Régression linéaire

R. Flamary, S. Canu, A. Rakotomamonjy

4 mars 2015

Sommaire

Prédiction linéaire

Fonction linéaire

Fonction $f : \mathbb{R}^d \to \mathbb{R}$, de la forme

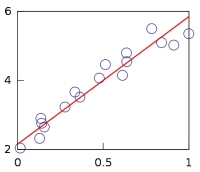
$$f(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i + b = \mathbf{x}^\top \mathbf{w} + b = [\mathbf{x}^\top 1]\boldsymbol{\alpha} \tag{1}$$

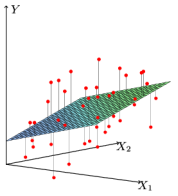
avec $\mathbf{w} \in \mathbb{R}^d$ un vecteur définissant un hyperplan dans $\mathbb{R}^d$ et $b \in \mathbb{R}$ un biais qui déplace la fonction perpendiculairement à l'hyperplan, et $\boldsymbol{\alpha} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$ est un vecteur de $\mathbb{R}^{d+1}$ contenant la concaténation de $\mathbf{w}$ et b.

Objectifs

- Régression : $f(\cdot) \in \mathbb{R}$.
- Classification : signe$(f(\cdot)) \in \{-1, 1\}$.

Régression linéaire





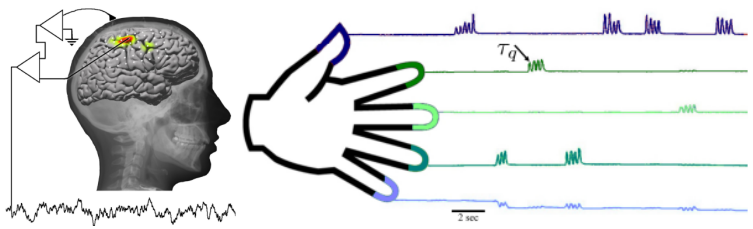
Objectif

Apprendre une fonction linéaire $f(\cdot)$ permettant de prédire une valeur réelle $y \in \mathbb{R}$ à partir d'une observation $\mathbf{x} \in \mathbb{R}^d$.
En pratique on cherche à déterminer les coefficients $(\mathbf{w}, b)$ de $f(\cdot)$ à partir d'un ensemble d'apprentissage $\{\mathbf{x}_i, y_i\}_{i=1,\dots,n}$.

Exemples

- Prédiction du taux d'insuline à partir de l'IMC et de l'épaisseur de la peau sur le triceps (Données PIMA).
- Prédiction du nombre de chenilles processionnaires par arbre, à partir de l'altitude, de la pente, du diamètre de l'arbre ...

Exemple d'application Interface Cerveau-Machine



BCI Competition IV, Données 4

- Données : enregistrement de signaux cérébraux ECoG et de la position des doigts du sujet avec un gant.
- But de la compétition : prédire les mouvement des 5 doigts de la main du sujet à partir de signaux ECoG.
- Les meilleures performances ont été obtenues avec des méthodes de régression linéaire.

Représentation des données

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top & 1 \\ \mathbf{x}_2^\top & 1 \\ \vdots & \vdots \\ \mathbf{x}_i^\top & 1 \\ \vdots & \vdots \\ \mathbf{x}_n^\top & 1 \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1d} & 1 \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2d} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{id} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{nd} & 1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}$$

Données d'apprentissage

Exemples :

- $\mathbf{x}_i \in \mathbb{R}^d$ observations pour $i = 1, \dots, n$.
- $y_i \in \mathbb{R}$ valeur à prédire $i = 1, \dots, n$.

Forme matricielle :

- $\mathbf{X} \in \mathbb{R}^{n \times (d+1)}$ telle que $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{e}]^\top$ avec $\mathbf{e} \in \mathbb{R}^d$ et $e_i = 1, \forall i$
- $\mathbf{y} \in \mathbb{R}^n$ telle que $\mathbf{y} = [y_1, y_2, \dots, y_n]^\top$.
- $\boldsymbol{\alpha} \in \mathbb{R}^{d+1}$est un vecteur de tel que $\boldsymbol{\alpha} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$

Minimisation de l'erreur de prédiction

Principe

On a le modèle suivant :

$$y = \mathbf{x}^\top \mathbf{w} + b \tag{2}$$

Et on cherche à estimer les paramètres $(\mathbf{w}, b)$ de la fonction $f(\cdot)$ à partir des données d'apprentissage.

