Apprentissage automatique

Régression linéaire - modèle

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

MODÈLE DE RÉGRESSION LINÉAIRE

Sujets: modèle, biais, poids

- Le modèle de **régression linéaire** est le suivant :

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

où $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$

- La prédiction correspond donc à un hyperplan de dimension D (donc une droite si $D{=}1$)

MODÈLE DE RÉGRESSION LINÉAIRE

Sujets: modèle, biais, poids

- Le modèle de **régression linéaire** est le suivant :

biais

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

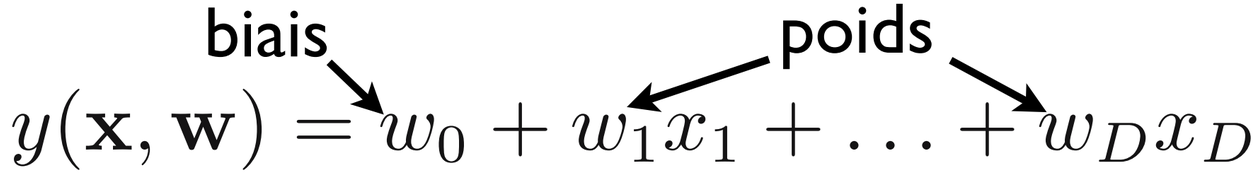
où $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$

- La prédiction correspond donc à un hyperplan de dimension D (donc une droite si $D{=}1$)

MODÈLE DE RÉGRESSION LINÉAIRE

Sujets: modèle, biais, poids

- Le modèle de **régression linéaire** est le suivant :

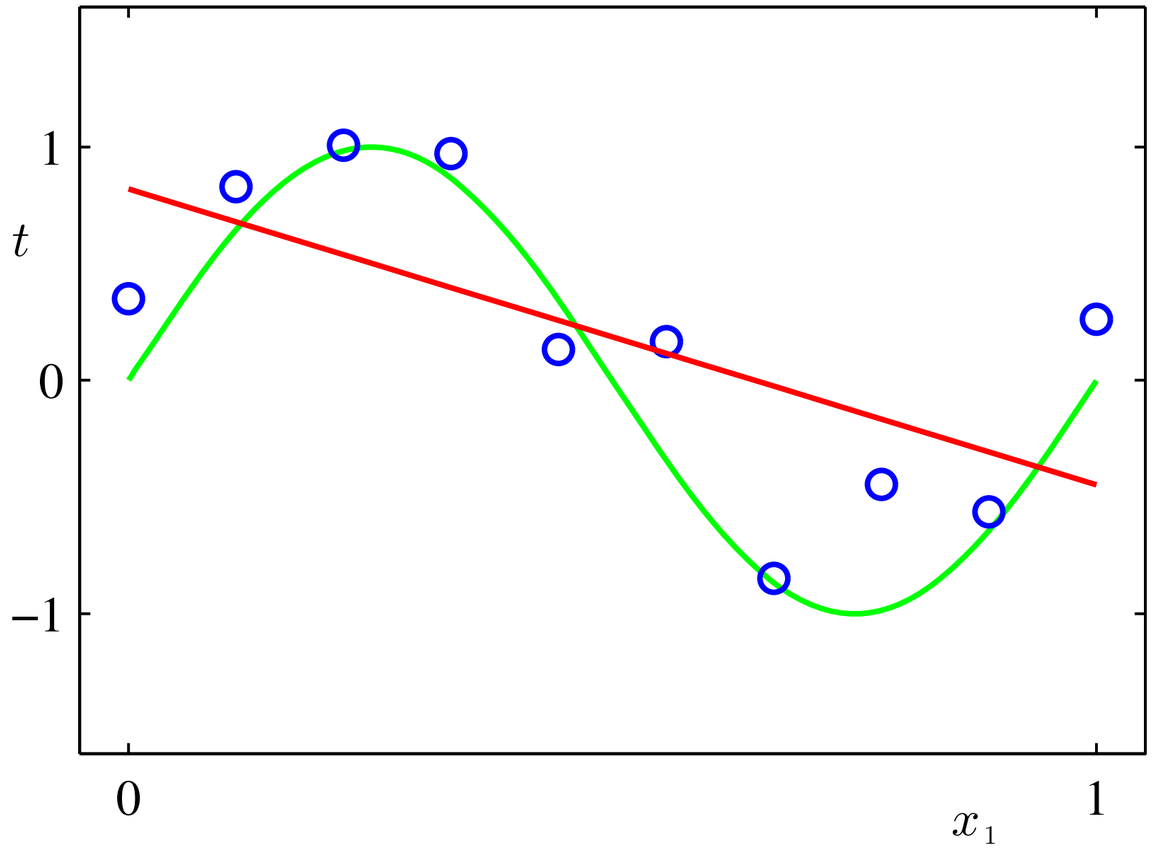


où $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$

- La prédiction correspond donc à un hyperplan de dimension D (donc une droite si $D{=}1$)

