

Machine Learning

January 25, 2017

1 La regression dans tous ses états

1.1 Introduction à la régression multiple

La régression multiple permet d'étudier la liaison entre une variable (à expliquer) Y et un ensemble de p variables explicatives X_1, \dots, X_p .

Le modèle de la régression multiple est définie par

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon \quad (1)$$

où les coefficients de régression β_j sont des paramètres fixe mais inconnus, et ϵ un terme aléatoire suivant une $\mathcal{N}(0, \sigma^2)$

1.1.1 Données et modèle statistique

On dispose de n observations des variables X_1, \dots, X_p . Soit

$$X = \begin{bmatrix} y_1 & x_{11} & \dots & x_{p1} \\ \dots & \dots & \dots & \dots \\ y_m & x_{1m} & \dots & x_{pm} \end{bmatrix} \quad (2)$$

(Tableau d'observations)

La modèle statistique est définie comme

Pour chaque individu t , on considère que la valeur y_i prise par Y est une réalisation d'une variable aléatoire Y_i définie par

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi} + \epsilon_i \quad (3)$$

où ϵ_i est un terme aléatoire suivant une $\mathcal{N}(0, \sigma^2)$. Il faut supposer de plus que les $\epsilon_1, \dots, \epsilon_m$ sont indépendantes les un les autres.

Sur l'exemple des automobiles, on considère que le prix Y_i de la voiture i suit une loi normale de moyenne

$$\mu_i = \beta_0 + \beta_1 PUISSANCE_i + \dots + \beta_p LARGEUR_i + \epsilon. \quad (4)$$

1.1.2 Estimations des paramètres du modèle

À l'aide des n observations des variables Y, X_1, \dots, X_p , nous allons chercher à estimer les paramètres β_0, \dots, β_p du modèle.

On cherche $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ tel que

$$\hat{\beta} = \operatorname{argmin} \|y - X\beta\|_2^2 = \operatorname{argmin} (y - X\beta)^T (y - X\beta) \quad (5)$$

Géométriquement, cas où $p = 1$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 \quad (6)$$

On cherche $\hat{\beta}_0$ et $\hat{\beta}_1$ tel que $\sum \epsilon_i^2$ soit minimal.

$$\frac{\partial L}{\partial \beta} = -X^T y - X^T y + 2X^T X \hat{\beta} = 0 \quad (7)$$

Ainsi

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (8)$$

remarque 1: Notons que $\hat{\beta}$ est de la forme Hy

remarque 2:

$$\hat{\beta} = (X^T X)^{-1} X^T y = (X^T X)(X^T X)^{-2} X^T y = X^T \hat{\alpha} \quad (9)$$

Cela signifie que $\hat{\beta}$ s'exprime comme la combinaison linéaire des individus

$$\hat{\beta} = \sum_{i=1}^n \hat{\alpha}_i x_i \quad (10)$$

Remarque 3 Posons

$$X = V \Sigma U^T \quad (11)$$

(Décomposition en valeurs singuliers)

$$V^T V = I \quad (12)$$

et Σ est diagonale et $U^T U = I$ Posons

$$\Lambda = \Sigma^T \Sigma \quad (13)$$

matrice diagonale des valeurs de $X^T X$. Ainsi

$$\begin{aligned} \hat{\beta}^{OLS} &= (X^T X)^{-1} X^T y = (U \Sigma V^T V \Sigma U^T)^{-1} U \Sigma V^T y \\ &= U \Sigma^{-1} \Sigma^{-1} U^T U \Sigma V^T y \\ &= U \Lambda^{-1} \Sigma V^T y \\ &= \sum_{j=1}^P \frac{v_j^T y}{\sqrt{\lambda_j}} \mu_j \end{aligned} \quad (14)$$

Interprétation géométrique de la régression multiple...

Pour un nouvel individu x

$$\hat{y}(x) = x^t \hat{\beta} \quad (15)$$

Remarque 4

En termes de prédiction, on peut aussi obtenir des expressions duales

$$\hat{y} = x^T \hat{\beta} = \sum_{j=1}^n \alpha_j x_j^T x \quad (16)$$

Cette expression duale a la particularité de ne dépendre que des produits scalaires entre observations.

1.1.3 Qualité des estimations

On souhaite évaluer la précision des estimations

$$\begin{aligned}\mathbb{E}[\hat{\beta}] &= \mathbb{E}[(X^T X)^{-1} X^T Y] = (X^T X)^{-1} X^T \mathbb{E}[Y] \\ &= (X^T X)^{-1} X^T \mathbb{E}[X\beta + \epsilon] \\ &= (X^T X)^{-1} (X^T X) \beta \\ &= \beta\end{aligned}\tag{17}$$

$$\begin{aligned}Var(\hat{\beta}) &= var((X^T X)^{-1} X^T Y) \\ &= (X^T X)^{-1} X^T var(Y) X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1}\end{aligned}\tag{18}$$

où

$$var(Y) = \sigma^2 I\tag{19}$$

Le MSE (Mean Square Error) d'un estimateur $\hat{\beta}$ d'un vecteur β est défini

$$\begin{aligned}MSE(\hat{\beta}) &= \mathbb{E}[tr(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] \\ &= \mathbb{E}[(\hat{\beta} - \beta)^T (\hat{\beta} - \beta)] \\ &= \mathbb{E}[||\hat{\beta} - \beta||_2^2] \\ &= [\mathbb{E}[\hat{\beta} - \beta]^T][\mathbb{E}[\hat{\beta} - \beta]] + \mathbb{E}[(\hat{\beta} - \mathbb{E}[\hat{\beta}])(\hat{\beta} - \mathbb{E}[\hat{\beta}])^T] \\ &= biais^2(\hat{\beta}) + tr(var(\hat{\beta})) \\ &= \sigma^2 tr(X^T X)^{-1} \\ &= \sigma^2 \sum_{j=1}^P \frac{1}{\lambda_j}\end{aligned}\tag{20}$$

