Data Mining

DM - Résumé

I. Apprentissage non supervisé

1. Mesure de distance entre clusters

Plus proche voisin :

$$D(\mathcal{C}_1, \mathcal{C}_2) = \min\{D(x_i, x_j), x_i \in \mathcal{C}_1, x_j \in \mathcal{C}_2\}$$

Distance des centres de gravité :

$$D(\mathcal{C}_1, \mathcal{C}_2) = D(\mu_1, \mu_2)$$

Distance moyenne :

$$D(\mathcal{C}_1, \mathcal{C}_2) = \frac{\sum_{x_i \in \mathcal{C}_1} \sum_{x_j \in \mathcal{C}_2} D(x_i, x_j)}{n_1 n_2}$$

Diamètre maximum :

$$D(\mathcal{C}_1, \mathcal{C}_2) = \max\{D(x_i, x_j), x_i \in \mathcal{C}_1, x_j \in \mathcal{C}_2\}$$

Distance de Ward :

$$D(\mathcal{C}_1, \mathcal{C}_2) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D(\mu_1, \mu_2)$$

2. Qualité d'un clustering

Inertie intra-cluster (within) :

Variance des points dans leur cluster

$$J_k = \sum_{x_i \in \mathcal{C}_k} D^2(x_i, \mu_k) \qquad J_w = \sum J_k$$

Inertie inter-cluster (between) :

Eloignement des centres de gravité

$$J_b = \sum N_k D^2(\mu_k, \mu)$$

II. Apprentissage supervisé

1. Décision Bayésienne

a. Définitions et formules

$\{\mathcal{C}_1, \dots, \mathcal{C}_K\}$
Classes

$\mathcal{X} = \mathbb{R}^d$
Espace des caractéristiques

$\mathcal{A} = \{a_1, \dots, a_m\}$
Ensemble des actions

$\ell_{jk} = l(a_j, \mathcal{C}_k)$
Coût de l'action a_j pour une observation de $\mathcal{C}_k$

$\mathbb{P}(\mathcal{C}_k)$
Proba. à priori

$\mathbb{P}(\mathcal{C} = \mathcal{C}_k | X = x) = \frac{\mathbb{P}(x|\mathcal{C}_k)\mathbb{P}(\mathcal{C}_k)}{\mathbb{P}(x)}$
Proba. à posteriori

$\mathbb{P}(X = x | \mathcal{C} = \mathcal{C}_k)$
Loi conditionnelle de x à $\mathcal{C}_k$

$\mathbb{P}_X(x) = \sum_{k=1}^{K} \mathbb{P}(x|\mathcal{C}_k)\mathbb{P}(\mathcal{C}_k)$
Loi marginale de x

$R(a_j|x) = \sum_{k=1}^{K} l_{jk} \mathbb{P}(\mathcal{C}_k|x)$
Risque conditionnel de l'action a_j pour x

$R_{moy}(D) = \int_{\mathcal{X}} R(D(x)|x) \mathbb{P}_X(x)\, dx$
Risque moyen d'une règle de décision

$$D : \begin{array}{ccc} \mathcal{X} & \to & \mathcal{A} \\ x & \mapsto & \underset{j=1..m}{\operatorname{argmin}} R(a_j|x) \end{array}$$
Règle de décision de bayes

b. Discrimination entre 2 classes

$$\mathcal{L}(x) = \frac{\mathbb{P}(x|\mathcal{C}_1)}{\mathbb{P}(x|\mathcal{C}_2)} \qquad D(x) = \begin{cases} a_1 & \text{si} \quad \mathcal{L}(x) \geq \eta \\ a_2 & \text{si} \quad \mathcal{L}(x) < \eta \end{cases} \quad \text{avec} \quad \eta = \frac{(\ell_{12} - \ell_{22})\mathbb{P}(\mathcal{C}_2)}{(\ell_{21} - \ell_{11})\mathbb{P}(\mathcal{C}_1)}$$