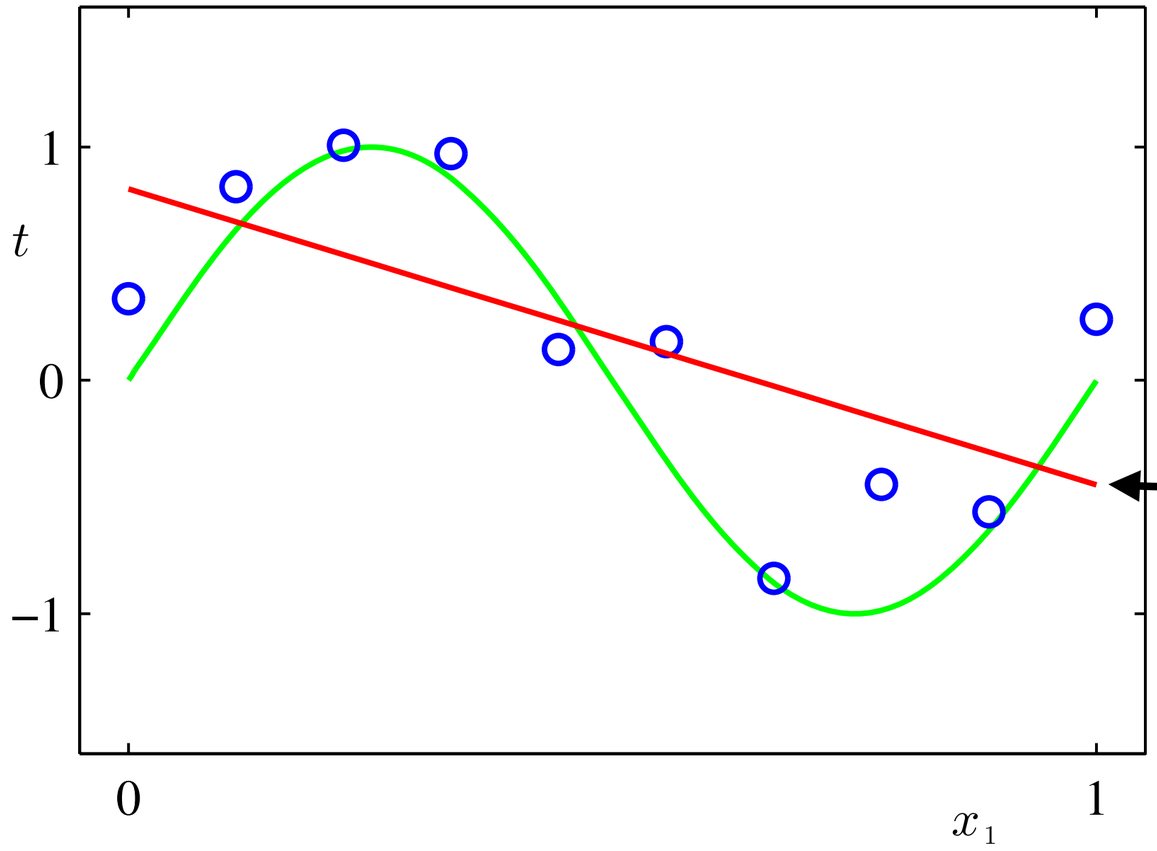
MODÈLE DE RÉGRESSION LINÉAIRE

Sujets: exemple $1D$



MODÈLE DE RÉGRESSION LINÉAIRE

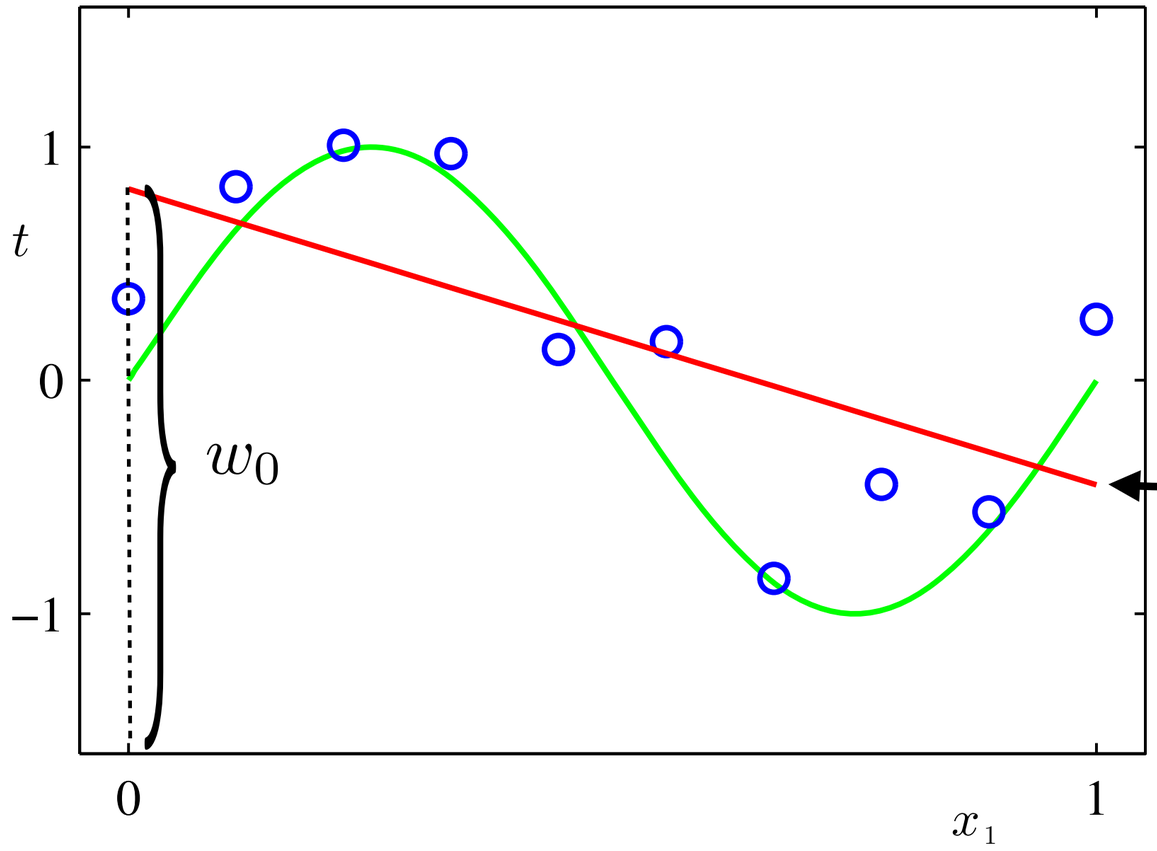
Sujets: exemple $1D$



$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1$$

MODÈLE DE RÉGRESSION LINÉAIRE

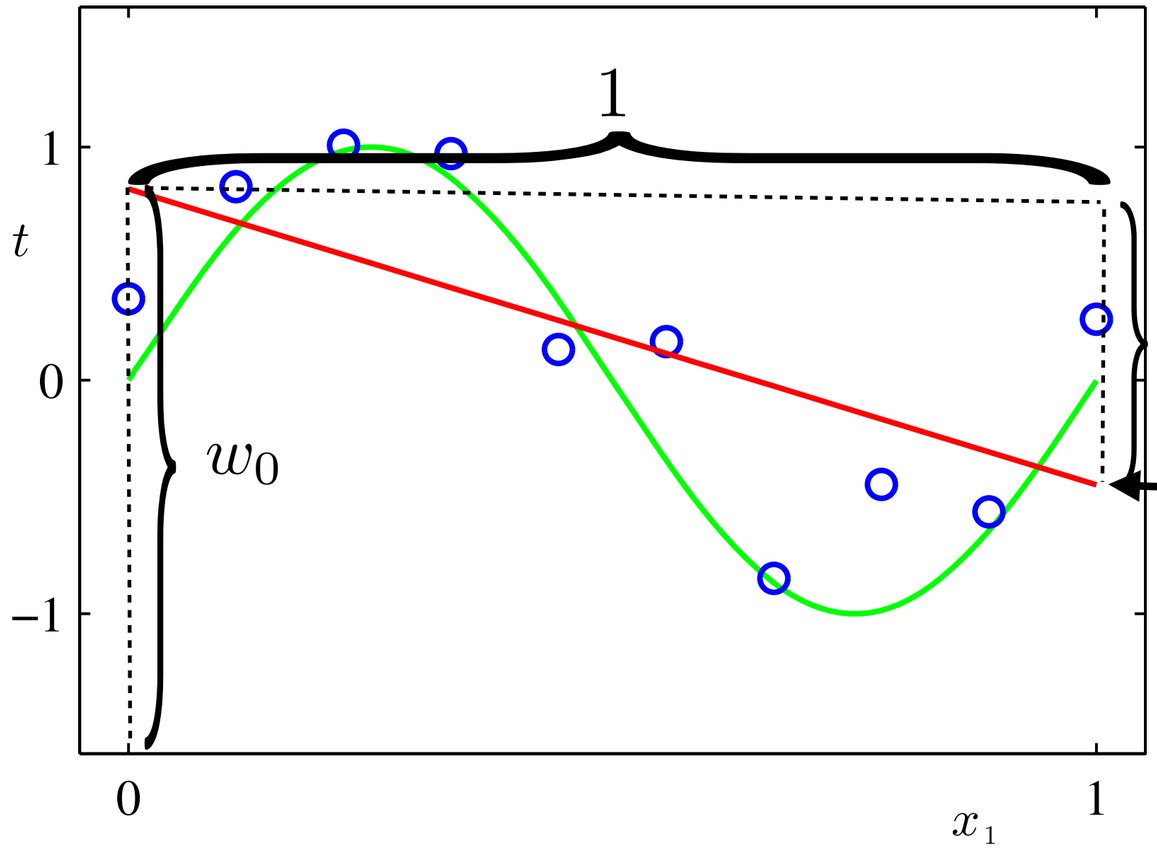
Sujets: exemple $1D$



$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1$$

MODÈLE DE RÉGRESSION LINÉAIRE

Sujets: exemple $1D$



$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1$$

Apprentissage automatique

Régression linéaire - fonctions de base (*basis functions*)

MODÈLE DE RÉGRESSION LINÉAIRE

Sujets: modèle

- Le modèle de **régression linéaire** est le suivant :

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

où $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$

- La prédiction correspond donc à un hyperplan de dimension D
  - un hyperplan peut ne pas être assez flexible pour faire une bonne prédiction

