1 Rappels

Il n'y a que quelques points clefs à retenir de ce cours. On les liste informellement ci-dessous. Des justifications et des exemples plus détaillés sont proposés plus loin.

Regression linéaire. On s'intéresse à résoudre un problème de regression, c'est à trouver f tel que $f(x) \approx y$. Dans le cas de la regression linéaire, on cherche des fonctions de la forme $f_\theta(x) = x^T\theta$. On dispose d'une base de données (x_i, y_i), et on minimise l'erreur moyenne sur ces données

$$\min_\theta \frac{1}{n}\sum_{k=1}^{n} \|f_\theta(x_i) - y_i\|_2^2$$

Dans ce cours on suppose de plus que y a une dépendance linéaire bruitée à x :

$$y_i = x_i^T\theta^* + \varepsilon_i, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)\ i.i.d.$$

Pour simplifier l'étude, on peut reformuler ce problème avec des notations matricielles X, Y : chaque ligne correspond à un échantillon x_i ou y_i. On gère l'ordonnée à l'origine soit en centrant les vecteurs soit en fixant la première coordonnée de chaque échantillon $x_{i,1} = 1$. Le problème s'écrit alors

$$\min_\theta \|X\theta - Y\|_2^2.$$

Ordinary Least Square. Lorsque X est de rang plein (pour ses colonnes), le problème admet une solution unique

$$\hat{\theta} = (X^TX)^{-1}X^TY$$

Ridge regression. Lorsque X n'est pas de rang plein, on peut rajouter une régularisation $\mathcal{L}_2$ qui rend le problème soluble

$$\min_\theta \|X\theta - Y\|_2^2 + \lambda\|\theta\|_2^2,$$

$$\hat{\theta} = (X^TX + \lambda I)^{-1}X^TY.$$

Lasso regression. Si on sait que seules certaines coordonnées des échantillons x_i sont utiles pour prédire y_i, on peut faire de la sélection de variables. Une façon simple est d'utiliser la régularisation $\mathcal{L}_1$, qui force la plupart des coordonnées de θ a être nulles

$$\min_\theta \|X\theta - Y\|_2^2 + \lambda\|\theta\|_1.$$

p-value. On peut vouloir tester certaines hypothèses sur la valeur de θ^* (par exemple tester si l'ordonnée à l'origine $\theta_1 = 0$). Notons cette hypothèse $\mathcal{H}_0$. Un outil souvent utilisé est la p-value : on construit une variable aléatoire T telle que

- on peut déterminer la loi de T sous l'hypothèse $\mathcal{H}_0$ et évaluer sa fonction cumulative $\mathbb{P}(T \leq t)$;

- on peut évaluer T sur nos échantillons (x_i, y_i) (notons t son évaluation).

On définit alors p-value $= \mathbb{P}(|T| \geq |t|)$ (Cf. https://stackoverflow.com/questions/28921661/p-value-significance-level-and-hypothesis, 2ème post pour une explication intuitive).

Intervalle de confiance. On est toujours incertain de la valeur exacte de θ^*, mais on peut vouloir controler sa dispersion. Un outil standard est l'intervalle de confiance : on construit des variables aléatoires A, B telles qu'avec haute probabilité $1 - \alpha$

$$\mathbb{P}(A \leq \theta^* \leq B) = 1 - \alpha.$$

2 Pré-requis

Ce cours suppose quelques pré-requis en algèbre linéaire, en probabilité, et en calcul différentiel, dont on rappelle certains ci-dessous. Si ces résultats ne paraissent pas évidents, et que l'on ne sait pas les retrouver et les démontrer rapidement, des sites comme https://math.stackexchange.com/ ou https://mathworld.wolfram.com/ sont des sources d'informations complémentaires utiles.