Si les données sont mal-conditionnées, alors \Rightarrow petit valeurs de $X^T X \Rightarrow$ instabilité des coefficients de regression \Rightarrow Explosion du MSE \Rightarrow écart entre β et $\hat{\beta}$

Illustration de la multi-linéarité

Considérons le modèle suivant

$$Y = \beta_1 X_1 + \beta_2 X_2 + \Sigma\tag{21}$$

On suppose que les données sont standardisées

$$cor(X_1, X_2) = r_{12}, cor(X_j, Y) = r_{jy}\tag{22}$$

L'estimation des moindres carrées

$$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)\tag{23}$$

est donné par

$$(X^T X) \hat{\beta} = X^T Y\tag{24}$$

Comme les données sont standardisées

$$\begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \sim \begin{pmatrix} r_{1y} \\ r_{2y} \end{pmatrix} \quad (25)$$

L'inverse de $X^T X$ est donnée par

$$C = (X^T X)^{-1} = \frac{1}{1 - r_{12}^2} \begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix} \quad (26)$$

On rappelle que

$$var(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \quad (27)$$

donc

$$var(\hat{\beta}_j) = \sigma^2 C_j \quad (28)$$

Alors, une forte corrélation entre X_1 et X_2 est indiquée par $|r_{12}| \rightarrow 1$. Ceci implique que $var(\hat{\beta}_j) \rightarrow +\infty$. Ainsi, MSE qui explose.

De manière générale, on peut montrer que dans le cas de p variables explicatives, des éléments diagonaux de $C = (X^T X)^{-1}$ sont égaux à

$$C_j = \frac{1}{1 - R_j^2} \quad (29)$$

où R_j^2 est le coefficient de détermination entre X_j et les p autres variables. S'il existe une forte multi-colinéarité entre X_j et les $(p-1)$ autres variables. $R_j^2 \rightarrow 1$

Exemple II

(Voir photos)

1.2 Facteur de shrinkage

On rappelle que

$$\hat{\beta}^{OLS} = (X^T X)^{-1} X^T y = \sum_{j=1}^P \frac{v_j^T y}{\sqrt{\lambda_j}} \mu_j \sum_{j=1}^P z_j \quad (30)$$

Dans sa forme générale, on définit un estimateur générale de β par

$$\hat{\beta}^{shi} = \sum_{j=1}^P f(\lambda_j) z_j \quad (31)$$

$f(\lambda_j)$ est un facteur de shrinkage qui va jouer sur le MSE. Ici, on se concentre sur les facteurs de shrinkage indépendant de y . Malheureusement, on a

$$\hat{\beta}^{shi} = U \Sigma^{-1} D V^T y = H^{shi} y \quad (32)$$

avec

$$D = \text{diag}(f(\lambda_1), \dots, f(\lambda_p)) \quad (33)$$

Étudions l'influence de (Photos)

$$\begin{aligned}
\hat{\beta} &= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \{ \|y - X\beta\| + \lambda \|\beta\|_2^2 \} \\
&= (X^T X + \lambda I_p)^{-1} X^T y \\
&= \sum_{j=1}^p \frac{\lambda_j}{\lambda_j + \lambda} \frac{v_j^T y}{\sqrt{\lambda_j}} \\
&= \sum_{j=1}^p f(\lambda_j) z_j
\end{aligned} \tag{34}$$

$$\hat{y}_\lambda = X \hat{\beta}_\lambda^{RR} = V \Sigma U^T U (\Lambda + \lambda I)^{-1} \Sigma \Lambda y = V \Sigma (\Lambda + \lambda I)^{-1} \Sigma V^T y \tag{35}$$

1.2.1 Choix de modèle (λ) + ridge path validation croisée

- Ensemble d'apprentissage permet de construire le modèle
- Ensemble de test permet d'évaluer la qualité du modèle

Nécessite de déterminer λ sur la base d'un critère objectif

Un modèle devrait prédire efficacement des individus qui n'ont pas servi à sa construction

Pour évaluer la qualité du modèle, on utilise des stratégies de validation croisée.

K-fold cross validation

1. Partition du jeu de données en K parties de taille égale T_1, \dots, T_K
2. Pour chaque $k = 1, \dots, K$ construire $\hat{\beta}_\lambda^{-k}$ sans T_k
3. $\hat{y}_\lambda^{-k} = T_k \hat{\beta}_\lambda^{-k}$
4. Calcul d'erreurs en test

$$CV_{\text{erreur}}(\lambda, k) = \frac{1}{n_k} \sum (y_k - \hat{y}_\lambda^{-k})^2 \tag{36}$$

Ainsi, Taux d'erreur global(λ)

$$= \frac{1}{K} \sum_{k=1}^K CV_{\text{erreur}}(\lambda, k) \tag{37}$$

Ainsi,

$$\lambda^* = \underset{\lambda}{\operatorname{argmin}} (Taux d'erreur global(\lambda)) \tag{38}$$

