\text{transformation linéaire}} \underbrace{\begin{pmatrix} \mathbf{w} \\ \boldsymbol{\epsilon} \end{pmatrix}}_{\text{gaussien}}$$

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

- En résumé, l'apprentissage bayésien c'est
  - calculer la loi a posteriori $p(\text{«modèle»} | \text{«données»})$

On remarque que $p(\mathbf{w}, \text{«données»})$ est gaussien, puisque

$$\underbrace{\begin{pmatrix} \mathbf{w} \\ \mathbf{t} \end{pmatrix}}_{\text{donc gaussien}} = \begin{pmatrix} \mathbf{w} \\ \mathbf{\Phi w} + \boldsymbol{\epsilon} \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{\Phi} & \mathbf{I} \end{pmatrix}}_{\text{transformation linéaire}} \underbrace{\begin{pmatrix} \mathbf{w} \\ \boldsymbol{\epsilon} \end{pmatrix}}_{\text{gaussien}}$$

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésien

- En résumé, l'apprentissage bayésien c'est
  - calculer la loi a posteriori $p(\text{«modèle»}|\text{«données»})$

On remarque que $p(\mathbf{w}, \text{«données»})$ est gaussien avec paramètre $\boldsymbol{\mu}$

$$\boldsymbol{\mu} = \begin{pmatrix} \mathbb{E}[\mathbf{w}] \\ \mathbb{E}[\boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\epsilon}] \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}$$

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

- En résumé, l'apprentissage bayésien c'est
  - calculer la loi a posteriori $p(\text{«modèle»}|\text{«données»})$

On remarque que $p(\mathbf{w}, \text{«données»})$ est gaussien avec paramètre $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \text{cov}[\mathbf{w}] & \text{cov}[\mathbf{w}, \boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\epsilon}] \\ \text{cov}[\boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\epsilon}, \mathbf{w}] & \text{cov}[\boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\epsilon}] \end{pmatrix}$$

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

- En résumé, l'apprentissage bayésien c'est
  - calculer la loi a posteriori $p(\text{«modèle»}|\text{«données»})$

On remarque que $p(\mathbf{w}, \text{«données»})$ est gaussien avec paramètre $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \alpha^{-1}\mathbf{I} & \text{cov}[\mathbf{w}, \boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\epsilon}] \\ \text{cov}[\boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\epsilon}, \mathbf{w}] & \text{cov}[\boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\epsilon}] \end{pmatrix}$$

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

- En résumé, l'apprentissage bayésien c'est
  - calculer la loi a posteriori $p(\text{«modèle»}|\text{«données»})$

On remarque que $p(\mathbf{w}, \text{«données»})$ est gaussien avec paramètre $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \alpha^{-1}\mathbf{I} & \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}} \\ \mathrm{cov}[\boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\epsilon}, \mathbf{w}] & \mathrm{cov}[\boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\epsilon}] \end{pmatrix}$$

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

- En résumé, l'apprentissage bayésien c'est
  - calculer la loi a posteriori $p(\text{«modèle»} | \text{«données»})$

On remarque que $p(\mathbf{w}, \text{«données»})$ est gaussien avec paramètre $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \alpha^{-1}\mathbf{I} & \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}} \\ \alpha^{-1}\boldsymbol{\Phi} & \mathrm{cov}[\boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\epsilon}] \end{pmatrix}$$

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

- En résumé, l'apprentissage bayésien c'est
  - calculer la loi a posteriori $p(\text{«modèle»}|\text{«données»})$

On remarque que $p(\mathbf{w}, \text{«données»})$ est gaussien avec paramètre $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \alpha^{-1}\mathbf{I} & \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}} \\ \alpha^{-1}\boldsymbol{\Phi} & \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I} \end{pmatrix}$$

LOI DE PROBABILITÉ GAUSSIENNE

Sujets: loi conditionnelle d'une gaussienne

RAPPEL

- Soit $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)^{\mathrm{T}}$ une variable aléatoire gaussienne, de moyenne et matrice de covariance

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

- La **loi conditionnelle** $p(\mathbf{x}_a|\mathbf{x}_b)$ est aussi gaussienne :

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \end{aligned}$$

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

$$\begin{aligned}\boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)\\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}\end{aligned}$$

- En résumé, l'apprentissage bayésien c'est
  - calculer la loi a posteriori $p(\text{«modèle»}|\text{«données»})$

$p(\mathbf{w}, \text{«données»})$ est gaussien avec paramètres

$$\boldsymbol{\mu} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \alpha^{-1}\mathbf{I} & \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}} \\ \alpha^{-1}\boldsymbol{\Phi} & \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I} \end{pmatrix}$$

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

$$\begin{aligned}\boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)\\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}\end{aligned}$$

- En résumé, l'apprentissage bayésien c'est
  - calculer la loi a posteriori $p(\text{«modèle»}|\text{«données»})$

$p(\mathbf{w}, \text{«données»})$ est gaussien avec paramètres

$$\boldsymbol{\mu} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \alpha^{-1}\mathbf{I} & \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}} \\ \alpha^{-1}\boldsymbol{\Phi} & \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I} \end{pmatrix}$$

donc $p(\mathbf{w}|\text{«données»})$ est gaussien avec paramètres

$$\boldsymbol{\mu}_{\mathbf{w}|\mathbf{t}} = \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\mathbf{t}$$

$$\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{t}} = \alpha^{-1}\mathbf{I} - \alpha^{-2}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\boldsymbol{\Phi}$$

Apprentissage automatique

Apprentissage bayésien - exemple : régression linéaire bayésienne

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

RAPPEL

- En résumé, l'apprentissage bayésien c'est
  - calculer la loi a posteriori $p(\text{«modèle»}|\text{«données»})$

$p(\mathbf{w}, \text{«données»})$ est gaussien avec paramètres

$$\boldsymbol{\mu} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \alpha^{-1}\mathbf{I} & \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}} \\ \alpha^{-1}\boldsymbol{\Phi} & \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I} \end{pmatrix}$$

donc $p(\mathbf{w}|\text{«données»})$ est gaussien avec paramètres

$$\boldsymbol{\mu}_{\mathbf{w}|\mathbf{t}} = \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\mathbf{t}$$

$$\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{t}} = \alpha^{-1}\mathbf{I} - \alpha^{-2}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\boldsymbol{\Phi}$$

RÉGRESSION LINÉAIRE BAYÉSIENNE

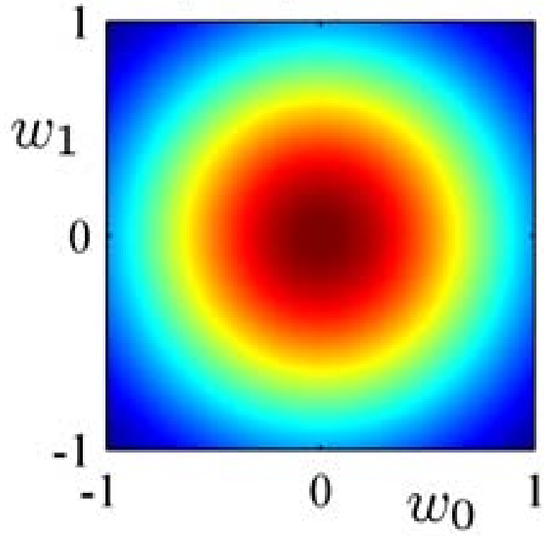
Sujets: régression linéaire bayésienne

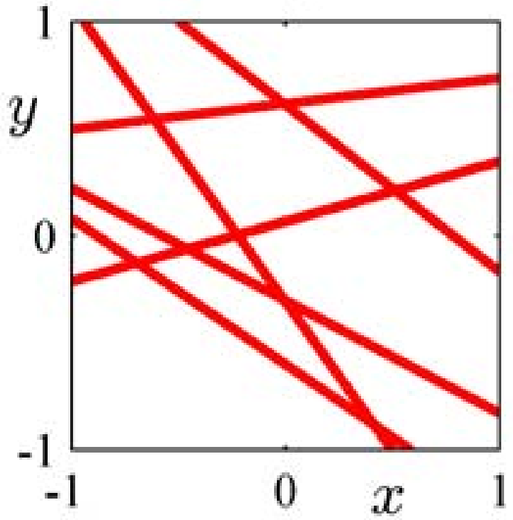
- Exemple (D=1)

$$p(\mathbf{w}|\text{«données»}) \propto p(\text{«données»}|\mathbf{w})\ p(\mathbf{w})$$