Méthode des moindres carrés

On cherche la fonction $f(\cdot)$ minimisant l'erreur de prédiction au carré sur les exemples d'apprentissage :

$$\min_f \quad \frac{1}{2} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 \tag{3}$$

en utilisant la forme de $f(\cdot)$, on obtient le problème d'optimisation suivant :

$$\min_{\mathbf{w},b} \quad \frac{1}{2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w} - b)^2 \tag{4}$$

Interprétation des moindres carrés

$$J(\mathbf{w}, b) = \frac{1}{2} \sum_{i=1}^{n} \underbrace{(y_i - \mathbf{x}_i^T \mathbf{w} - b)}_{\varepsilon_i}^2$$

Ce problème peut s'interpréter comme la recherche de l'hyperplan $\mathbf{x}^T \mathbf{w} + b = y$ passant "au mieux" (au sens des moindres carrés) parmi le nuage des observations $(\mathbf{x}_i, y_i)$ $i = 1, n$.

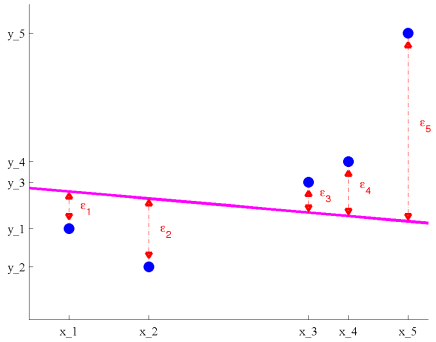


Figure : Les résidus de la régression

Résidus et moindres carrés

Résidu

Le résidu est l'erreur de prédiction :

$$\epsilon_i = y_i - \mathbf{x}_i^\top \mathbf{w} - b \tag{5}$$

Tout comme pour les exemples d'apprentissage, il peut se mettre sous forme matricielle $\boldsymbol{\epsilon} \in \mathbb{R}^n$ tel que :

$$\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\alpha} \tag{6}$$

Moindres carrés

Le problème des moindres carrés revient finalement à minimiser :

$$\min_{\alpha} \quad \|\boldsymbol{\epsilon}\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|^2 \tag{7}$$

avec $\|\cdot\|$ la norme euclidienne d'un vecteur telle que $\|\boldsymbol{\epsilon}\|^2 = \sum_{i=1}^n \epsilon_i^2$

Rappels d'optimisation

Problème à optimiser

Nous voulons résoudre le problème

$$\min_{\alpha} \quad J(\boldsymbol{\alpha}) \qquad \text{avec} \quad J(\boldsymbol{\alpha}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}||^2$$

Nous supposons pour le moment que la fonction $J(\boldsymbol{\alpha})$ est convexe.

Minimum d'une fonction convexe

Soit $J(\boldsymbol{\alpha})$ une fonction convexe de $\mathbb{R}^d$ dans $\mathbb{R}$. $\alpha^\star$ est un minimum de la fonction $J(\boldsymbol{\alpha})$ si et seulement si

$$\nabla J(\boldsymbol{\alpha}^\star) = \mathbf{0} \tag{8}$$

où $\nabla J(\boldsymbol{\alpha}) \in \mathbb{R}^d$ est le gradient de la fonction en $\boldsymbol{\alpha}$ tel que

$$\nabla J(\boldsymbol{\alpha})_i = \frac{\partial J(\boldsymbol{\alpha})}{\partial \alpha_i} \quad \forall i$$

Version matricielle du cout

Le problème de moindres carrés se réécrit de la manière suivante :

$$\min_{\alpha} J(\boldsymbol{\alpha}) \qquad \text{avec} \quad J(\boldsymbol{\alpha}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}||^2$$

soit en développant :

$$\begin{aligned}
\min_{\alpha} \quad J(\boldsymbol{\alpha}) &= \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}||^2 \\
&= \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}) \\
&= \frac{1}{2}\mathbf{y}^\top\mathbf{y} - \frac{1}{2}\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{y} - \frac{1}{2}\mathbf{y}^\top\mathbf{X}\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha} \\
&= \frac{1}{2}\mathbf{y}^\top\mathbf{y} - \boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{y} + \frac{1}{2}\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha}
\end{aligned}$$

Calcul matriciel : $\mathbf{y}^\top\mathbf{y}$ est un scalaire, $\mathbf{X}^\top\mathbf{y}$ est un vecteur de $\mathbb{R}^d$ et $\mathbf{X}^\top\mathbf{X}$ est une matrice $d \times d$.