On appelle que

$$\hat{\beta}^{RR} = \operatorname{argmin}\{\|y - x\|_2^2 + \lambda\|\beta\|_2^2\} = (X^T X + \lambda I)^{-1} X^T y \quad (39)$$

Proposition $\forall \lambda > 0$ on a

$$(X^T X + \lambda I_p)^{-1} X^T = X^T (X X^T + \lambda I_n)^{-1} \quad (40)$$

Preuve

$$(X^T X + \lambda I_p)^{-1} X^T (X X^T + \lambda I_n) = X^T \quad (41)$$

$$(X^T X + \lambda I_p)^{-1} (X^T X + \lambda I_p) X^T = X^T \quad (42)$$

Il vient de cette proposition une formulation duale pour $\hat{\beta}^{RR}$

$$\begin{aligned} \hat{\beta}^{RR} &= X^T (X X^T + \lambda I_n)^{-1} y \\ &= X^T \hat{\alpha}^{RR} \end{aligned} \quad (43)$$

Remarque: $\hat{\beta}^{RR}$ s'exprime comme une combinaison linéaire des observations.
En termes de prédiction

$$\hat{y}_\lambda = X \hat{\beta}_\lambda^{RR} = X X^T \hat{\alpha}^{RR} \quad (44)$$

$$h(x) = \sum_{i=1}^n \alpha_i x_i^T x \quad (45)$$

Conclusion

On remarque que $\hat{\alpha}$ et \hat{y} ne s'expriment qu'au (Photo)

2 Méthode à noyaux

Introduction Supposons que l'on recherche à résoudre un problème de régression non-linéaire.

Il est tout à fait possible d'utiliser une fonction de redescription $\Phi : \mathbb{R}^p \rightarrow \mathcal{F}$ (espace de description). La solution du problème de régression linéaire dans l'espace de redescription est de la forme

$$\begin{aligned} h(x) &= \langle \beta, \Phi(x) \rangle \\ &= \sum_{j=1}^n \beta_j \Phi_j(x) + \beta_0 \end{aligned} \quad (46)$$

avec $\Phi(x) = \{\Phi_1(x), \Phi_2(x), \dots\}$

En résolvant comme précédemment dans l'espace de redescription, devient la représentation duale.

$$h(x) = \sum_{i=1}^n \Phi(x_i)^T \Phi(x) \quad (47)$$

En rapprochant $h(x) = \beta^T \Phi(x)$ et $h(x) = \sum_{i=1}^n \alpha_i \Phi(n_i)^T \Phi(x)$ devient

$$\beta^T = \sum_{i=1}^n \alpha_i \Phi(x_i)^T \quad (48)$$

Commentaire 1: Comment trouver une fonction de redescription adéquate?

Commentaire 2: Coût calculatoire des produits scalaires dans un espace \mathcal{F} dont la dimension peut être très grande.

Plutôt que de représenter les observations par un tableau individus x variables, on peut les représenter par une matrice de similarité de dimension $n \times n$

$$[K]_y = k(x_i, x_j) \quad (49)$$

$$[X] \sim [K] \quad (50)$$

$$(x_i)_{n \times p} \sim (k(x_i, x_n))_{n \times n} \quad (51)$$

Remarque: Mesure de similarité quelle que soit la nature et la complexité des objets.(images, graphs, séquences ADN,...) Remarque 2: La matrice K est toujours de dimension $n \times n \forall$ la taille de l'objet.

Dans ce cours nous nous restreignons à une classe particulières de fonction k .

Définition Fonction semi définie positive

Une fonction k semi définie positive sur l'ensemble X est une fonction $k : X, X \rightarrow \mathcal{R}$ symétrique

$$\forall (x, x') \in X, k(x, x') = k(x', x) \quad (52)$$

et qu'il satisfait pour (Photo)

et $(a_1, \dots, a_n) \in \mathcal{R}^n$

$$\sum_{i,j=1}^n a_i a_j k(x_i, x_j) \geq 0 \quad (53)$$

En d'autres termes $K = (k(x_i, x_j))$ est semi définie positive.

Exemple 1: Le plus simple des noyaux semi définie positif

Soit $X \in \mathbb{R}^P$ et la fonction $k : \mathcal{X}^2 \rightarrow \mathbb{R}$ définie par

$$\forall (x, x') \in (\mathbb{R}^P)^2, k(x, x') = \langle x, x' \rangle_{\mathbb{R}^P} \quad (54)$$

Symétrie

$$\langle x, x' \rangle_{\mathbb{R}^P} = \langle x', x \rangle_{\mathbb{R}^P} \quad (55)$$

Théorème de Moore-Aronszajn, 1950

(Photo)

Définition: Soit \mathcal{H}

(Photo)

2. La fonction k est une fonction noyau reproduisante

$$\forall f \in \mathcal{H} \text{ on a } \langle f, k(n, \cdot) \rangle = f(x) \quad (56)$$

Le produit scalaire est alors défini comme suit:

Soit f et $g \in \mathcal{H}$ définies par

$$f(x) = \sum_{i=1}^n \alpha_i k(x_i, x) \quad (57)$$

et

$$g(x) = \sum_{j=1}^n \beta_j k(x_j, x) \quad (58)$$

alors

$$\begin{aligned} \langle f, g \rangle &= \sum_{j=1}^n \sum_{i=1}^n \alpha_i \beta_j k(x_i, x_j) \\ &= \sum_{i=1}^n \alpha_i g(x_i) \\ &= \sum_{j=1}^n \beta_j f(x_j) \end{aligned} \quad (59)$$

Representation theorem (Kimeldorf & Wabhin 1970)

Soit \mathcal{H} un RKHS associé à un noyau k . Soit $\Omega[0, +\infty[\rightarrow \mathbb{R}$ une fonction strictement croissante Soit \mathcal{X} un ensemble et $l(\mathcal{X} \times \mathbb{R}) \rightarrow \mathbb{R}$ une fonction de coût.