FONCTION DE BASE

Sujets: fonctions de base (*basis functions*)

- On peut introduire une non-linéarité comme suit :

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

où les $\phi_j(\mathbf{x})$ sont des **fonctions de base** (*basis functions*)

- Cas linéaire : $\phi_j(\mathbf{x}) = x_j$ et $M = D + 1$

FONCTION DE BASE

Sujets: fonctions de base (*basis functions*)

- Pour simplifier la notation, on va supposer que $\phi_0(\mathbf{x}) = 1$

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

FONCTION DE BASE

Sujets: fonctions de base (*basis functions*)

- Pour simplifier la notation, on va supposer que $\phi_0(\mathbf{x}) = 1$

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

$$(w_0, \ldots, w_{M-1})^{\mathrm{T}}$$

FONCTION DE BASE

Sujets: fonctions de base (*basis functions*)

- Pour simplifier la notation, on va supposer que $\phi_0(\mathbf{x}) = 1$

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

$\mathbf{w}$: $(w_0, \ldots, w_{M-1})^{\mathrm{T}}$

$\boldsymbol{\phi}(\mathbf{x})$: $(\phi_0(\mathbf{x}), \ldots, \phi_{M-1}(\mathbf{x}))^{\mathrm{T}}$

FONCTION DE BASE

Sujets: fonctions de base polynomiales

- Exemple : fonctions de bases polynomiales ($1D$)

$$\phi_j(x) = x^j$$

- On retrouve alors la régression polynomiale

FONCTION DE BASE

Sujets: fonctions de base gaussiennes

- Exemple : fonctions de base gaussiennes

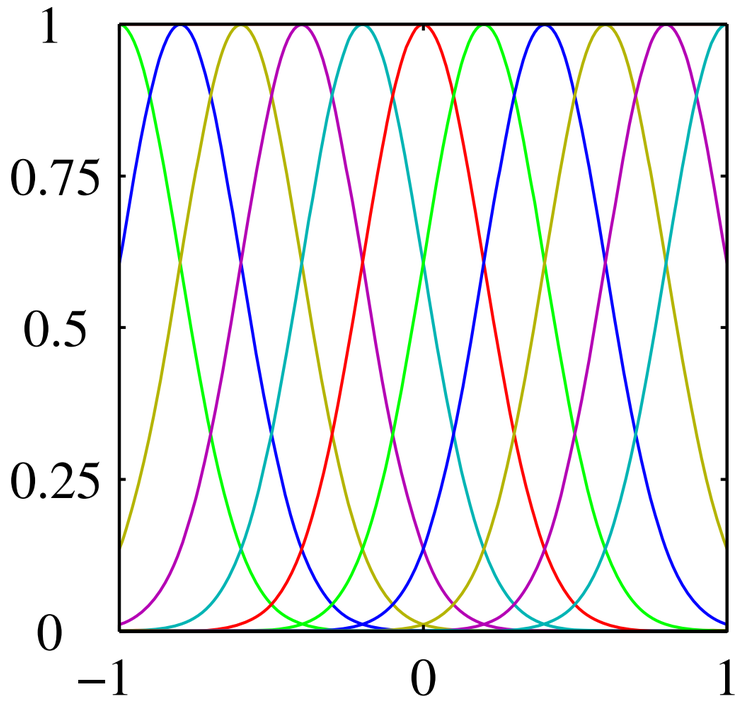
$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$

où μ_j et s doivent être spécifiés

FONCTION DE BASE

Sujets: fonctions de base gaussiennes

- Exemple : fonctions de base gaussiennes
  - exemple en $1D$



Apprentissage automatique

Régression linéaire - maximum de vraisemblance

MAXIMUM DE VRAISEMBLANCE

Sujets: formulation probabiliste

- Pour entraîner le modèle $y(\mathbf{x}, \mathbf{w})$, nous passerons par une formulation probabiliste :