Échantillons de $p(\mathbf{w}|\text{«données»})$

$p(\mathbf{w}|\text{«données»})$



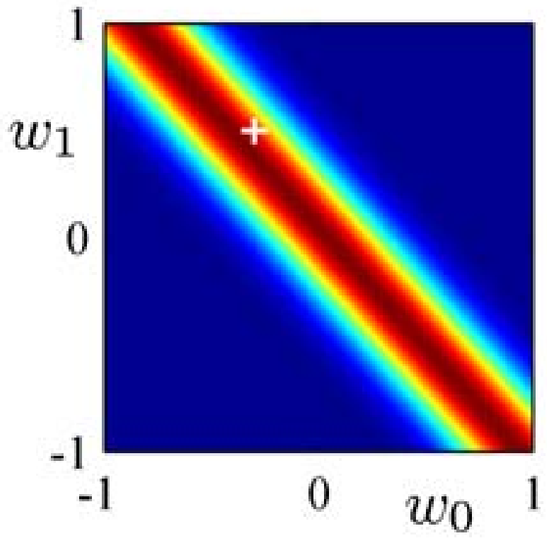


RÉGRESSION LINÉAIRE BAYÉSIENNE

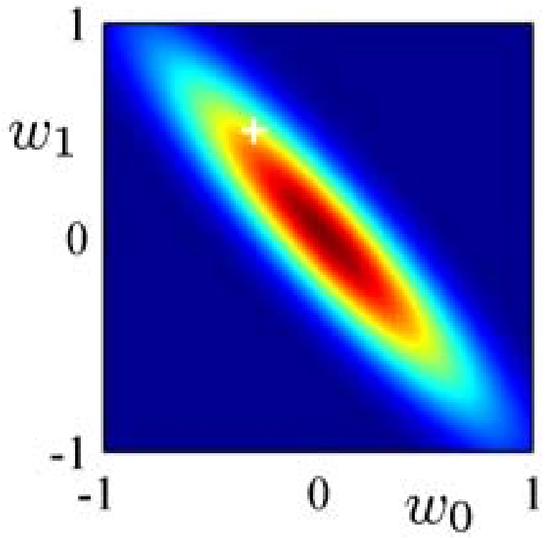
Sujets: régression linéaire bayésienne

- Exemple ($D=1$)

$$p(\mathbf{w}|\text{«données»}) \propto p(\text{«données»}|\mathbf{w})\ p(\mathbf{w})$$

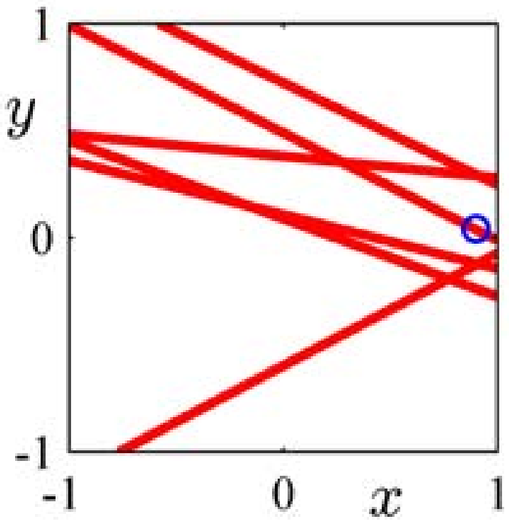


$p(\mathbf{w}|\text{«données»})$



Échantillons de

$p(\mathbf{w}|\text{«données»})$

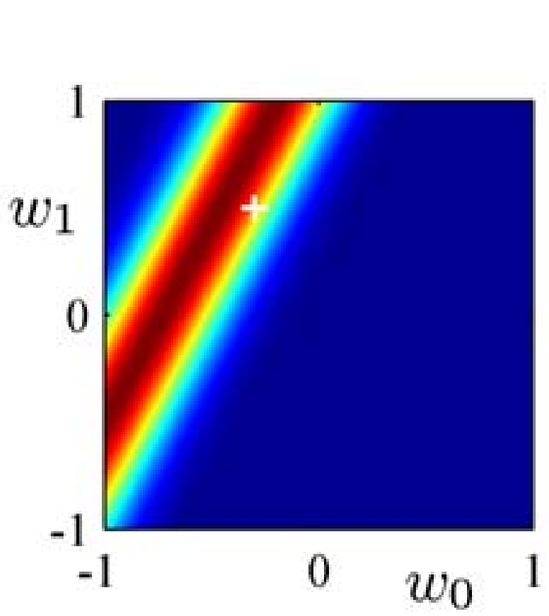


RÉGRESSION LINÉAIRE BAYÉSIENNE

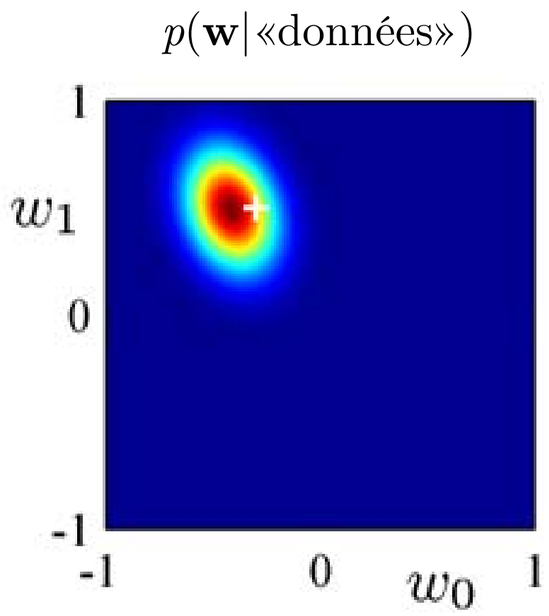
Sujets: régression linéaire bayésienne

- Exemple ($D=1$)

$$p(\mathbf{w}|\text{«données»}) \propto p(\text{«données»}|\mathbf{w})\ p(\mathbf{w})$$