Alors toute solution du problème d'optimisation suivant

$$\min_f J_f = l((x_1, y_1, f(x_1)), \dots, (x_n, y_n, f(x_n))) + \lambda \Omega(\|f\|_{\mathcal{H}}) \quad (60)$$

admet une représentation de la forme $f^* = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$

Remarque 1

$$l((x_1, y_1, f(x_1)), \dots, (x_n, y_n, f(x_n))) + \lambda \Omega(\|f\|_{\mathcal{H}}) \quad (61)$$

conduit à la regression rigide.

Remarque 2: Quelque soit la dimension de l'espace \mathcal{H} , la solution évolue dans $\text{span}\{k(x_i, \cdot), 1 \leq i \leq n\}$ de dimension n connue, bien que \mathcal{H} puisse être de dimension infinie. \Rightarrow Développement d'algorithme efficace.

Preuve du théorème du représentant Supposons $f \in \mathcal{H}$ projetée sur $\text{span}\{k(x_i, \cdot), 1 \leq i \leq n\}$ Soit f_s (la composante dans l'espace engendré) et f_{\perp} la composante orthogonale

$$f = f_s + f_{\perp} \quad (62)$$

Ainsi

$$\|f\|_{\mathcal{H}}^2 = \|f_s\|_{\mathcal{H}}^2 + \|f_{\perp}\|_{\mathcal{H}}^2 \geq \|f_s\|_{\mathcal{H}}^2 \quad (63)$$

Puisque Ω est strictement croissante, on a

$$\Omega(\|f\|_{\mathcal{H}}^2) \geq \Omega(\|f_s\|_{\mathcal{H}}^2) \quad (64)$$

Cela signifie que Ω est minimisée pour des fonctions $f \in \text{span}\{k(x_i, \cdot), 1 \leq i \leq n\}$

Par ailleurs, des propriétés reproduisantes de k on obtient

$$f(x_i) = \langle f, k(x_i, \cdot) \rangle = \langle f_s, k(x_i, \cdot) \rangle_{\mathcal{H}} + \langle f_{\perp}, k(x_i, \cdot) \rangle_{\mathcal{H}} = \langle f_s, k(x_i, \cdot) \rangle = f_s(x_i) \quad (65)$$

Par conséquent,

$$l((x_1, y_1, f(x_1)), \dots, (x_n, y_n, f(x_n))) = ((x_1, y_1, f_s(x_1)), \dots, (x_n, y_n, f_s(x_n))) \quad (66)$$

Donc l ne dépend que des composantes de $f \in \text{span}\{k(x_i, \cdot), 1 \leq i \leq n\}$ et Ω est minimisée que si $f \in$ cet espace

Conclusion: $J(f)$ est minimisée que si $f \in \text{span}\{k(x_i, \cdot), 1 \leq i \leq n\}$ et la solution est donc de la forme

$$f^* = \sum_{i=1}^n \alpha_i k(x_i, \cdot) \quad (67)$$

Kernel Rigide Regression

Un point d'une fonctionnelle

Soit \mathcal{H} un RKHS de noyau k . On considère le problème suivant

$$f^* = \underset{f \in \mathcal{H}}{\text{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^n (y_i - f(x_i))^2 + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2 \right\} \quad (68)$$

= terme d'attachement aux données + terme de régularisation.

Par le théorème de représentant, on sait que

$$f^* = \sum_{j=1}^n \alpha_j k(x_j, \cdot) \quad (69)$$

(Photo)

Exercice: Montrer que

(Photo)

Kernel Ridge Regression

(Photo)

Franchement, le choix de la fonction Φ se ramène au choix d'une fonction k sd.

Exemple démonstratif:

similarity: fonction noyau

$$k(x_i, x_j) = (1 + x_i^T x_j)^2 \quad (70)$$

On va construire explicitement l'espace induit par la fonction k .

$$\begin{aligned} k(x_i, x_j) &= 1 + x_{i1}^2 x_{j1}^2 + 2x_{i1} x_{j1} x_{i2} x_{j2} + x_{i2}^2 x_{j2}^2 + 2x_{i1} x_{j1} x_{i2} x_{j2} \\ &= [1 x_{i1}^2 \sqrt{2} x_{i1} x_{i2} x_{i2}^2 \sqrt{2} x_{i1} \sqrt{2} x_{i2}]^T [1 x_{j1}^2 \sqrt{2} x_{j1} x_{j2} x_{j2}^2 \sqrt{2} x_{j1} \sqrt{2} x_{j2}] \\ &= \Phi(x_i)^T \Phi(x_j) \end{aligned} \quad (71)$$

Astude du noyau

$$\hat{\alpha}^{RR} = (XX^T + \lambda I_n)^{-1}y \quad (72)$$

$$\hat{\alpha}^{RR} = (K + \lambda I_n)^{-1}y \quad (73)$$

Autres exemples de noyaux, Noyau polynomial

$$k(x, y) = (c + x_i^T x_j)^p \quad (74)$$

Noyau gaussien

$$\exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) \quad (75)$$

3 Analyse discriminante

Introduction Soit Y une variable à expliquer QUALITATIVE à K catégorie $y \in 1, \dots, K$

Soient X_1, \dots, X_p les p variables explicatives.