Dérivées partielles du coût

$$\begin{array}{rccccc}
J(\boldsymbol{\alpha}) &= \frac{1}{2}\mathbf{y}^\top\mathbf{y} &- \boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{y} &+ \frac{1}{2}\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha} \\
\frac{\partial J(\boldsymbol{\alpha})}{\partial \alpha_i} &= 0 &- p_i &+ \frac{1}{2}\sum_{j=1}^{d+1}(M_{ij} + M_{ji})\alpha_j
\end{array}$$

avec $\mathbf{p} = \mathbf{X}^\top\mathbf{y}$ et $M = \mathbf{X}^\top\mathbf{X}$

- $\frac{\partial \boldsymbol{\alpha}^\top\mathbf{p}}{\partial \alpha_i} = \frac{\partial \sum_{j=1}^{d+1} p_j\alpha_j}{\partial \alpha_i} = p_i$
- $\frac{\partial \boldsymbol{\alpha}^\top\mathbf{M}\boldsymbol{\alpha}}{\partial \alpha_i} = \frac{\partial \sum_{j=1}^{d+1}\sum_{k=1}^{d+1}\alpha_j\alpha_k M_{jk}}{\partial \alpha_i} = \sum_{j=1}^{d+1}\alpha_j M_{ji} + \sum_{k=1}^{d+1}\alpha_k M_{ik}$

  car $(uv)' = uv' + u'v$ avec $u = \alpha_j$ et $v = \sum_{k=1}^{d+1}\alpha_k M_{jk}$

$$\nabla J(\widehat{\boldsymbol{\alpha}}) = -\mathbf{p} + \mathbf{M}\boldsymbol{\alpha} \qquad = -\mathbf{X}^\top\mathbf{y} + \mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\alpha}}$$

La minimisation du cout

La minimisation du cout $J(\boldsymbol{\alpha})$ est réalisée lorsque le gradient du cout s'annule, soit lorsque :

$$\nabla J(\widehat{\boldsymbol{\alpha}}) = 0 \Leftrightarrow -\mathbf{X}^\top \mathbf{y} \; + \; \mathbf{X}^\top \mathbf{X} \widehat{\boldsymbol{\alpha}} = 0$$

La solution du problème de minimisation des moindres carrés est le vecteur $\widehat{\boldsymbol{\alpha}}$ défini par :

$$\widehat{\boldsymbol{\alpha}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

Hypothèse

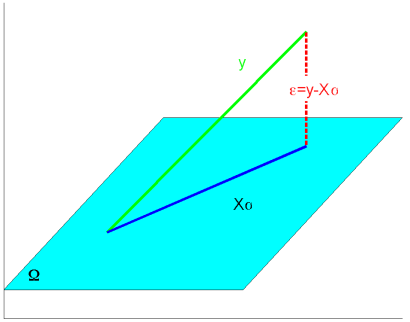
$\mathbf{X}$ est une matrice de rang $d+1$ et donc la matrice $\mathbf{X}^\top \mathbf{X}$ est inversible.
Dans les faits, ceci implique que $n > d+1$ et donc que l'on ait plus d'exemples d'apprentissage que de variables.

La régression du point de vue géométrique

- $\Omega = \text{span}\{\mathbf{X}\}$

  le sous espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de la matrice $\mathbf{X}$.

  $\mathbf{z} \in \Omega \Leftrightarrow \exists\, \boldsymbol{\alpha} \in \mathbb{R}^{d+1} \quad \mathbf{z} = \mathbf{X}\boldsymbol{\alpha}$



La régression est la projection du vecteur $\mathbf{y}$ sur le sous espace vectoriel engendré par les observations $\mathbf{X}$. En effet, on a :

$$\mathbf{X}\widehat{\boldsymbol{\alpha}} = H\mathbf{y} \qquad \text{avec} \qquad H = \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \tag{9}$$

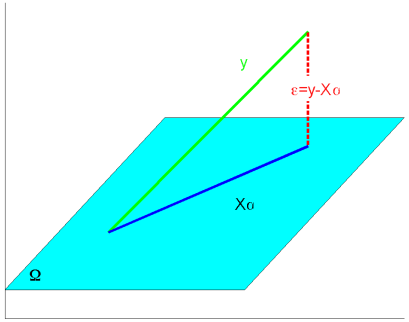
Estimation et orthogonalité

$$J(\alpha) = \frac{1}{2} \|\boldsymbol{\epsilon}\|^2$$

$\|\boldsymbol{\epsilon}\|^2$ est minimal quand on choisi $\boldsymbol{\alpha}$ tel que $\mathbf{z} = \mathbf{X}\boldsymbol{\alpha}$ est la projection orthogonale de $\mathbf{y}$ sur Ω