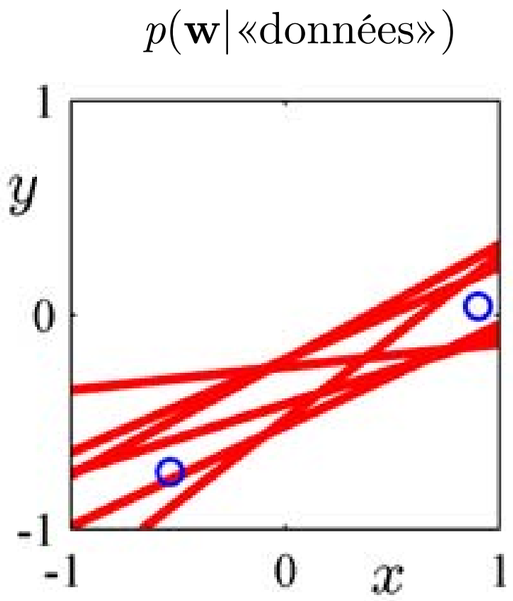


$p(\mathbf{w}|\text{«données»})$



Échantillons de

$p(\mathbf{w}|\text{«données»})$

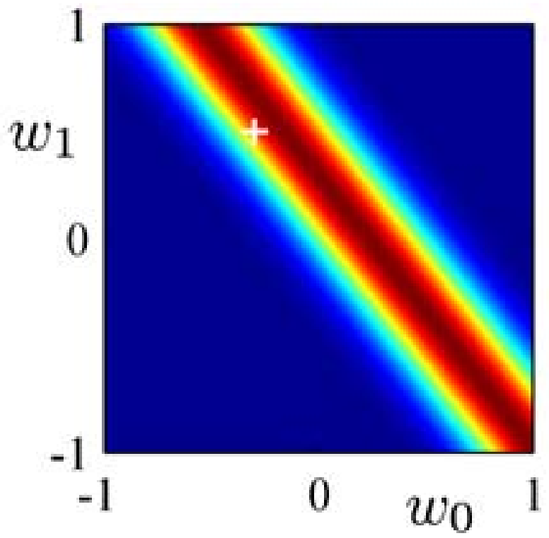


RÉGRESSION LINÉAIRE BAYÉSIENNE

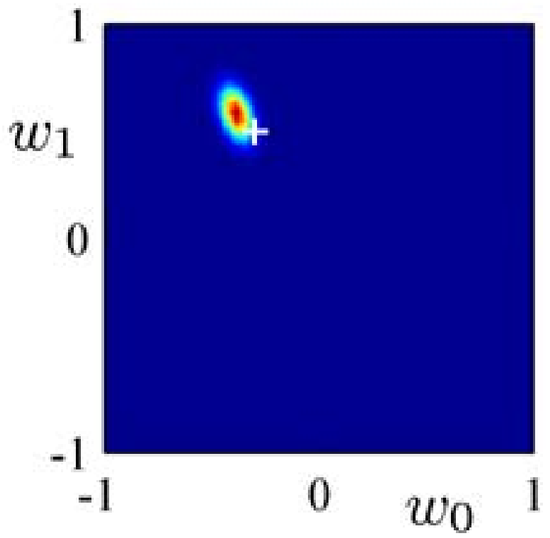
Sujets: régression linéaire bayésienne

- Exemple ($D=1$)

$$p(\mathbf{w}|\text{«données»}) \propto p(\text{«données»}|\mathbf{w})\ p(\mathbf{w})$$

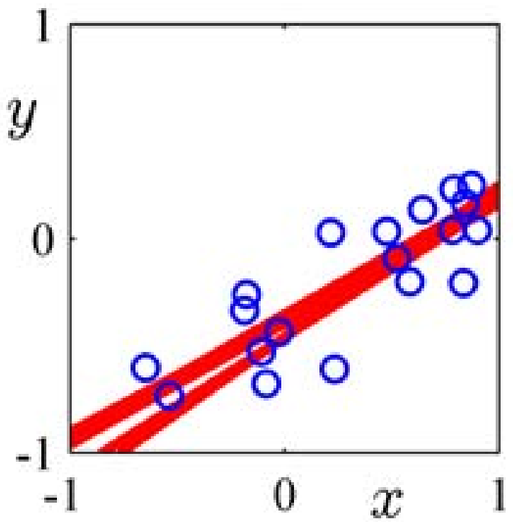


$p(\mathbf{w}|\text{«données»})$



Échantillons de

$p(\mathbf{w}|\text{«données»})$



Apprentissage automatique

Apprentissage bayésien - loi prédictive a posteriori

APPRENTISSAGE BAYÉSIEN

Sujets: apprentissage bayésien

- En résumé, l'apprentissage bayésien c'est
  - définir $p(\text{«modèle»})$
  - définir $p(\text{«données»}|\text{«modèle»})$
  - calculer et manipuler la loi a posteriori

$$p(\text{«modèle»}|\text{«données»}) \propto p(\text{«données»}|\text{«modèle»})\ p(\text{«modèle»})$$

afin de faire des prédictions sur de nouvelles données

APPRENTISSAGE BAYÉSIEN

Sujets: apprentissage bayésien

- En résumé, l'apprentissage bayésien c'est
  - définir $p(\text{«modèle»})$
  - définir $p(\text{«données»}|\text{«modèle»})$
  - calculer et manipuler la loi a posteriori

$$p(\text{«modèle»}|\text{«données»}) \propto p(\text{«données»}|\text{«modèle»})\ p(\text{«modèle»})$$

afin de faire des prédictions sur de nouvelles données

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

RAPPEL

- En résumé, l'apprentissage bayésien c'est
  - calculer la loi a posteriori $p(\text{«modèle»}|\text{«données»})$

$p(\mathbf{w}, \text{«données»})$ est gaussien avec paramètres

$$\boldsymbol{\mu} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \alpha^{-1}\mathbf{I} & \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}} \\ \alpha^{-1}\boldsymbol{\Phi} & \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I} \end{pmatrix}$$

donc $p(\mathbf{w}|\text{«données»})$ est gaussien avec paramètres

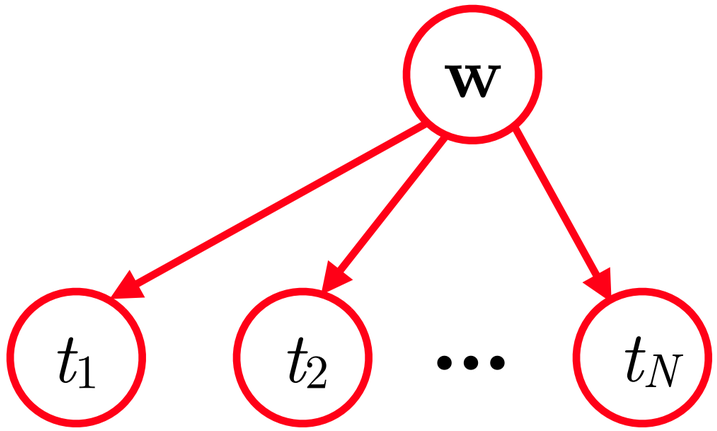
$$\boldsymbol{\mu}_{\mathbf{w}|\mathbf{t}} = \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\mathbf{t}$$

$$\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{t}} = \alpha^{-1}\mathbf{I} - \alpha^{-2}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\boldsymbol{\Phi}$$

PRÉDICTION BAYÉSIENNE

Sujets: prédiction bayésienne

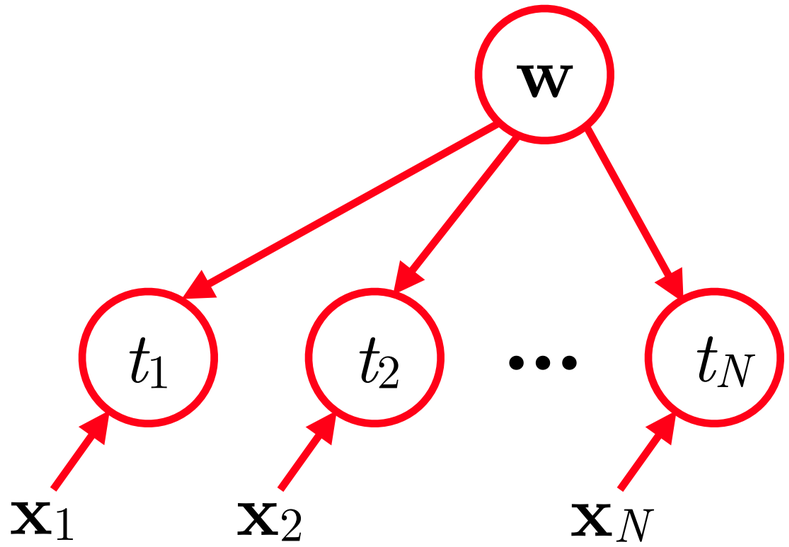
- Notre modèle suppose que les données ont été générées par le réseau bayésien suivant :