Objectif 1: L'analyse discriminante descriptive: Proposer un système de représentation qui permet de discerner le plus possible les différents groupes d'individus. *Rightarrow* L'analyse factorielle discriminante

Objective 2: Construire une fonction de classement qui permet de prédire le groupe d'appartenance d'un individu à partir des valeurs prises par les variables explicatives. *Rightarrow* Analyse Discriminante Bayésienne

Soit n_j le nombre d'observation appartenant au groupe j . Soit n le nombre d'observations total.

$$g = (g_1, \dots, g_p) \quad (76)$$

les centres de gravité du nuage global. noyaux des X_j calculé à partir de toutes les observations.

$$m_k = (m_{k1}, \dots, m_{kp}) \quad (77)$$

centre de gravité du nuage des individus de la class k .

$$V = \frac{1}{n} \sum_{i=1}^n (x_i - y)(x_i - y)^T \quad (78)$$

matrice de covariance des données complètes.

$$B = \frac{1}{n} \sum_{k=1}^K n_k (g_k - g)(g_k - g)^T \quad (79)$$

$$W = \frac{1}{n} \sum_{k=1}^K n_k V_k \quad (80)$$

avec

$$V_k = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} (x_i - m_k)(x_i - m_k)^T \quad (81)$$

qui est la matrice de covariance intraclasse.

On peut montrer(exercice) que(voir TP)

$$V = B + W \quad (82)$$

Construisons la variance globale du nuage de points.

$$\begin{aligned} \sigma^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - y)(x_i - y)^T \\ &= \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{L}_k} (x_i - g)^*(x_i - g) \\ &= \frac{1}{n} \sum_{k=1}^K SS(k) \end{aligned} \quad (83)$$

$$\begin{aligned} SS(k) &= \sum_{i \in \mathcal{L}_k} (x_i - g)^T (x_i - g) \\ &= \sum_{i \in \mathcal{L}_k} (x_i - m_k + m_k - g)^T (x_i - m_k + m_k - g) \\ &= \sum_{i \in \mathcal{L}_k} (||x_i - m_k||^2 + ||m_k - g||^2 + 2(x_i - m_k)(m_k - g)) \\ &= \sum_{i \in \mathcal{L}_k} (||x_i - m_k||^2 + ||m_k - g||^2) \end{aligned} \quad (84)$$

Donc, la variance totale s'écrit

$$\begin{aligned} \sigma^2 &= \frac{1}{n} \left[\sum_{k=1}^K \sum_{i \in \mathcal{L}_k} ||x_i - m_k||^2 + \sum_{k=1}^K n_k ||m_k - m||^2 \right] \\ &= \frac{1}{n} \sum_{k=1}^K n_k \frac{1}{n_k} \sum_{i \in \mathcal{L}_k} ||x_i - m_k||^2 + \sum_{k=1}^K ||m_k - g||^2 \\ &= \frac{1}{n} \sum_{k=1}^K n_k \left[\frac{1}{n_k} \sum_{i \in \mathcal{L}_k} ||x_i - m_k||^2 + ||m_k - g||^2 \right] \end{aligned} \quad (85)$$

Si l'on definit une operation de projection $\Pi = vv^T$ avec $v^T v = 1$

$$\begin{aligned} \sigma_{within}^2 &= \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{L}_k} (\Pi_{x_i} - \Pi_{m_k})^T (\Pi_{x_i} - \Pi_{m_k}) \\ &= \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{L}_k} (vv^T x_i - vv^T m_k)^T (vv^T x_i - vv^T m_k) \\ &= \frac{1}{n} \sum_{k=1}^K (x_i - m_k)^T vv^T vv^T (x_i - m_k) \end{aligned} \quad (86)$$

$$\begin{aligned}
\sigma_w^2 &= \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{L}_k} (x_i - m_k)^T v v^T (x_i - m_k) \\
&= \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{L}_k} v^T (x_i - m_k) (x_i - m_k)^T v \\
&= v^T \left[\frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{L}_k} (x_i - m_k) (x_i - m_k)^T \right] v \\
&= v^T \left[\frac{1}{n} \sum_{k=1}^K n_k \sum_{i \in \mathcal{L}_k} (x_i - m_k) (x_i - m_k)^T \right] v \\
&= v^T v
\end{aligned} \tag{87}$$

Pour la variance between projetée sur V

$$\begin{aligned}
\sigma_{between}^2(v) &= \frac{1}{n} \sum_{k=1}^K n_k (\Pi m_k - \Pi y)^T (\Pi m_k - \Pi y) \\
&= v^T \left[\sum_{k=1}^K n_k \frac{1}{n} (m_k - y) (m_k - y)^T \right] v
\end{aligned} \tag{88}$$

Comme $V = B + W$

$$\sigma_{total}^2(v) + v^T V v \tag{89}$$

L'analyse discriminante factorielle est définie par le problème d'optimisation suivant

$$\max \left(\frac{v^T B v}{v^T W v} \right) \tag{90}$$

On remarque si v^* est solution alors αv^* est également solution

$$\max(v^T B v), \text{ sc } v^T W v = 1 \tag{91}$$

Remarque: On aurait pu également considérer le problème d'optimisation suivant

$$\max \left(\frac{v^T B v}{v^T V v} \right) \Leftrightarrow \max v^T B v \text{ sc } v^T V v = 1 \tag{92}$$

Considerons le lagrangien associée à ce problème d'optimisation

$$L = v^T B v - \lambda (v^T V v - 1) \tag{93}$$

En annulant la dérivée de L par rapport à v , devient

$$\frac{\partial L}{\partial v} = 2Bv - \lambda V v = 0 \tag{94}$$

Ainsi,

$$V^{-1}Bv = \lambda v \quad (95)$$

La solution est obtenue en considérant v le premier vecteur propre.