$$\forall \mathbf{z} \in \Omega \quad \mathbf{z}^\top \boldsymbol{\epsilon} = 0$$

$$\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\alpha}}) = 0$$



$$\begin{aligned} \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\alpha}}) = 0 \quad &\Leftrightarrow \quad \mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{X}\widehat{\boldsymbol{\alpha}} = 0 \\ &\Leftrightarrow \quad \mathbf{X}^\top \mathbf{X}\widehat{\boldsymbol{\alpha}} = \mathbf{X}^\top \mathbf{y} \\ &\Leftrightarrow \quad \widehat{\boldsymbol{\alpha}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \end{aligned}$$

Mesures de performance

Comment mesurer la performance d'une régression ?

Soit $\mathbf{y}$ les valeurs à prédire et $\hat{\mathbf{y}}$ les prédictions.
Les deux mesures suivantes sont souvent utilisées :

Erreur quadratique moyenne

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- 0 quand la prédiction est parfaite.
- Pas une mesure normalisée (dépend de la variance de $\mathbf{y}$)

Coefficient de corrélation

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sigma_y \sigma_{\hat{y}}}$$

- $\bar{y} = \frac{1}{n}\sum_i y_i$ moyenne de $\mathbf{y}$ et σ_y sa variance.
- 1 quand la prédiction est parfaite.
- Mesure normalisée.

Attention

Toujours calculer la mesure de performances sur des données de test qui n'ont pas été utilisées pour l'apprentissage sinon **risque de sur-apprentissage**.

Moindres carrés régularisés

Rappel

On cherche à minimiser l'erreur de prédiction :

$$\min_{\alpha} \quad J(\boldsymbol{\alpha}) \qquad \text{avec} \quad J(\boldsymbol{\alpha}) = \frac{1}{2}||\mathbf{y} - X\boldsymbol{\alpha}||^2$$

La solution du problème est

$$\widehat{\boldsymbol{\alpha}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

Problème

- Lorsque $n < d+1$, la matrice $\mathbf{X}^\top\mathbf{X}$ n'est pas inversible.
- Il existe une infinité de solutions, le problème est mal posé.

$\Rightarrow$ régularisation.

Régression Ridge

Problème d'optimisation

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \mathbf{w} - b)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \tag{10}$$

- On ajoute un terme de régularisation $\|\mathbf{w}\|^2$ pondéré par un coefficient de régularisation λ.
- Le paramètre λ permet de limiter le sur-apprentissage si choisi judicieusement.
- Cette régularisation aura pour effet de promouvoir les paramètres $\mathbf{w}$ de norme minimale et de rendre le problème strictement convexe.
- Choisir $\lambda = 0$ permet de revenir à la régression des moindres carrés.

Version matricielle du problème

Version matricielle du coût

$$\min_{\alpha} \quad \left\{ J'(\alpha) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}||^2 + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \mathbf{S}\boldsymbol{\alpha} \right\} \tag{11}$$

avec $\mathbf{S} \in \mathbb{R}^{(d+1)\times(d+1)}$ une matrice dont le terme général

$$S_{i,j} = \begin{cases} 1 & \text{si } i = j \text{ et } i \leq d \\ 0 & \text{sinon} \end{cases} \tag{12}$$

$\mathbf{S}$ est donc une matrice diagonale unitaire dont le dernier terme diagonal est nul.
Le passage du problème (10) au problème (11) peut se fait

$$\boldsymbol{\alpha}^\top \mathbf{S}\boldsymbol{\alpha} = \sum_{i,j=1}^{d+1} \alpha_i \alpha_j S_{i,j} = \sum_{i=1}^{d} \alpha_i^2 = \sum_{i=1}^{d} \mathbf{w}_i^2 = \|\mathbf{w}\|^2 \tag{13}$$

Dérivées partielles du coût

$$\begin{array}{rcccccl} J'(\boldsymbol{\alpha}) & = & \frac{1}{2}\mathbf{y}^\top\mathbf{y} & - & \boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{y} & + & \overbrace{\frac{1}{2}\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha} + \frac{\lambda}{2}\boldsymbol{\alpha}^\top\mathbf{S}\boldsymbol{\alpha}}^{\frac{1}{2}\boldsymbol{\alpha}^\top(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{S})\boldsymbol{\alpha}} \\ \frac{\partial J(\boldsymbol{\alpha})}{\partial \alpha_i} & = & 0 & - & p_i & + & \frac{1}{2}\sum_{j=1}^{d+1}(M_{ij} + M_{ji})\alpha_j \end{array}$$

avec $\mathbf{p} = \mathbf{X}^\top\mathbf{y}$ et $M = \mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}$