PRÉDICTION BAYÉSIENNE

Sujets: prédiction bayésienne

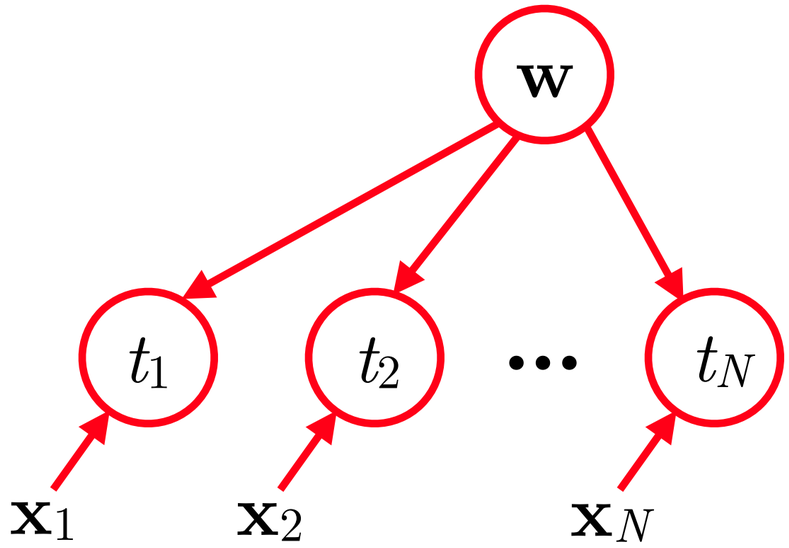
- Notre modèle suppose que les données ont été générées par le réseau bayésien suivant :



PRÉDICTION BAYÉSIENNE

Sujets: prédiction bayésienne

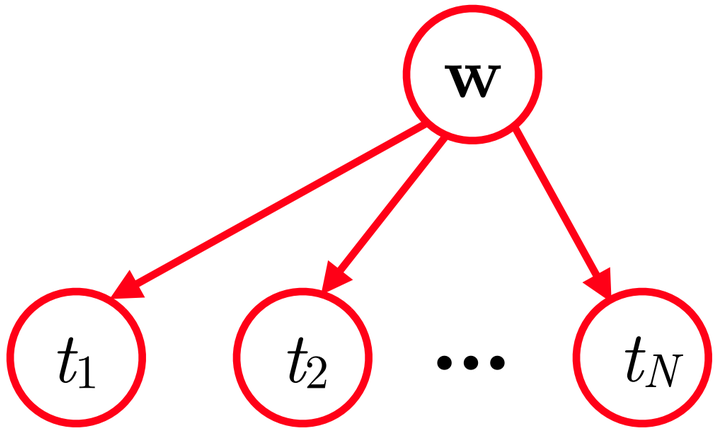
- Faire une prédiction d'une nouvelle cible t consiste à faire l'inférence pour cette cible, étant données $\mathbf{t}=(t_1,\ldots,t_N)^{\mathrm{T}}$



PRÉDICTION BAYÉSIENNE

Sujets: prédiction bayésienne

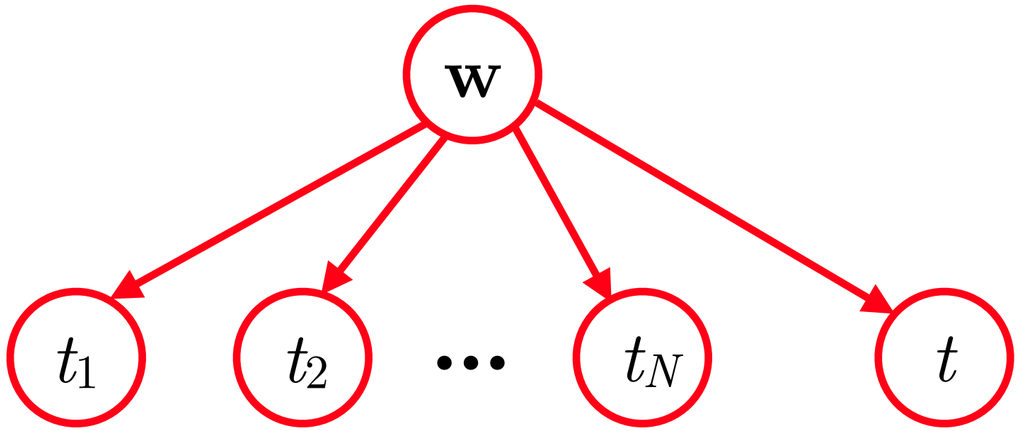
- Faire une prédiction d'une nouvelle cible t consiste à faire l'inférence pour cette cible, étant données $\mathbf{t}=(t_1,\ldots,t_N)^{\mathrm{T}}$



PRÉDICTION BAYÉSIENNE

Sujets: prédiction bayésienne

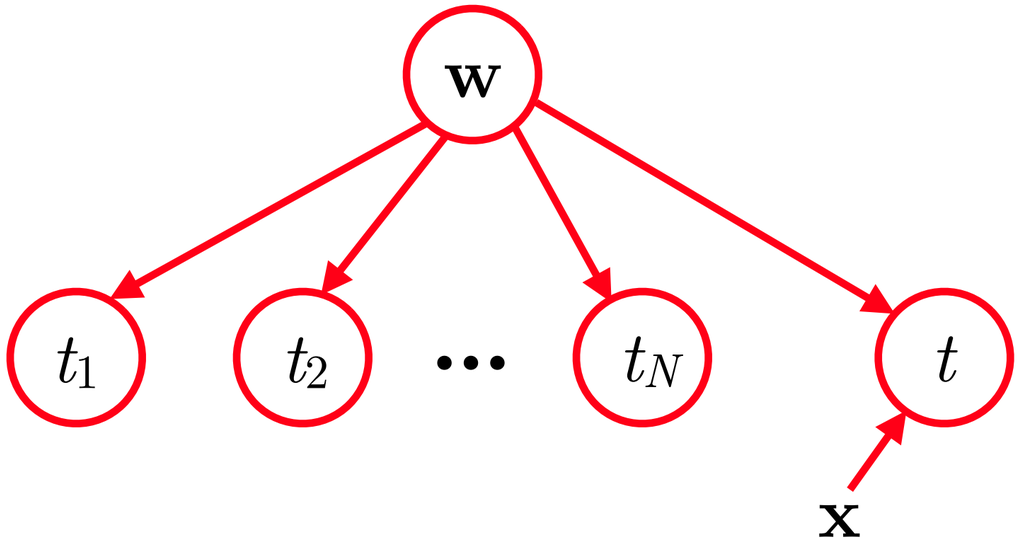
- Faire une prédiction d'une nouvelle cible t consiste à faire l'inférence pour cette cible, étant données $\mathbf{t}=(t_1,\ldots,t_N)^{\mathrm{T}}$



PRÉDICTION BAYÉSIENNE

Sujets: prédiction bayésienne

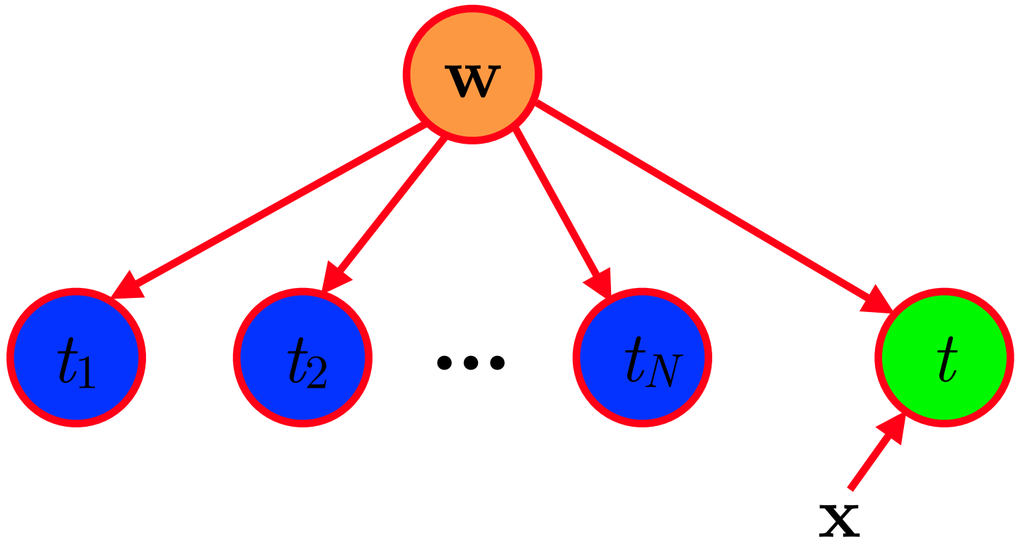
- Faire une prédiction d'une nouvelle cible t consiste à faire l'inférence pour cette cible, étant données $\mathbf{t}=(t_1,\ldots,t_N)^{\mathrm{T}}$