Commentaire: Le nombre de vecteurs propres associé à des valeurs propres non nulles est égal à $K - 1$ (du fait du rang de B)

3.1 Analyse Discriminante Bayesienne

On cherche à construire une règle de prédiction sur l'affectation d'un individu à l'un des groupes. On s'intéresse à estimer

$$p_k(x) = P(y = k | X_1 = x_1, \dots, X_p = x_p) \quad (96)$$

On effectuera l'observation un groupe le plus probable. On utilise la formule de Bayesienne pour calculer $p_k(x)$ avec l'hypothèse que le $X = (X_1, \dots, X_p) \sim \mathcal{N}(\mu_k, \Sigma_k)$ sur chaque population et en fonction des probabilités a priori $q_k = \mathbb{P}(y = k)$ d'appartenance aux différentes sous-population. La formule de Bayes permet d'obtenir la probabilité a posteriori $p_k(x)$ en fonction de la densité $f_k(x)$ et des probabilités q_k .

$$p_k(x) = \mathbb{P}(Y = k | X = x) = \frac{f_k(x)q_k}{\sum_{k=1}^K f_k(x)q_k} \quad (97)$$

3.1.1 Cas de matrice de covariances homogène

Ici, on suppose que $X = (X_1, \dots, X_p) \sim \mathcal{N}(\mu_h, \Sigma)$ sur la sous population \mathcal{P}_k (Photo)

En remplaçant la moyenne μ_h par leur estimation,

$$\hat{\mu}_h = m_h \quad (98)$$

$$\hat{q}_k = \frac{n_k}{n} \quad (99)$$

Et la matrice de covariance par son estimation

$$S = \frac{1}{n - K} \sum_{k=1}^K n_k V_k \quad (100)$$

On obtient des fonctions discriminantes linéaire $d_h(x)$ suivante

$$d_h(x) = \frac{1}{2} m_k^T S^{-1} m_k - h^T S^{-1} x + \ln(\hat{q}_h) \quad (101)$$

on en déduit une estimation des probabilités a posteriori

$$\hat{p}_h(x) = \frac{\exp(d_h(x))}{\sum_{k=1}^K \exp(d_k(x))} \quad (102)$$

Un nouvel individu sur lequel a été observé les valeurs $x = (x_1, \dots, x_p)$ est affecté au groupe h pour lequel $d_h(x)$ est maximum.

3.1.2 Cas où les covariances sont hétérogène

On suppose maintenant que $X = (X_1, \dots, X_p)$ est une loi multinomial $\mathcal{N}(\mu_h, \Sigma_h)$ sur chaque sous population \mathcal{P}_h . Avec un raisonnement analogue et en considérant

$$\hat{\mu} \quad (103)$$

On obtient des fonctions discriminantes quadratiques définie par

$$d_h(x) = -\frac{1}{2}(x - m_k)^T S_h^{-1}(x - m_h) \quad (104)$$

(Photo)

Pour chacun entre la version linéaire ou quadratique, on procède par cross validation.

4 Les Support Vector Machines(SVM)

Introduction Dans ce cours, on se concentre sur des problématiques de classification binaire. L'approche proposée initialement par (Boser et al. 1995) repose sur la notion intuitive d'hyperplan séparable de marge maximale. Ce nouveau concept est motivé par la question suivante:

Comment construire à partir d'une base d'apprentissage finie une fonction de décision capable de prédire efficacement la classe d'appartenance d'un individu qui n'a pas servi à construire le modèle.

4.1 SVM Linéaire et classe linéaire séparable

Soient n observations d'écrites par p variables séparable en 2 classes. À chaque observation x_i est donc associé un label $y \in \{-1, 1\}$.

Soit un hyperplan défini par

$$\mathcal{H} = \{x | f(x) = x^T \beta + \beta_0 = 0\} \quad (105)$$

Si les données sont linéairement séparable, alors. Il existe une infinité d'hyperplan tel que

$$y_i f(x_i) > 0 \quad (106)$$

L'hyperplan séparateur optimal est défini par les vecteurs β et β_0 comme suit:

Trouver β et β_0 qui rendent maximal la distance minimale des individus à l'hyperplan.

La distance d'un individu de \mathcal{C}_1 à l'hyperplan \mathcal{H}

$$d(x, \mathcal{H}) = x^T \beta - (-\beta_0) = x^T \beta + \beta_0 \quad (107)$$

La distance d'un individu de \mathcal{C}_{-1} à l'hyperplan \mathcal{H}

$$d(x, \mathcal{H}) = -\beta_0 - x^T \beta \quad (108)$$

Ainsi, $x^T \beta + \beta_0$ représente la distance "signée" de x à l'hyperplan \mathcal{H} . Cette distance est positive si $x \in \mathcal{C}_1$ et négative si $x \in \mathcal{C}_{-1}$. De plus, $y_i f(x_i) \geq 0, \forall i$. Si les données sont linéairement séparables.