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

  - équivaut à supposer que les cibles sont une version bruitée de la prédiction du vrai modèle

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon$$

Gaussienne de moyenne 0 et variance β^{-1}

MAXIMUM DE VRAISEMBLANCE

Sujets: formulation probabiliste

- Soit notre ensemble d'entraînement
  $\mathcal{D} = \{(\mathbf{x}_1, t_1), \ \ldots \ , \ (\mathbf{x}_N, t_N)\}$
  - on va également noter $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
    et $\mathbf{t} = (\ t_1, \ldots, t_N\ )^{\mathrm{T}}$

- En faisant l'hypothèse i.i.d., on a :

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

MAXIMUM DE VRAISEMBLANCE

Sujets: maximum de vraisemblance

- Lors de l'entraînement, on cherche le $\mathbf{w}$ maximisant la (log-)probabilité des données d'entraînement

$$\begin{aligned}\ln p(\mathbf{t}|\mathbf{w},\beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}) \\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})\end{aligned}$$

où

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

MAXIMUM DE VRAISEMBLANCE

Sujets: maximum de vraisemblance

- On sait que le gradient de la somme des pertes

$$\nabla E_D(\mathbf{w}) = -\sum_{n=1}^{N} \{t_n - \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n)\} \, \phi(\mathbf{x}_n)^{\mathrm{T}}$$

à la valeur $\mathbf{w}$ minimisante doit être égale à 0 :

$$0 = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} - \mathbf{w}^{\mathrm{T}} \left(\sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n) \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \right)$$

MAXIMUM DE VRAISEMBLANCE

Sujets: maximum de vraisemblance, *design matrix*

- En isolant $\mathbf{w}$, on trouve que le $\mathbf{w}$ minimisant la somme des pertes (maximisant la log-probabilité) est

$$\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$$

où $\mathbf{\Phi}$ est appelée la *design matrix* :

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

MAXIMUM DE VRAISEMBLANCE

Sujets: maximum de vraisemblance, *design matrix*

- En isolant $\mathbf{w}$, on trouve que le $\mathbf{w}$ minimisant la somme des pertes (maximisant la log-probabilité) est

$$\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$$

- Il faudrait aussi vérifier qu'il s'agit bel et bien d'un minimum de $E_D(\mathbf{w})$ (et non un maximum ou un pointselle)
  - se fait en calculant les dérivées secondes
  - plus précisément, on montre que la matrice des dérivées secondes (matrice hessienne) est définie positive

Apprentissage automatique

Régression linéaire - régularisation

MAXIMUM DE VRAISEMBLANCE

Sujets: maximum de vraisemblance

- Maximiser la log-probabilité

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

$$= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})$$

équivaut à minimiser la somme des pertes de l'erreur au carré (*squared error*) :

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

RÉGULARISATION

Sujets: régularisation, *weight decay*, régression de Ridge

- Afin de contrôler les risques de sur-apprentissage, on préfère ajouter un terme de régularisation

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

  - équivaut au maximum a posteriori dans la formulation probabiliste
  - le terme de régularisation est souvent appelé ***weight decay***
  - la régression avec un terme de régularisation est aussi appelée **régression de Ridge**

RÉGULARISATION

Sujets: régularisation, *weight decay*, régression de Ridge

- Afin de contrôler les risques de sur-apprentissage, on préfère ajouter un terme de régularisation

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\underbrace{\mathbf{w}^{\mathrm{T}}\mathbf{w}}_{\|\mathbf{w}\|^2}$$

  - équivaut au maximum a posteriori dans la formulation probabiliste
  - le terme de régularisation est souvent appelé ***weight decay***
  - la régression avec un terme de régularisation est aussi appelée **régression de Ridge**

RÉGULARISATION

Sujets: régularisation, *weight decay*, régression de Ridge

- On peut montrer que la solution (maximum a posteriori) est alors :

$$\mathbf{w} = \left(\lambda \mathbf{I} + \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^{\mathrm{T}} \mathbf{t}$$

  - dans le cas $\lambda = 0$, on retrouve la solution du maximum de vraisemblance
  - si $\lambda > 0$, permet également d'avoir une solution plus stable numériquement (si $\mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi}$ n'est pas inversible)