PRÉDICTION BAYÉSIENNE

Sujets: prédiction bayésienne

- Faire une prédiction d'une nouvelle cible t consiste à faire l'inférence pour cette cible, étant données $\mathbf{t}=(t_1,\ldots,t_N)^{\mathrm{T}}$

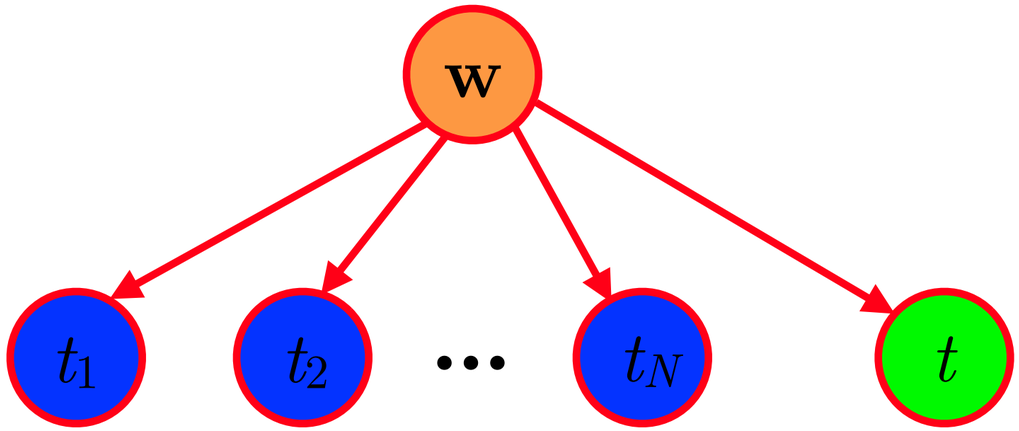


- à prédire
- connus
- inconnu

PRÉDICTION BAYÉSIENNE

Sujets: prédiction bayésienne

- Faire une prédiction d'une nouvelle cible t consiste à faire l'inférence pour cette cible, étant données $\mathbf{t}=(t_1,\ldots,t_N)^{\mathrm{T}}$



- à prédire
- connus
- inconnu

LOI PRÉDICTIVE A POSTERIORI

Sujets: loi prédictive a posteriori

- On doit calculer la **loi prédictive a posteriori** :

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) \, \mathrm{d}\mathbf{w}$$

- On voit que c'est équivalent à un ensemble (infini) où chaque modèle $p(t|\mathbf{w}, \beta)$ est pondéré par $p(\mathbf{w}|\mathbf{t}, \alpha, \beta)$

LOI PRÉDICTIVE A POSTERIORI

Sujets: loi prédictive a posteriori

- On doit calculer la **loi prédictive a posteriori** :

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) \, \mathrm{d}\mathbf{w}$$

- Régression linéaire : $t = \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}) + \epsilon$
  - donc t est gaussien, avec paramètres

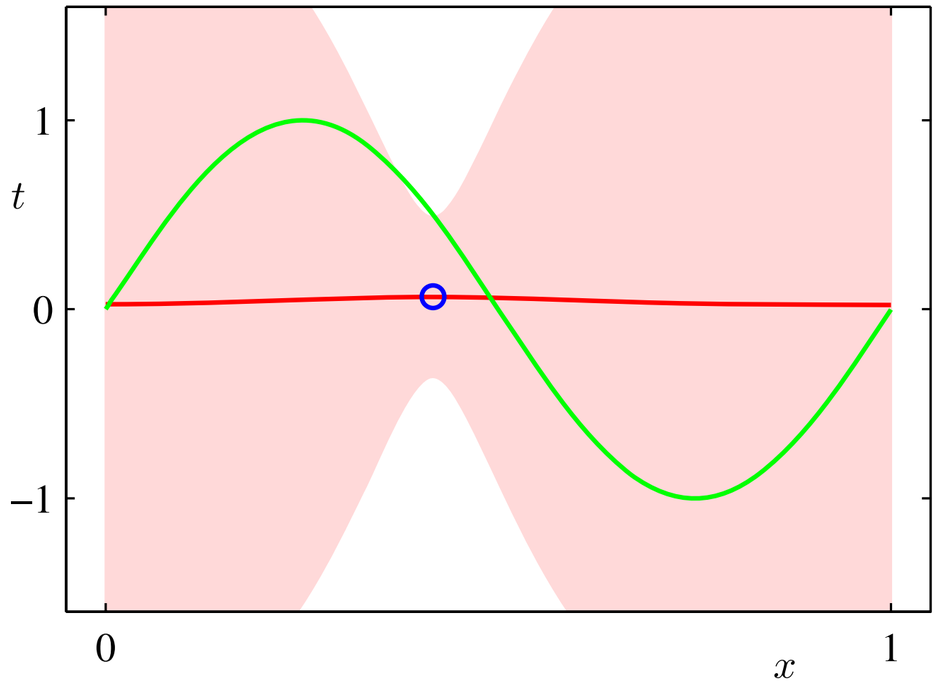
$$\mu_{t|\mathrm{t}} = \mathbb{E}[\mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}) + \epsilon|\mathrm{t}] = \boldsymbol{\mu}_{\mathbf{w}|\mathrm{t}}^{\mathrm{T}} \phi(\mathbf{x})$$

$$\sigma_{t|\mathrm{t}}^2 = \mathrm{var}(\mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}) + \epsilon|\mathrm{t}) = \phi(\mathbf{x})^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{w}|\mathrm{t}} \phi(\mathbf{x}) + \frac{1}{\beta}$$

LOI PRÉDICTIVE A POSTERIORI

Sujets: loi prédictive a posteriori

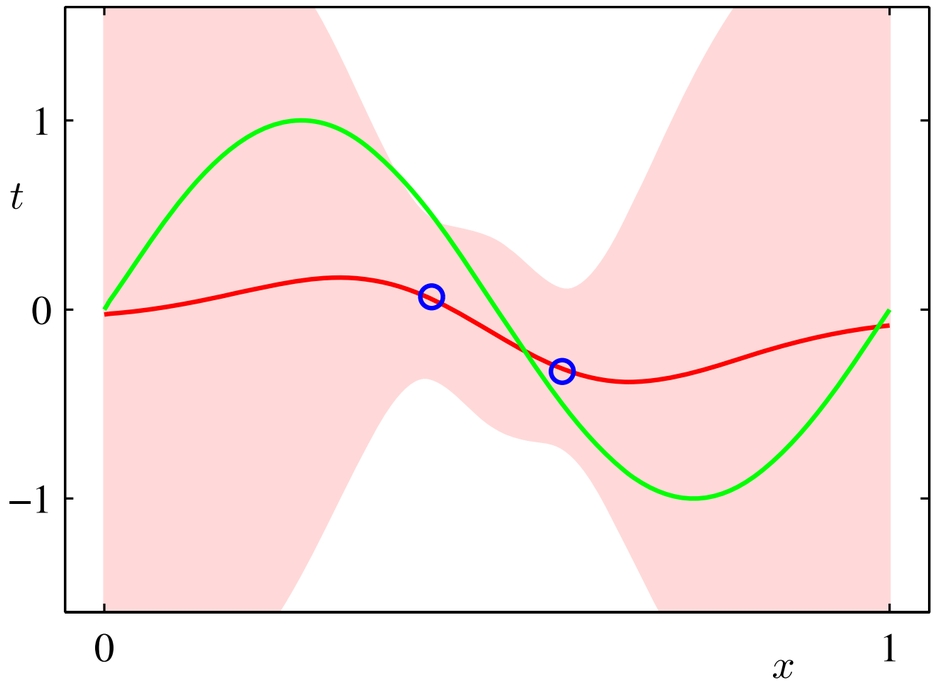
- En plus d'une prédiction $\mu_{t|\mathrm{t}}$, on a notre incertitude sur notre prédiction $\sigma^2_{t|\mathrm{t}}$