Conclusion

L'hyperplan séparateur optimal maximise les distances signées entre les points et l'hyperplan \mathcal{H} . On cherche donc à maximiser l'épaisseur (de la marge par rapport à β et β_0).

$$\|\beta\| = 1 \quad (109)$$

Si x_i à droite de (Photo)

Ainsi, maximiser C sert à minimiser $\|\beta\|_2$ (lien avec la ridge) Par conséquent, dans le cas de 2 classes linéairement séparables. La recherche de l'hyperplan séparateur optimal sert à résoudre le problème d'optimisation quadratique suivant:

$$\min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 \text{ s.c. } y_i(x_i^T \beta + \beta_0) \geq 1, \forall i = 1, \dots, n \quad (110)$$

Les contraintes de finissent une marge de par et d'autre de l'HSO les valeurs x_i tel que

$$y_i(x_i^T \beta + \beta_0) = 1 \quad (111)$$

Sont situés sur la marge et sont appelés "support vectors".

Il faut maintenant résoudre ce problème d'optimisation.

Définissons le Lagrangien associé

$$L(\beta, \beta_0, \alpha) = \frac{1}{2} \|\beta\|^2 \sum_{i=1}^n \alpha [y_i(x_i^T \beta + \beta_0) - 1] \quad (112)$$

où les $x_i \geq 0$ sont les multiplicateurs de Lagrange.

On doit minimiser par rapport à β et β_0 et maximiser par rapport à α . En annulant la dérivée du Lagrangien par rapport à β et β_0 , on obtient

$$\frac{\partial L}{\partial \beta} = \beta - \sum_{i=1}^n \alpha_i y_i x_i \quad (113)$$

$$\frac{\partial L}{\partial \beta_0} = - \sum_{i=1}^n \alpha_i y_i \quad (114)$$

En développant cette formule et injecter (113) et (114), on atteint la formulation duale du Lagrangien

$$L(\alpha) = \sum_{i=1}^n \alpha_i - \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j \quad (115)$$

qui ne dépend pas de β et β_0 . On cherche ensuite à maximiser $L(\alpha)$ par rapport à α sous les contraintes

$$\begin{cases} \alpha_i \geq 0 \\ \sum_{i=1}^n \alpha_i y_i = 0 \end{cases} \quad i = 1, \dots, n \quad (116)$$

C'est un problème d'optimisation quadratique sous contraintes linéaire. Ainsi *quadprog*/*ipop*.

Soit α^* la solution. La condition *KKT* devient ici

$$\alpha_i^* [y_i(x_i^T \beta_0^* + \beta_0^*) - 1] = 0 \quad (117)$$

Si $\alpha_2^* > 0$,

$$\Rightarrow y_i(x_i^T \beta^* + \beta_0^*) = 1 \quad (118)$$

\Rightarrow la contrainte est saturée $\Rightarrow x_i$ est sur la marge et est un vecteur de support.

Posons

$$S = \{i | \alpha_i^* > 0\} \quad (119)$$

L'ensemble des i associé aux vecteurs de support. Inversement, si x_i n'est pas un vecteur de support alors,

$$y_i(x_i^T \beta + \beta_0) > 1 \quad (120)$$

et donc $\alpha_i^* = 0$.

Une fois résolu le problème dual, on en déduit la solution (β^*, β_0^*) du problème initial.

En effet, on a

$$\beta^* = \sum_{i=1}^n \alpha_i^* y_i x_i = \sum_{i \in S} \alpha_i^* y_i x_i \quad (121)$$

Par ailleurs, pour tout $i \in S$,

$$y_i(x_i^T \beta^* + \beta_0^*) = 1 \Rightarrow \beta_0^* = \frac{1}{y_i} - x_i^T \beta^* = y_i - x_i^T \beta^* \quad (122)$$

Remarque importante

La solution du problème initial s'exprime donc uniquement en fonction des support vectors(SV).

Conclusion L'HSO a donc pour équation

$$\begin{aligned} f(x) &= x^T \beta^* + \beta_0^* \\ &= \sum_{i=1}^n \alpha_i^* y_i x_i^T x + \beta_0^* \\ &= \sum_{i \in S} \alpha_i^* y_i x_i^T x + \beta_0^* \end{aligned} \quad (123)$$

La fonction de décision correspondante est

$$d(x) = \text{sign}(f^*(x)) \quad (124)$$

Remarque fondamentale

La fonction de décision et le problème d'optimisation ne s'exprime qu'à travers de produits scalaires entre observations

4.2 SVM Linéaire cas non-linéairement séparable

Dans le cas non linéairement séparable, l'HSO doit établir un compromis entre largeur de marge "raisonnable" et taux d'erreur. La solution consiste à autoriser certains individus à avoir une marge plus petite que 1 voire négative. Les contraintes doivent être retranchés de façon douce en introduisant des variables ressorts

$\epsilon_i, i = 1, \dots, n$. On va remplacer

$$y_i(x_i^T \beta + \beta_0) > 1 \quad (125)$$

par

$$y_i(x_i^T \beta + \beta_0) > 1 - \epsilon_i \quad (126)$$

On définit donc le problème d'optimisation suivant

$$\min \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^n \epsilon_i \quad (127)$$

sc

$$\{ y_i(x_i^T \beta + \beta_0) \geq 1 - \epsilon_i \quad i = 1, \dots, n \quad \epsilon_i \geq 0 \quad i = 1, \dots, n \quad (128)$$

Remarque 1: Le paramètre γ est un paramètre de régularisation permettant d'établir un compromis entre largeur de marge "raisonnable" et taux d'erreur.