Apprentissage automatique

Régression linéaire - prédictions multiples

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 vecteur réel $\mathbf{t} \in \mathbb{R}^K$
    - exemple : prédiction de la valeur d'une action à la bourse
      - ✓ $\mathbf{x}$: vecteur contenant l'information sur l'activité économique de la journée
      - ✓ $\mathbf{t}$: la valeur de plusieurs actions le lendemain

PRÉDICTIONS MULTIPLES

Sujets: modèle pour prédictions multiples

- Le modèle doit maintenant prédire un vecteur :

$$\mathbf{y}(\mathbf{x}, \mathbf{w}) = \mathbf{W}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

où $\mathbf{W}$ est une matrice $M \times K$

- Chaque colonne de $\mathbf{W}$ peut être vue comme le vecteur $\mathbf{w}_k$ du modèle $y(\mathbf{x}, \mathbf{w}_k)$ pour la k^{e} cible

PRÉDICTIONS MULTIPLES

Sujets: formulation probabiliste pour prédictions multiples, modèle multitâche

- On suppose encore un modèle gaussien

$$p(\mathbf{t}|\mathbf{x}, \mathbf{W}, \beta) = \mathcal{N}(\mathbf{t}|\mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\mathbf{I})$$

où on suppose que les cibles sont indépendantes

- Un modèle faisant des prédictions multiples est parfois appelé un **modèle multitâche**

PRÉDICTIONS MULTIPLES

Sujets: formulation probabiliste pour prédictions multiples

- Soit notre ensemble d'entraînement
  $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{t}_1), \ \ldots \ , \ (\mathbf{x}_N, \mathbf{t}_N)\}$
  - on va également noter $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
    et $\mathbf{T}$ est une matrice dont les rangées sont les vecteurs $\mathbf{t}_1, \ \ldots \ , \ \mathbf{t}_N$

- En faisant l'hypothèse i.i.d., on a :

$$\ln p(\mathbf{T}|\mathbf{X}, \mathbf{W}, \beta) = \sum_{n=1}^{N} \ln \mathcal{N}(\mathbf{t}_n | \mathbf{W}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\mathbf{I})$$

$$= \frac{NK}{2} \ln \left(\frac{\beta}{2\pi} \right) - \frac{\beta}{2} \sum_{n=1}^{N} \left\| \mathbf{t}_n - \mathbf{W}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\|^2$$

MAXIMUM DE VRAISEMBLANCE

Sujets: formulation probabiliste pour prédictions multiples

- On peut démontrer que le maximum de vraisemblance est :

$$\mathbf{W}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{T}$$

- On peut voir le résultat comme la concaténation (colonne par colonne) des solutions pour chaque tâche

$$\mathbf{w}_k = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}_k$$

où $\mathbf{t}_k = (t_{1,k}, \ldots, t_{N,k})^{\mathrm{T}}$

Apprentissage automatique

Régression linéaire - théorie de la décision

MODÈLE DE RÉGRESSION LINÉAIRE

Sujets: formulation probabiliste

RAPPEL

- En régression linéaire, on suppose que :

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

où

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

- Lorsqu'on doit faire une prédiction pour une nouvelle entrée $\mathbf{x}$, on prédit alors la moyenne, i.e. $y(\mathbf{x},\mathbf{w})$

THÉORIE DE LA DÉCISION

Sujets: théorie de la décision

- Pourquoi est-ce que prédire la moyenne ($y(\mathbf{x}, \mathbf{w})$) est la bonne chose à faire ?

- La **théorie de la décision** va nous éclairer sur le sujet

- On va maintenant noter $\hat{y}(\mathbf{x})$ la prédiction (**décision**) que l'on va faire pour une entrée $\mathbf{x}$

  - $\hat{y}(\mathbf{x})$ pourrait être différente de $y(\mathbf{x}, \mathbf{w})$

THÉORIE DE LA DÉCISION

Sujets: théorie de la décision

- Sachant que :
  - chaque paire $(\mathbf{x}, t)$ est échantillonnée d'une distribution $p(\mathbf{x}, t)$
  - la perte qui nous intéresse est $L(t, \hat{y}(\mathbf{x})) = \{\hat{y}(\mathbf{x})\text{-}t\}^2$
- La perte espérée de notre prédiction (décision) sera

$$\mathbb{E}[L] = \iint \{\hat{y}(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t$$

THÉORIE DE LA DÉCISION

Sujets: théorie de la décision

- Pour trouver $\hat{y}(\mathbf{x})$ optimale on commence par noter que :

$$\begin{aligned}\{\hat{y}(\mathbf{x}) - t\}^2 &= \{\hat{y}(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\ &= \{\hat{y}(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{\hat{y}(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2\end{aligned}$$