- $\frac{\partial \boldsymbol{\alpha}^\top \mathbf{p}}{\partial \alpha_i} = \frac{\partial \sum_{j=1}^{d+1} p_j \alpha_j}{\partial \alpha_i} = p_i$
- $\frac{\partial \boldsymbol{\alpha}^\top \mathbf{M}\boldsymbol{\alpha}}{\partial \alpha_i} = \frac{\partial \sum_{j=1}^{d+1}\sum_{k=1}^{d+1} \alpha_j \alpha_k M_{jk}}{\partial \alpha_i} = \sum_{j=1}^{d+1} \alpha_j M_{ji} + \sum_{k=1}^{d+1} \alpha_k M_{ik}$

  car $(uv)' = uv' + u'v$ avec $u = \alpha_j$ et $v = \sum_{k=1}^{d+1} \alpha_k M_{jk}$

$$\nabla J(\widehat{\boldsymbol{\alpha}}) = -\mathbf{p} + \mathbf{M}\boldsymbol{\alpha} \qquad = -\mathbf{X}^\top\mathbf{y} \; + \; (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S})\widehat{\boldsymbol{\alpha}}$$

La minimisation du cout

La minimisation du cout $J(\boldsymbol{\alpha})$ est réalisée lorsque le gradient du cout s'annule, soit lorsque :

$$\nabla J(\widehat{\boldsymbol{\alpha}}) = 0 \Leftrightarrow -\mathbf{X}^\top \mathbf{y} + (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{S})\widehat{\boldsymbol{\alpha}} = 0$$

La solution du problème de minimisation des moindres carrés est le vecteur $\widehat{\boldsymbol{\alpha}}$ défini par :

$$\widehat{\boldsymbol{\alpha}} = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{S}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

Régularisation

La matrice $\mathbf{S}$ ajoute λ sur la diagonale de $\mathbf{X}^\top \mathbf{X}$, ce qui la rend inversible. Le problème est donc bien posé et une solution unique existe.

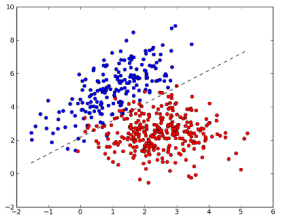
Classification linéaire

Objectif

- Apprendre une fonction linéaire $f(\cdot)$ permettant de prédire une valeur binaire $y \in \{-1, 1\}$ à partir d'une observation $\mathbf{x} \in \mathbb{R}^d$.
- En pratique on cherche à déterminer les coefficients $(\mathbf{w}, b)$ de $f(\cdot)$ à partir d'un ensemble d'apprentissage $\{\mathbf{x}_i, y_i\}_{i=1,\dots,n}$.
- La classe prédite est le signe de la onction de prédiction $f(\cdot)$

Exemples

- Reconnaissance de caractères.
- Aide au diagnostique.
- Inspection de pièces.



Mise en oeuvre

Minimisation des moindres carrés

- On utilise directement les valeurs $\{-1, 1\}$ et on tente de les prédire avec une fonction linéaire.
- $\mathbf{y}$ contient uniquement -1 ou 1
- Problème d'optimisation :

$$\min_{\alpha} \quad \|\boldsymbol{\epsilon}\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|^2 \tag{14}$$

Solution du problème

$$\widehat{\boldsymbol{\alpha}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

Ce qui est exactement la solution pour une régression classique. La transposition pour la régression ridge est évidente.

Relations avec la décision bayésienne

Régression pour la classif.

Estimation :

$$\boldsymbol{\alpha} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$$

- b estimé en même temps que $\mathbf{w}$.
- $\mathbf{X}^\top \mathbf{X}$ matrice de corrélation.
- $\mathbf{X}^\top \mathbf{y} = n_1 \boldsymbol{\mu}_1 - n_2 \boldsymbol{\mu}_2$.
- n_i nb. d'exemple de chaque classe.

Décision bayesienne

Estimation :

$$\begin{aligned} \mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ b &= \frac{1}{2}(\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1) \\ &\quad + \log(P(\omega_1)) - \log(P(\omega_2)) \end{aligned}$$

- Cas 3 du cours : $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$.
- $\boldsymbol{\Sigma}$ matrice de covariance.
- $\boldsymbol{\mu}_i$ moyenne de chaque classe.
- $\log(P(\omega_i))$ encode le nombre d'exemples de chaque classe.