c. Classification binaire à coût 0-1 et rejet

	Classification binaire	Classification binaire avec rejet (a_3)
Risques	$R(a_1\|x) = 1 - \mathbb{P}(\mathcal{C}_1\|x)$ $R(a_2\|x) = 1 - \mathbb{P}(\mathcal{C}_2\|x)$	$R(a_1\|x) = \mathbb{P}(\mathcal{C}_2\|x)$ $R(a_2\|x) = \mathbb{P}(\mathcal{C}_1\|x)$ $R(a_3\|x) = \alpha$
Décision	$D(x) = \begin{cases} a_1 \text{ si } \mathbb{P}(\mathcal{C}_1\|x) > \frac{1}{2} \\ a_2 \text{ sinon} \end{cases}$	$D(x) = \begin{cases} a_1 \text{ si } \mathbb{P}(\mathcal{C}_1\|x) > \mathbb{P}(\mathcal{C}_2\|x) \text{ et } 1 - \alpha \\ a_2 \text{ si } \mathbb{P}(\mathcal{C}_3\|x) > \mathbb{P}(\mathcal{C}_1\|x) \text{ et } 1 - \alpha \\ a_3 \text{ sinon} \end{cases}$ $\alpha \in [0; 0{,}5] \Leftrightarrow [100\% ; 0\%]$ rejet

d. Cas gaussien et fonctions discriminantes

i. Matrices de covariance identiques : LDA

$$g_j(x) = \ln \mathbb{P}(\mathcal{C}_j|x)\mathbb{P}(x) = w_j^\top x + w_{j_0} \quad w_j = \Sigma^{-1}\mu_j \quad w_{j_0} = \ln \mathbb{P}(\mathcal{C}_j) - \frac{1}{2}\mu_j^\top \Sigma^{-1}\mu_j$$

ii. Matrices de covariance différentes : QDA

$$g_j(x) = x^\top W_j^\top x + w_j^\top x + w_{j_0} \quad W_j = \frac{1}{2}\Sigma_j^{-1}\mu_j \quad w_j = \Sigma_j^{-1}\mu_j \quad w_{j_0} = \frac{1}{2}\mu_j^\top \Sigma_j^{-1}\mu_j - \ln|\Sigma_j| + \ln \mathbb{P}(\mathcal{C}_j)$$

III. Régression logistique

$$\log \frac{\mathbb{P}(\mathcal{C}_1|x)}{\mathbb{P}(\mathcal{C}_2|x)} = [1 \quad x_i^\top]\theta = \phi_i^\top \theta \Rightarrow p_i = \mathbb{P}(\mathcal{C}_1|x) = \frac{e^{\phi_i^\top \theta}}{1 + e^{\phi_i^\top \theta}} \quad \mathbb{P}(\mathcal{C}_2|x) = 1 - p_i \quad z_i = \begin{cases} 1 \Leftrightarrow \mathcal{C}_1 \\ 0 \Leftrightarrow \mathcal{C}_2 \end{cases}$$

$$\tilde{\mathcal{L}} = \sum_{i=1}^{N} z_i \log p_i + (1 - z_i)\log(1 - z_i) = \sum_{i=1}^{N} z_i \phi_i^\top \theta - \log\left(1 + e^{\phi_i^\top \theta}\right) \quad \max \tilde{\mathcal{L}} \Leftrightarrow \min J = -\tilde{\mathcal{L}}$$

IV. SVM

$x_i \in \mathbb{R}^p \quad y_i \in \{-1; 1\}$ On veut une frontière de décision linéaire $f(x) = w^\top x + b$

La distance du point x au plan H séparant les données ($f(x) = 0$) est :

$$d(x, H) = \frac{w^\top(x - x_0)}{\|w\|} = \frac{w^\top x - w^\top x_0}{\|w\|} = \frac{w^\top x + b}{\|w\|}$$