- Ainsi : $\int\int \{\hat{y}(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x}\mathrm{d}t =$

$$\int\int \{\hat{y}(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x}\mathrm{d}t + \int\int \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x}\mathrm{d}t$$

$$+ \int\int 2\{\hat{y}(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} p(\mathbf{x}, t) \mathrm{d}\mathbf{x}\mathrm{d}t$$

THÉORIE DE LA DÉCISION

Sujets: théorie de la décision

- Ensuite, on remarque que :

$$\int\int 2\{\hat{y}(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\}p(\mathbf{x}, t)\mathrm{d}\mathbf{x}\mathrm{d}t$$

$$= \int\int 2\{\hat{y}(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\}p(t|\mathbf{x})p(\mathbf{x})\mathrm{d}t\mathrm{d}\mathbf{x}$$

$$= \int 2\{\hat{y}(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}p(\mathbf{x})\left(\int \{\mathbb{E}[t|\mathbf{x}] - t\}p(t|\mathbf{x})\mathrm{d}t\right)\mathrm{d}\mathbf{x}$$

$$= \int 2\{\hat{y}(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}p(\mathbf{x})\left(\mathbb{E}[t|\mathbf{x}] - \int t\ p(t|\mathbf{x})\mathrm{d}t\right)\mathrm{d}\mathbf{x} = 0$$

THÉORIE DE LA DÉCISION

Sujets: théorie de la décision

- Donc on a que :

$$\int\int \{\hat{y}(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x}\mathrm{d}t =$$

$$\int\int \{\hat{y}(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x}\mathrm{d}t + \int\int \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x}\mathrm{d}t$$

$$\cancel{+ \int\int 2\{\hat{y}(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} p(\mathbf{x}, t) \mathrm{d}\mathbf{x}\mathrm{d}t}$$

- Le mininum est donc atteint lorsque $\hat{y}(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}]$

THÉORIE DE LA DÉCISION

Sujets: théorie de la décision

- Puisqu'on ne connaît pas la vraie distribution $p(\mathbf{x}, t)$
  (et donc ni $p(t|\mathbf{x})$, ni $\mathbb{E}[t|\mathbf{x}]$), le mieux qu'on puisse faire
  est d'utiliser notre modèle de $p(t|\mathbf{x})$

$$\hat{y}(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = y(\mathbf{x}, \mathbf{w})$$

- Donc, si on veut une petite perte de la différence au carré (en espérance), prédire la moyenne est un bon choix selon la théorie de la décision

THÉORIE DE LA DÉCISION

Sujets: théorie de la décision

- Pour d'autres choix de perte, la décision optimale sera différente
  - pour la perte $L(t,\hat{y}(\mathbf{x})) = |\hat{y}(\mathbf{x})\text{-}t|$, la décision $\hat{y}(\mathbf{x})$ devrait être la médiane de $p(t|\mathbf{x},\mathbf{w})$
- Par chance, la médiane d'une gaussienne est aussi la moyenne!
  - pour d'autres choix de modèle probabiliste $p(t|\mathbf{x},\mathbf{w})$, ce pourrait ne pas être le cas

Apprentissage automatique

Régression linéaire - décomposition biais-variance

GÉNÉRALISATION

Sujets: généralisation

- Analysons la performance espérée de généralisation d'un modèle donné $y(\mathbf{x}, \mathbf{w})$ de régression :

$$\mathbb{E}_{(\mathbf{x},t)}[L(t, y(\mathbf{x}, \mathbf{w}))] = \int \int \{y(\mathbf{x}, \mathbf{w}) - t\}^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x} \mathrm{d}t$$

  où $p(\mathbf{x}, t)$ est la vraie distribution des exemples $(\mathbf{x}, t)$

- Changement à la notation : on note explicitement selon quelles variables aléatoires on fait l'espérance, en ajoutant l'information en indice

GÉNÉRALISATION

Sujets: généralisation

- Analysons la performance espérée de généralisation d'un modèle donné $y(\mathbf{x}, \mathbf{w})$ de régression :

$$\mathbb{E}_{(\mathbf{x},t)}[L(t, y(\mathbf{x}, \mathbf{w}))] = \int \int \{y(\mathbf{x}, \mathbf{w}) - t\}^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x} \mathrm{d}t$$

où $p(\mathbf{x},t)$ est la vraie distribution des exemples $(\mathbf{x},t)$

- Changement à la notation : on note explicitement selon quelles variables aléatoires on fait l'espérance, en ajoutant l'information en indice