LOI PRÉDICTIVE A POSTERIORI

Sujets: loi prédictive a posteriori

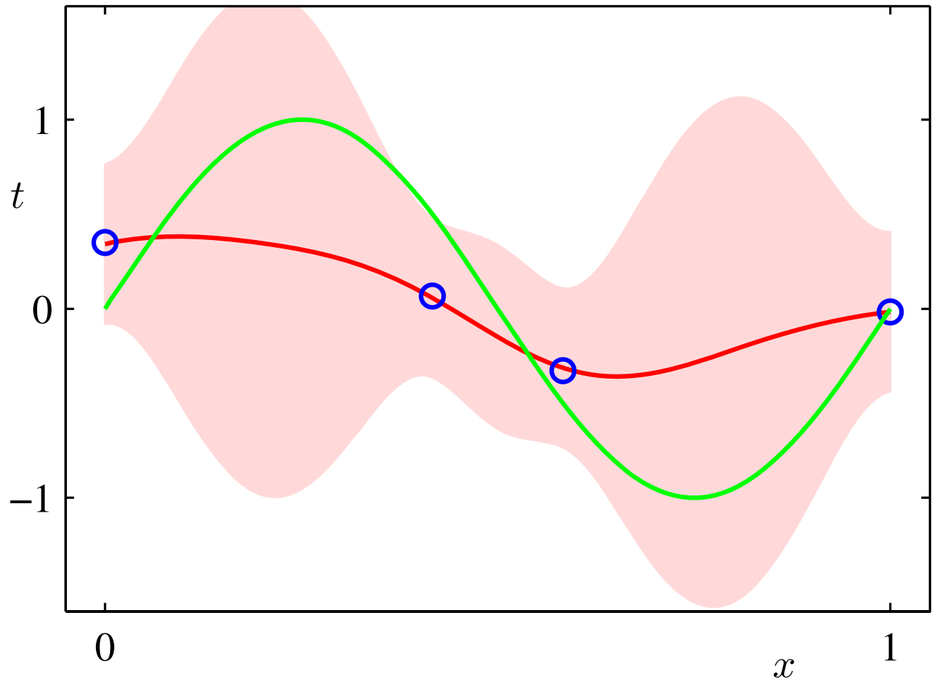
- En plus d'une prédiction $\mu_{t|\mathrm{t}}$, on a notre incertitude sur notre prédiction $\sigma^2_{t|\mathrm{t}}$



LOI PRÉDICTIVE A POSTERIORI

Sujets: loi prédictive a posteriori

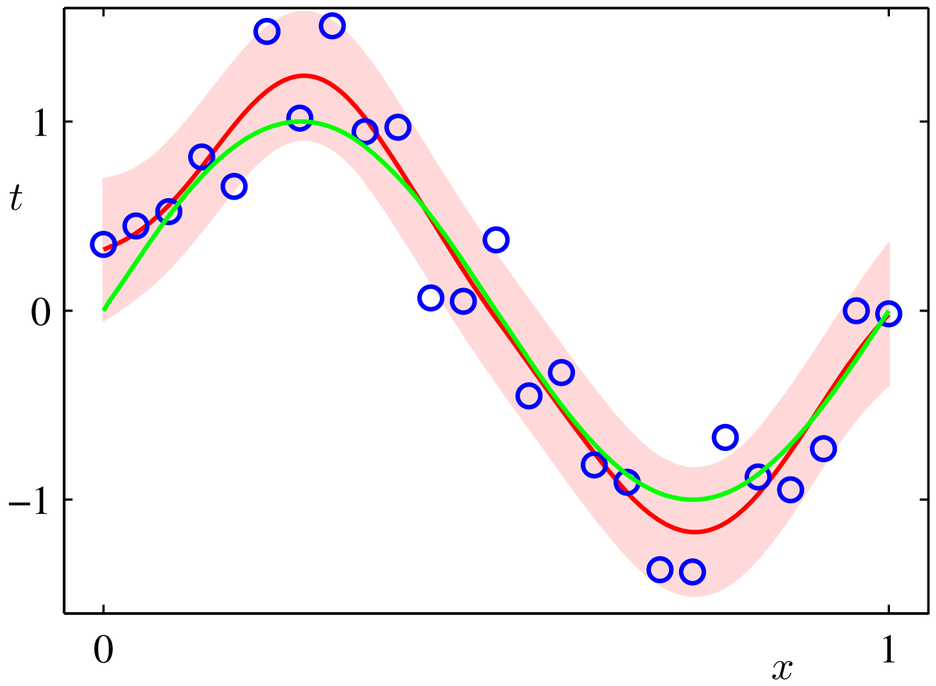
- En plus d'une prédiction $\mu_{t|\mathrm{t}}$, on a notre incertitude sur notre prédiction $\sigma^2_{t|\mathrm{t}}$



LOI PRÉDICTIVE A POSTERIORI

Sujets: loi prédictive a posteriori

- En plus d'une prédiction $\mu_{t|\mathrm{t}}$, on a notre incertitude sur notre prédiction $\sigma^2_{t|\mathrm{t}}$



Apprentissage automatique

Apprentissage bayésien - régression à noyau bayésienne

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: régression linéaire bayésienne

RAPPEL

- En résumé, l'apprentissage bayésien c'est
  - calculer la loi a posteriori $p(\text{«modèle»}|\text{«données»})$

$p(\mathbf{w}, \text{«données»})$ est gaussien avec paramètres

$$\boldsymbol{\mu} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \alpha^{-1}\mathbf{I} & \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}} \\ \alpha^{-1}\boldsymbol{\Phi} & \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I} \end{pmatrix}$$

donc $p(\mathbf{w}|\text{«données»})$ est gaussien avec paramètres

$$\boldsymbol{\mu}_{\mathbf{w}|\mathrm{t}} = \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\mathrm{t}$$

$$\boldsymbol{\Sigma}_{\mathbf{w}|\mathrm{t}} = \alpha^{-1}\mathbf{I} - \alpha^{-2}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\boldsymbol{\Phi}$$

LOI PRÉDICTIVE A POSTERIORI

Sujets: loi prédictive a posteriori

RAPPEL

- On doit calculer la **loi prédictive a posteriori** :

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) \, \mathrm{d}\mathbf{w}$$

- Régression linéaire : $t = \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}) + \epsilon$
  - donc t est gaussien, avec paramètres

$$\mu_{t|\mathrm{t}} = \mathbb{E}[\mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}) + \epsilon|\mathrm{t}] = \boldsymbol{\mu}_{\mathbf{w}|\mathrm{t}}^{\mathrm{T}} \phi(\mathbf{x})$$

$$\sigma_{t|\mathrm{t}}^2 = \mathrm{var}(\mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}) + \epsilon|\mathrm{t}) = \phi(\mathbf{x})^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{w}|\mathrm{t}} \phi(\mathbf{x}) + \frac{1}{\beta}$$

RÉGRESSION BAYÉSIENNE À NOYAU

Sujets: régression bayésienne à noyau

- On pourrait utiliser le truc du noyau pour obtenir une **version à noyau** de la régression linéaire bayésienne

$$\mu_{t|\mathrm{t}} = \phi(\mathbf{x})^{\mathrm{T}} \boldsymbol{\mu}_{\mathbf{w}|\mathrm{t}}$$

$$\sigma^2_{t|\mathrm{t}} = \phi(\mathbf{x})^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{w}|\mathrm{t}} \phi(\mathbf{x}) + \frac{1}{\beta}$$

RÉGRESSION BAYÉSIENNE À NOYAU

Sujets: régression bayésienne à noyau

- On pourrait utiliser le truc du noyau pour obtenir une **version à noyau** de la régression linéaire bayésienne

$$\boldsymbol{\mu}_{\mathbf{w}|\mathbf{t}} = \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\mathbf{t}$$

$$\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{t}} = \alpha^{-1}\mathbf{I} - \alpha^{-2}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\boldsymbol{\Phi}$$

$$\mu_{t|\mathbf{t}} = \phi(\mathbf{x})^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{w}|\mathbf{t}}$$

$$\sigma^2_{t|\mathbf{t}} = \phi(\mathbf{x})^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{t}}\phi(\mathbf{x}) + \frac{1}{\beta}$$