On veut $y_i f(x_i) = |f(x_i)| > 1 \Rightarrow f(x) \in]-\infty; -1[\cup]1; +\infty[$ La marge est $M = \frac{2}{\|w\|}$

$$\begin{array}{lll} \min & \frac{1}{2}\|w\|^2 & \text{maximisation de la marge} \\ \text{s.c.} & y_i f(x_i) \geq 1 & \text{points bien classés} \end{array} \Leftrightarrow \begin{array}{ll} \max & \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \\ \text{s.c.} & \alpha_i \geq 0 \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \end{array}$$

Les vecteurs supports sont ceux tel que $y_i f(x_i) = 1$

b se détermine grâce aux vecteurs supports. $M = \sqrt{\frac{1}{\sum \alpha_i}}$

Cas non séparable :

On rajoute une variable de relâchement ξ_i et un coefficient C de pénalité

$$\begin{array}{ll} \min & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{s.c.} & y_i f(x_i) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{array} \Leftrightarrow \begin{array}{ll} \max & \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \\ \text{s.c.} & 0 \leq \alpha_i \leq C \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \end{array}$$

$\alpha_i = C$ pour les points mal classés.

V. Dérivées

1. Dérivée première

a. Gradient

$$\nabla J(x_0) = \left[\frac{\partial J}{\partial x_1}, \dots, \frac{\partial J}{\partial x_n}\right]^\top$$

Propriété : Au point x_0, $\nabla J(x_0)$ est $\perp$ à la ligne de niveau, son sens va dans le sens de J croissant.

b. Dérivée directionnelle

$$\boxed{D_x J(x,d) = \lim_{\varepsilon \to 0} \frac{J(x+\varepsilon d) - J(x)}{\varepsilon} = \left.\frac{d\,\overbrace{J(x+\varepsilon d)}^{\varphi(\varepsilon)}}{d\varepsilon}\right|_{\varepsilon=0} = \varphi'(0) = \nabla J(x)^\top d}$$

c. Règles de calcul pour la dérivée

$D(J_1 + \alpha J_2) = DJ_1 + \alpha J_2$ | $D\big(J_1(J_2)\big) = D_z J_1(z = J_2)\, D_x J_2(x)$ | $\nabla\, a^\top x = a$ | $\nabla\, x^\top A x = 2Ax$

2. Dérivées secondes

a. Dérivée directionnelle au sens de Gâteaux

$$D^2 J(x,d) = \lim_{\varepsilon \to 0} \frac{DJ_x(x+ed) - DJ_x(x)}{\varepsilon}$$

b. Matrice Hessienne

$$[H_J(x)]_{ij} = \frac{\partial^2 J(x)}{\partial x_i \partial x_j}$$

Calcul partique : A partir de la dérivée de $\varphi(\varepsilon) = D_x J(x + \varepsilon d)$, identifier $\varphi'(0) = d^\top H_X(x)^\top d$

c. Développements limités

$$J(x+d) \approx J(x) + \varepsilon\,\nabla J(x)^\top d \approx J(x) + \nabla_x J(x)^\top d + \frac{1}{2} d^\top H d$$

VI. Généralités sur l'optimisation

p variables, m contraintes d'égalité, n contraintes d'inégalité

Contraintes C	$V \mapsto \mathbb{R}^{m+n}$ $\quad \begin{cases} h_i(x) = 0 \\ g_j(x) \le 0 \end{cases} \Leftrightarrow \begin{cases} H(x) = (h_1, \dots, h_n)^\top = 0 \\ G(x) = (g_1, \dots, g_m)^\top \le 0 \end{cases}$
Domaine de faisabilité Ω	$\Omega = \{x \in \mathbb{R}^d\,; H(x) = 0 \text{ et } G(x) \le 0\}$
Fonction coût	$J : \Omega \mapsto \mathbb{R}$
Domaine de la fonction coût	$\operatorname{dom} J = \{x \in \Omega\,; -\infty < J(\theta) < +\infty\}$ *Si $\emptyset$, fonction coût impropre : pas de solution*
Minimum global θ^*	$J(\theta^*) \le J(\theta) \quad \forall\, \theta$
Minimum local $\hat{\theta}$	$J(\hat{\theta}) \le J(\theta) \quad \forall\, \theta \mid \lVert\hat{\theta} - \theta\rVert \le \varepsilon$