GÉNÉRALISATION

Sujets: généralisation

- Analysons la performance espérée de généralisation d'un modèle donné $y(\mathbf{x}, \mathbf{w})$ de régression :

$$\mathbb{E}_{(\mathbf{x},t)}[L(t, y(\mathbf{x}, \mathbf{w}))] = \int \int \{y(\mathbf{x}, \mathbf{w}) - t\}^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x} \mathrm{d}t$$

  où $p(\mathbf{x}, t)$ est la vraie distribution des exemples $(\mathbf{x}, t)$

- On s'intéresse plus spécifiquement au cas où le modèle $\mathbf{w}$ est celui obtenu après s'être entraîné sur $\mathcal{D}$
  - en régression linéaire, lorsque $\mathbf{w} = \left(\lambda \mathbf{I} + \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^{\mathrm{T}} \mathbf{t}$

GÉNÉRALISATION

Sujets: généralisation

- Analysons la performance espérée de généralisation d'un modèle donné $y(\mathbf{x}, \mathbf{w})$ de régression :

$$\mathbb{E}_{(\mathbf{x},t)}[L(t, y(\mathbf{x}; \mathcal{D}))] = \int \int \{y(\mathbf{x}; \mathcal{D}) - t\}^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x} \mathrm{d}t$$

où $p(\mathbf{x}, t)$ est la vraie distribution des exemples $(\mathbf{x}, t)$

- Pour traiter le cas général, on va plutôt noter notre modèle $y(\mathbf{x}; \mathcal{D})$
  - c'est le modèle obtenu après avoir entraîné sur $\mathcal{D}$

GÉNÉRALISATION

Sujets: généralisation

- Analysons la performance espérée de généralisation d'un modèle donné $y(\mathbf{x}, \mathbf{w})$ de régression :

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{(\mathbf{x},t)}[L(t, y(\mathbf{x}; \mathcal{D}))]\right] = \mathbb{E}_{\mathcal{D}}\left[\int\int \{y(\mathbf{x}; \mathcal{D}) - t\}^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x}\mathrm{d}t\right]$$

où $p(\mathbf{x}, t)$ est la vraie distribution des exemples $(\mathbf{x}, t)$

- Dans ce qui suit, on va noter $h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}]$ le meilleur modèle possible, i.e. celui qu'on cherche (voir diapositives sur la théorie de la décision)

DÉCOMPOSITION BIAIS-VARIANCE

Sujets: biais, variance, bruit

- On peut montrer que :

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{(\mathbf{x},t)}[L(t, y(\mathbf{x};\mathcal{D}))]\right] = \ (\text{bias})^2 + \text{variance} + \text{noise}$$

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{noise} = \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

DÉCOMPOSITION BIAIS-VARIANCE

Sujets: biais

- On peut montrer que :

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{(\mathbf{x},t)}[L(t, y(\mathbf{x};\mathcal{D}))]\right] = (\text{bias})^2 + \text{variance} + \text{noise}$$

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{noise} = \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

À quel point le modèle «moyen» donné par l'algorithme d'apprentissage sera proche du meilleur modèle possible $h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}]$, i.e. est-ce que l'algorithme d'apprentissage est capable de modéliser $h(\mathbf{x})$

DÉCOMPOSITION BIAIS-VARIANCE

Sujets: variance

- On peut montrer que :

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{(\mathbf{x},t)}[L(t, y(\mathbf{x};\mathcal{D}))]\right] = (\text{bias})^2 + \text{variance} + \text{noise}$$

À quel point le modèle donné par l'algorithme d'apprentissage varie d'un ensemble d'entraînement à l'autre

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{noise} = \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

DÉCOMPOSITION BIAIS-VARIANCE

Sujets: bruit

- On peut montrer que :

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{(\mathbf{x},t)}[L(t, y(\mathbf{x};\mathcal{D}))]\right] = (\text{bias})^2 + \text{variance} + \text{noise}$$

À quel point il y a du bruit dans la cible à prédire, i.e.
à que point elle varie autour de son espérance conditionnelle
(ne dépend pas de l'algorithme d'apprentissage)

$$\text{noise} = \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t$$

DÉCOMPOSITION BIAIS-VARIANCE

Sujets: bruit

- On bon algorithme d'apprentissage aura un bon compromis entre son biais et sa variance
  - si la capacité augmente, le biais diminue et la variance augmente
  - si on régularise, la variance diminue, mais le biais augmente
- La décomposition biais-variance illustre donc plus formellement les phénomène de sur-apprentissage et sous-appprentissage
  - pas assez de capacité ⇒ biais très élevé ⇒ mauvaise généralisation
  - trop de capacité ⇒ variance très élevée ⇒ mauvaise généralisation

Apprentissage automatique

Régression linéaire - résumé

RÉGRESSION LINÉAIRE

Sujets: résumé de la régression linéaire

- Modèle : $y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$

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

- Entraînement : $\mathbf{w} = \left(\lambda \mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}$
  (maximum de vraisemblance si $\lambda=0$ ou maximum a posteriori si $\lambda>0$)
- Hyper-paramètre : λ
- Prédiction : $y(\mathbf{x},\mathbf{w})$