RÉGRESSION BAYÉSIENNE À NOYAU

Sujets: régression bayésienne à noyau

- On pourrait utiliser le truc du noyau pour obtenir une **version à noyau** de la régression linéaire bayésienne

$$\boldsymbol{\mu}_{\mathbf{w}|\mathrm{t}} = \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\mathrm{t}$$

$$\boldsymbol{\Sigma}_{\mathbf{w}|\mathrm{t}} = \alpha^{-1}\mathbf{I} - \alpha^{-2}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\boldsymbol{\Phi}$$

$$k(\mathbf{x}_n, \mathbf{x}_m) = \frac{1}{\alpha}\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_m)$$

$$\mu_{t|\mathrm{t}} = \phi(\mathbf{x})^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{w}|\mathrm{t}} = \mathbf{k}(\mathbf{x})^{\mathrm{T}}(\mathbf{K} + \tfrac{1}{\beta}\mathbf{I})^{-1}\mathrm{t}$$

$$\sigma^2_{t|\mathrm{t}} = \phi(\mathbf{x})^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{w}|\mathrm{t}}\phi(\mathbf{x}) + \tfrac{1}{\beta}$$

RÉGRESSION BAYÉSIENNE À NOYAU

Sujets: régression bayésienne à noyau

- On pourrait utiliser le truc du noyau pour obtenir une **version à noyau** de la régression linéaire bayésienne

$$\boldsymbol{\mu}_{\mathbf{w}|\mathrm{t}} = \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\mathrm{t}$$

$$\boldsymbol{\Sigma}_{\mathbf{w}|\mathrm{t}} = \alpha^{-1}\mathbf{I} - \alpha^{-2}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\boldsymbol{\Phi}$$

$$k(\mathbf{x}_n, \mathbf{x}_m) = \frac{1}{\alpha}\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_m)$$

$$\mu_{t|\mathrm{t}} = \phi(\mathbf{x})^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{w}|\mathrm{t}} = \mathbf{k}(\mathbf{x})^{\mathrm{T}}(\mathbf{K} + \tfrac{1}{\beta}\mathbf{I})^{-1}\mathrm{t}$$

$$\sigma^2_{t|\mathrm{t}} = \phi(\mathbf{x})^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{w}|\mathrm{t}}\phi(\mathbf{x}) + \tfrac{1}{\beta}$$

$$= k(\mathbf{x}, \mathbf{x}) + \tfrac{1}{\beta} - \mathbf{k}(\mathbf{x})^{\mathrm{T}}(\mathbf{K} + \tfrac{1}{\beta}\mathbf{I})^{-1}\mathbf{k}(\mathbf{x})$$

RÉGRESSION BAYÉSIENNE À NOYAU

Sujets: équivalence avec la régression normale

- La prédiction est la même que pour la régression à noyau non-bayésienne
  - dans ce cas-ci, l'apprentissage bayésien n'apporte pas d'avantages en terme de généralisation
  - par contre, il donne une estimation de la certitude sur la prédiction faite par le modèle (sa variance)

Apprentissage automatique

Apprentissage bayésien - processus gaussien

RÉGRESSION BAYÉSIENNE À NOYAU

Sujets: régression bayésienne à noyau

RAPPEL

- On pourrait utiliser le truc du noyau pour obtenir une **version à noyau** de la régression linéaire bayésienne

$$\mu_{t|\mathbf{t}} = \phi(\mathbf{x})^{\mathrm{T}} \boldsymbol{\mu}_{\mathbf{w}|\mathbf{t}} = \mathbf{k}(\mathbf{x})^{\mathrm{T}} (\mathbf{K} + \tfrac{1}{\beta}\mathbf{I})^{-1} \mathbf{t}$$

$$\sigma^2_{t|\mathbf{t}} = \phi(\mathbf{x})^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{t}} \phi(\mathbf{x}) + \tfrac{1}{\beta}$$

$$= k(\mathbf{x}, \mathbf{x}) + \tfrac{1}{\beta} - \mathbf{k}(\mathbf{x})^{\mathrm{T}} (\mathbf{K} + \tfrac{1}{\beta}\mathbf{I})^{-1} \mathbf{k}(\mathbf{x})$$

PROCESSUS GAUSSIEN

Sujets: processus gaussien

- On va dériver le même algorithme autrement, sans supposer $t = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + \epsilon$

- On va plutôt poser $t = y(\mathbf{x}) + \epsilon$ et traiter $y(\mathbf{x})$ comme une variable aléatoire
  - notre a priori sur la fonction $y(\mathbf{x})$ est qu'elle a été générée par un **processus gaussien**
  - en d'autres mots, on va utiliser un processus gaussien pour notre $p(\text{«modèle»})$

PROCESSUS GAUSSIEN

Sujets: processus gaussien

- Si $y(\mathbf{x})$ est généré d'un processus gaussien, alors la loi de probabilité de tout vecteur $\mathbf{y} = (y(\mathbf{x}_1),...,y(\mathbf{x}_N))^{\mathrm{T}}$ est

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K})$$

$$K_{nm} = k(\mathbf{x}_n, \mathbf{x}_m)$$

peu importe la valeur des $\mathbf{x}_n$ et leur nombre N

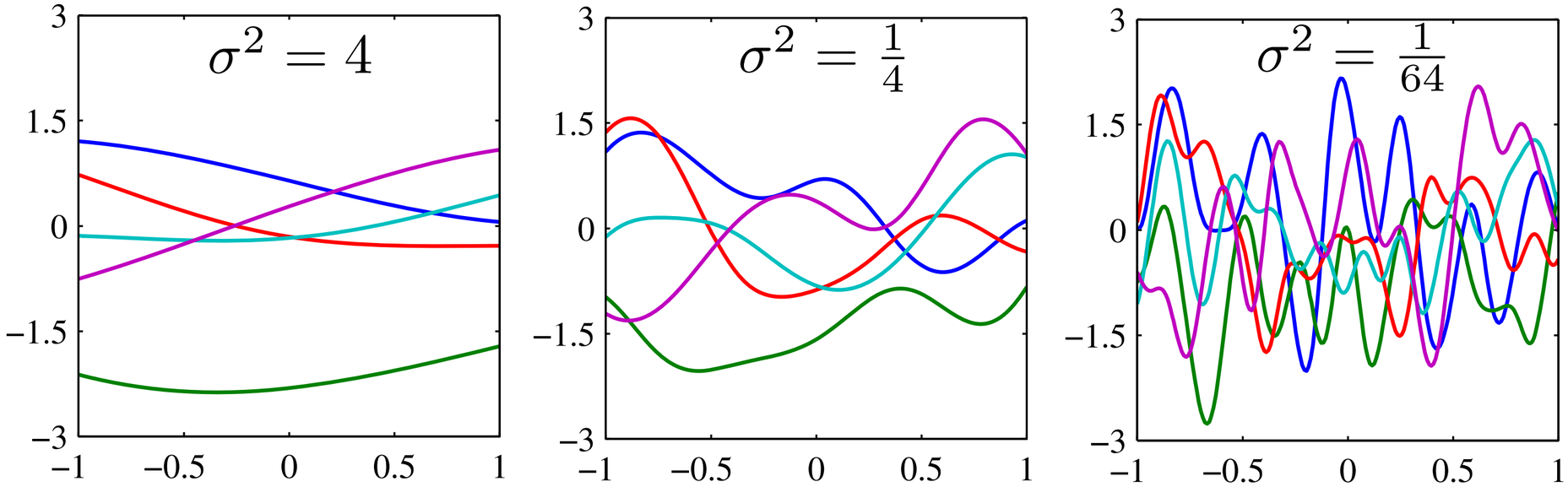
- Un processus gaussien est défini par le noyau $k(\mathbf{x}_n,\mathbf{x}_m)$
  - on l'appelle fonction de covariance : $\mathrm{cov}[y(\mathbf{x}_n),y(\mathbf{x}_m)] = k(\mathbf{x}_n,\mathbf{x}_m)$