Remarque 2:

$$y_i(x_i^T \beta + \beta_0) \geq 1 - \epsilon_i \Leftrightarrow \epsilon_i \geq 1 - y_i(x_i^T \beta + \beta_0) \quad (129)$$

Dans ce cas, le Lagrangian s'écrit

$$L(\beta, \beta_0, \epsilon, \alpha, \mu) = \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^n \epsilon_i - \sum_{i=1}^n \alpha_i (y_i(x_i^T \beta + \beta_0) - (1 - \epsilon_i)) - \sum_{i=1}^n \mu_i \epsilon_i \quad (130)$$

En annulant la dérivée du Lagrangian par rapport aux β, β_0 et ϵ_i

$$\frac{\partial L}{\partial \beta} = \beta - \sum_{i=1}^n \alpha_i y_i x_i = 0 \quad (131)$$

$$\frac{\partial L}{\partial \beta_0} = - \sum_{i=1}^n \alpha_i y_i = 0 \quad (132)$$

$$\frac{\partial L}{\partial \epsilon_i} = \gamma - \alpha_i - \mu_i = 0, i = 1, \dots, n \quad (133)$$

En développant cette formule et en injectant les constantes, on obtient

$$L(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j \quad (134)$$

même expression que dans le cas linéairement séparabl.

Le problème d'optimisation général des SVM consiste à maximiser () sous les contraintes:

$$\left\{ \begin{array}{ll} \alpha_i \geq 0 & i = 1, \dots, n \\ \mu_i \geq 0 & i = 1, \dots, n \\ \mu_i = \gamma - \alpha_i & i = 1, \dots, n \\ \sum_{i=1}^n \alpha_i y_i = 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} 0 \leq \alpha_i \leq \gamma \\ \sum_{i=1}^n \alpha_i y_i = 0 \end{array} \right. \quad (135)$$

En résumé, SVM est défini par le problème d'optimisation suivant

$$\text{whatever} \quad (136)$$

Soit α^* la solution de (), les conditions KKT se ramène ici KKT1

$$\alpha_i^* [y_i (x_i^T \beta^* + \beta_0^*) - (1 - \epsilon_i)] = 0 \quad (137)$$

KKT 2:

$$\mu_i^* \epsilon_i^* = 0 \quad (138)$$

Les SV correspondent aux x_i associés à des $\alpha_i^* > 0$. Ils vérifient

$$y_i (x_i^T \beta^* + \beta_0^*) = 1 - \epsilon_i^* \quad (139)$$

Les SV sont soit les x_i mal classés ($1 - \epsilon_i < 0$), soit les x_i à l'intérieur de la marge.

Si de plus $\mu_i^* > 0$ ($0 < \alpha_i < \gamma$), alors, $\epsilon_i = 0$ et dans ce cas, les SV x_i sont sur la marge.

Posons $S = \{i | \alpha_i^* > 0\}$ l'ensemble des indices associées aux SV.

Une fois résolu le problème dual

$$\beta^* = \sum_{i=1}^n \alpha_i^* y_i x_i = \sum_{i \in S} \alpha_i^* y_i x_i \quad (140)$$

Par ailleurs, pour obtenir β_0^* , il suffit de considérer un SV sur la marge ($0 < \alpha < \gamma$) puis d'utiliser KKT 1

$$\beta_0^* = \frac{1}{y_i} - x_i^T \beta^* = y_i - x_i^T \beta^* \quad (141)$$

La fonction de décision est la même que précédemment

$$f^*(x) = \sum_{i \in S} \alpha_i^* y_i x_i^T x + \beta_0^* \quad (142)$$

Rightarrow

$$d^*(u) = \text{sign}(f^*(x)) \quad (143)$$

4.3 SVM comme une problématique de fonction de coût pénalisé

$$\min_{\beta} \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^n \quad (144)$$

sc

$$y_i(x_i^T \beta + \beta_0) \geq 1 - \epsilon_i, i = 1, \dots, n \quad (145)$$

et

$$\epsilon_i \geq 0, i = 1, \dots, n \quad (146)$$

\Leftrightarrow

$$\min_{\beta} \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^n \epsilon_i \quad (147)$$

sc

$$\epsilon_i \geq 1 - y_i(x_i^T \beta + \beta_0) \quad (148)$$

$$\epsilon_i \geq 0 \quad (149)$$

\Leftrightarrow

$$\min_{\beta} \frac{1}{2} + \gamma \sum_{i=1}^n \max(0, 1 - y_i(x_i^T \beta + \beta_0)) \quad (150)$$

\Leftrightarrow

$$\min_{\beta} \sum_{i=1}^n \max(0, 1 - y_i(x_i^T \beta + \beta_0)) + \lambda \|\beta\|^2 \quad (151)$$

Premier terme Hinge Loss

Si l'on souhaite étendre les SVM linéaire à non-linéaire, il suffit d'appliquer le kernel trick.

Les problème d'optimisation associé au SVM non-linéaire est le suivant

$$\max_{\alpha} \sum_{i=1}^n \quad (152)$$