VII. Optimisation convexe sans contraintes

1. Condition d'optimalité

a. Existence de solution

J est coercivité	$\|x\| \to \infty \Rightarrow J(x) \to \infty$ (fonction infinie à l'infini)
J est propre	$\exists\, x \mid J(x) \in \mathbb{R}$
$\exists$ solution globale	Si J continue, propre, coercive, $\min_x J(x)$ admet une solution globale

b. Conditions d'optimalité

1er ordre	x_0 solution $\Rightarrow \nabla J(x_0) = 0$
2ème ordre	x_0 solution $\Rightarrow \nabla J(x_0) = 0$ et $H(x_0)$ définie positive
	$\nabla J(x_0) = 0$ et $H(x_0)$ définie positive $\Rightarrow$ x_0 solution locale
Convexité	J est convexe et x_0 respecte condition d'optimalité $\Rightarrow x_0 = x^*$ solution globale

2. Optimisation itérative

$$x_{k+1} = x_k + \rho_k d_k \qquad \lim_{k\to\infty} x_k = x^* = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} J(x)$$

i. Direction de descente d $\quad (\nabla J^\top d < 0)$

- **Gradient :** $d_k = -\nabla J$ $\quad \mathcal{O}(n)$
- **Gradient conjugué :** $d_k = -\nabla J + \beta_k d_{k-1}$ $\quad \mathcal{O}(n^2)$
- **Quasi-Newton :** $d_k = -B\nabla J$ $\quad$ (Ex : $B = diag(H)^{-1}$) $\quad \mathcal{O}(n^2)$
- **Newton :** $d_k = -H^{-1}\nabla J$ $\quad \mathcal{O}(n^3)$

VIII. Optimisation convexe sous contraintes

1. Avec contraintes d'égalités

Problème	Lagrangien	Conditions d'optimalité
$\begin{cases}\min\limits_{x\in\mathbb{R}^n} J(x)\\ s.c.\, H(x) = 0\end{cases}$	$\boxed{L(x,\lambda) = J(x) + \sum_{j=1}^{p} \lambda_j H_j(x)}$ *λ_j multiplicateurs de Lagrange*	$\nabla_x L = 0 \;\Rightarrow\; \nabla_x J(x) + \sum_{j=1}^{p} \lambda_i \nabla_x H_j(x) = 0$ $\nabla_{\lambda_j} L = 0 \;\Rightarrow\; H_j(x) = 0$

2. Avec contraintes d'inégalités

a. Lagrangien et conditions d'optimalité

Problème	Lagrangien	Conditions d'optimalité KKT	
$\begin{cases}\min\limits_{x\in\mathbb{R}^n} J(x)\\ \text{s.c.}\quad H(x) =\\ \text{et}\quad G(x) \leq\end{cases}$	$\boxed{L(x,\lambda,\mu) = J(x) + \sum_{j=1}^{p} \lambda_j H_j(x) + \sum_{i=1}^{q} \mu_i G_i(}$ *λ_j multiplicateurs de Lagrange*	Stationarité :	$\nabla L(x,\lambda) = 0$
		Adm. primale :	$G(x) \leq 0$ $H(x) = 0$
		Adm. duale :	$\mu_i \geq 0$
		Complémentarité :	$\mu_i g_i(x) = 0$

b. Problème dual

Le problème dual est $\boxed{\mathcal{L}(\mu,\lambda) = \min_x L(z,\lambda,\mu)}$