PROCESSUS GAUSSIEN

Sujets: processus gaussien

- Exemple de fonction $y(\mathbf{x})$ générée avec un noyau gaussien

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\|\mathbf{x} - \mathbf{x}'\|^2 / 2\sigma^2\right)$$



PROCESSUS GAUSSIEN

Sujets: processus gaussien

- On suppose que $t = y(\mathbf{x}) + \epsilon$ où ϵ suit une loi $\mathcal{N}(\epsilon|0, \beta^{-1})$
- Soit $\mathbf{t}_{N+1} = (\underbrace{t_1, \ldots, t_N}_{\text{données d'entraînement}}, \overbrace{t}^{\text{nouvelle cible à prédire, pour } \mathbf{x}})^{\mathrm{T}}$
- Alors $\mathbf{t}_{N+1}$ suit une loi gaussienne

$$p(\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1})$$

où $\mathbf{C}_{N+1} = \mathbf{K}_{N+1} + \beta^{-1}\mathbf{I}$

PROCESSUS GAUSSIEN

Sujets: processus gaussien

- On suppose que $t = y(\mathbf{x}) + \epsilon$ où ϵ suit une loi $\mathcal{N}(\epsilon|0, \beta^{-1})$

$p(\text{«données»}|\text{«modèle»})$

- Soit $\mathbf{t}_{N+1} = (\underbrace{t_1,\ldots,t_N}_{\text{données d'entraînement}}, \underbrace{t}_{\text{nouvelle cible à prédire, pour } \mathbf{x}})^{\mathrm{T}}$

- Alors $\mathbf{t}_{N+1}$ suit une loi gaussienne

$$p(\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1})$$

où $\mathbf{C}_{N+1} = \mathbf{K}_{N+1} + \beta^{-1}\mathbf{I}$

PROCESSUS GAUSSIEN

Sujets: processus gaussien

- Soit la loi gaussienne

$$p(\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1})$$

- On a donc que la loi conditionnelle de t étant données $\mathbf{t}=(t_1,\ldots,t_N)$ est une gaussienne de paramètres

$$\mu_{t|\mathbf{t}} = \mathbf{k}(\mathbf{x})^{\mathrm{T}}(\mathbf{K} + \tfrac{1}{\beta}\mathbf{I})^{-1}\mathbf{t}$$

$$\sigma^2_{t|\mathbf{t}} = k(\mathbf{x},\mathbf{x}) + \tfrac{1}{\beta} - \mathbf{k}(\mathbf{x})^{\mathrm{T}}(\mathbf{K} + \tfrac{1}{\beta}\mathbf{I})^{-1}\mathbf{k}(\mathbf{x})$$

PROCESSUS GAUSSIEN

Sujets: processus gaussien

- Soit la loi gaussienne

$$\begin{pmatrix} \mathbf{K} + \frac{1}{\beta}\mathbf{I} & \mathbf{k}(\mathbf{x}) \\ \mathbf{k}(\mathbf{x})^{\mathrm{T}} & k(\mathbf{x},\mathbf{x}) + \frac{1}{\beta} \end{pmatrix}$$

$$p(\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1})$$

- On a donc que la loi conditionnelle de t étant données $\mathbf{t}=(t_1,\ldots,t_N)$ est une gaussienne de paramètres

$$\mu_{t|\mathbf{t}} = \mathbf{k}(\mathbf{x})^{\mathrm{T}}(\mathbf{K} + \tfrac{1}{\beta}\mathbf{I})^{-1}\mathbf{t}$$

$$\sigma^2_{t|\mathbf{t}} = k(\mathbf{x},\mathbf{x}) + \tfrac{1}{\beta} - \mathbf{k}(\mathbf{x})^{\mathrm{T}}(\mathbf{K} + \tfrac{1}{\beta}\mathbf{I})^{-1}\mathbf{k}(\mathbf{x})$$

PROCESSUS GAUSSIEN

Sujets: processus gaussien

- Soit la loi gaussienne

$$\begin{pmatrix} \mathbf{K} + \frac{1}{\beta}\mathbf{I} & \mathbf{k}(\mathbf{x}) \\ \mathbf{k}(\mathbf{x})^{\mathrm{T}} & k(\mathbf{x},\mathbf{x}) + \frac{1}{\beta} \end{pmatrix}$$

$$p(\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1})$$

- On a donc que la loi conditionnelle de t étant données $\mathbf{t}=(t_1,\ldots,t_N)$ est une gaussienne de paramètres

$$\mu_{t|\mathrm{t}} = \mathbf{k}(\mathbf{x})^{\mathrm{T}}(\mathbf{K} + \tfrac{1}{\beta}\mathbf{I})^{-1}\mathrm{t}$$

$$\sigma^2_{t|\mathrm{t}} = k(\mathbf{x},\mathbf{x}) + \tfrac{1}{\beta} - \mathbf{k}(\mathbf{x})^{\mathrm{T}}(\mathbf{K} + \tfrac{1}{\beta}\mathbf{I})^{-1}\mathbf{k}(\mathbf{x})$$

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \end{aligned}$$

Apprentissage automatique

Apprentissage bayésien - résumé

RÉGRESSION LINÉAIRE BAYÉSIENNE

Sujets: résumé de la régression linéaire bayésienne

- Modèle : $y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x})$

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}) \qquad p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

$p(\text{«données»}|\text{«modèle»})$ $\qquad$ $p(\text{«modèle»})$

- Entraînement : inférence de $p(\mathbf{w}|\text{«données»})$

$$\boldsymbol{\mu}_{\mathbf{w}|\mathrm{t}} = \alpha^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\mathrm{t}$$

$$\boldsymbol{\Sigma}_{\mathbf{w}|\mathrm{t}} = \alpha^{-1}\mathbf{I} - \alpha^{-2}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I})^{-1}\boldsymbol{\Phi}$$

- Hyper-paramètres : α, β
- Prédiction : $\phi(\mathbf{x})^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{w}|\mathrm{t}}$ (variance : $\phi(\mathbf{x})^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{w}|\mathrm{t}}\phi(\mathbf{x}) + \frac{1}{\beta}$)

PROCESSUS GAUSSIEN

Sujets: résumé de la régression avec processus gaussien

- Modèle : $t = y(\mathbf{x}) + \epsilon$

$$p(t|\mathbf{x}, \beta) = \mathcal{N}(t|y(\mathbf{x}), \beta^{-1})$$

$p(\text{«données»}|\text{«modèle»})$

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K})$$

$p(\text{«modèle»})$

- Entraînement : calcul de $\mathbf{K}$
- Hyper-paramètres : β et ceux dans le noyau $k(\mathbf{x}_n, \mathbf{x}_m)$
- Prédiction : $\mathbf{k}(\mathbf{x})^{\mathrm{T}}(\mathbf{K} + \frac{1}{\beta}\mathbf{I})^{-1}\mathbf{t}$

  (variance : $k(\mathbf{x}, \mathbf{x}) + \frac{1}{\beta} - \mathbf{k}(\mathbf{x})^{\mathrm{T}}(\mathbf{K} + \frac{1}{\beta}\mathbf{I})^{-1}\mathbf{k}(\mathbf{x})$)

CHOIX DE LOI A PRIORI

Sujets: choix de loi a priori

- Si on ne connaît rien du problème à résoudre, il est préférable de choisir une loi a priori à haute entropie (*flat prior*)
  - on peut aussi le traiter comme un hyper-paramètre et faire de la sélection de modèle

- Sinon, il sera avantageux d'incorporer dans la loi a priori, toute information sur la solution
  - par contre, si l'information incorporée n'est pas juste, on risque d'en payer le prix avec une réduction de la performance

EXTENSIONS

Sujets: extension de l'apprentissage bayésien

- L'apprentissage bayésien est un principe applicable à tout modèle probabiliste
  - voir section 4.5 : régression logistique bayésienne
- On peut faire de la classification avec les processus gaussiens
  - voir section 6.4.5
- On peut optimiser les hyper-paramètres sans ensemble de validation
  - voir section 6.4.3