- **Linéarité de** $\mathbb{E}$. $\mathbb{E}(AX) = A\mathbb{E}(X)$, $\mathbb{E}(XA) = \mathbb{E}(X)A$, $\mathbb{E}(X + A) = \mathbb{E}(X) + A$
- **Covariance**. $\mathrm{cov}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^\top\right] = \left(\mathrm{cov}\left(x_i, x_j\right)\right)_{i,j}$
- $\mathbb{V}(aX + b) = a^2 V(X)$, $covV(AX + B) = Acov(X)A^T$
- **Loi normale**. $x \sim N(0,1) \Rightarrow \sigma x + \mu \sim N\left(\mu, \sigma^2\right)$
- $x \sim N\left(\mu, \sigma^2\right) \Rightarrow (x - \mu)/\sigma \sim N(0,1)$
- $X_1 \sim N\left(\mu_1, \sigma_1^2\right), X_2 \sim N\left(\mu_2, \sigma_2^2\right)$ independent $\Rightarrow X_1 + X_2 \sim N\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right)$
- **Noyau**. $Ker(A) = 0 \Leftrightarrow A$ est inversible
- $A \in \mathbb{R}^{np}$, $rang(A) = p$ alors A est injective : $Ker(A) = \{0\}$.
- **Transposition**. $(A^T)^T = A \quad (AB)^T = B^\top A^\top \quad (A + B)^T = A^\top + B^\top$
- D diagonale $\Rightarrow D^T = D$
- A symétrique inversible $\Rightarrow A^{-1}$ symétrique.
- $X^T X$ est symétrique positive (i.e. symétrique et à valeurs propres positives).
- **Valeurs propres**. A est inversible ssi ses valeurs propres sont non nulles.
- Si on note $vp(A)$ l'ensemble des valeurs propres de A, on a $vp(A + \lambda I) = \lambda + vp(A)$

- **Décomposition en valeurs singulières (SVD)**. $A \in \mathbb{R}^{np} \Rightarrow \exists U \in \mathbb{R}^{nn}\ \exists V \in \mathbb{R}^{pp}$ orthonormales et $\exists \Sigma \in \mathbb{R}^{np}$ diagonale telles que $A = U\Sigma V$.
- **Produit scalaire**. $(a|b) = a^T b, \quad \|a\|^2 = a^T a, \quad |(a|b)| \leq \|a\|\|b\|, \quad \|a\| = 0 \Rightarrow a = 0$
- **Convexité**. $f : \mathbb{R}^p \mapsto \mathbb{R}^n$ et $\nabla\nabla f \in \mathbb{R}^{pp}$ symétrique positive $\Rightarrow$ f convexe.
- **Gradient**. $\nabla_x(a^T x) = a \quad \nabla_x(x^T Ax) = (A^T + A)x$ en général $\quad \nabla_x(x^T Ax) = 2Ax$ si A est symétrique.

3 Applications

On propose la solution de quelques questions du quizz 2018/1029 comme application des deux sections précédentes. Cette sélection couvre la plupart des notions importantes du cours. On détaille chaque question le plus exhaustivement possible.

2) **What is the orthogonal projection of a vector $\mathbf{x} \in \mathbb{R}^n$ over** $\mathrm{Vect}\,(1_n)$**, where** $1_n = (1, \ldots, 1)^\top \in \mathbb{R}^n$ **?**

Par definition, si F est un sous espace vectoriel de E et $x \in E$, le projeté orthogonal sur F est défini par $p_F(x) = argmin_{y \in F}\|x - y\|^2$. Si $F = Vect(u)$, alors $p_{Vect(u)}(x) = argmin_\lambda \|x - \lambda u\|^2 u = \frac{(u|x)}{\|u\|^2} u$.

Justifions cette dernière égalité brièvement.

$$f(\lambda) = \|x - \lambda u\|^2 = (x - \lambda u)^T (x - \lambda u) = x^T x - 2\lambda u^T x + \lambda^2 u^T u = \|x\|^2 - 2\lambda(u|x) + \lambda^2\|u\|^2$$

Si λ est un minimum de f alors $\nabla_\lambda f(\lambda) = 0$. Or $\nabla_\lambda f(\lambda) = -2(u|x) + 2\lambda\|u\|^2 = 0 \Leftrightarrow \lambda = \frac{(u|x)}{\|u\|^2}$. On conclut avec la définition du projeté orthogonal donnée ci-dessus.

En prenant $u = 1_n$, on a $p_{Vect(1_n)}(x) = \frac{1}{n}(1_n^T x)1_n$.

7) **What is the solution of** $\left\{ \begin{array}{l} \max_{u \in \mathbb{R}^n, v \in \mathbb{R}^p} u^\top X v \\ \textbf{s.c. } \|u\|_2^2 = 1 \textbf{ et } \|v\|_2^2 = 1 \end{array} \right.$ **?**

L'astuce est d'utiliser la décomposition en valeurs singulières de X : $X = U^T \Sigma V$.

$$u^T X v = u^T U^T \Sigma V v = (Uu)^T \Sigma (Vv)$$

$u \mapsto Uu$ et $v \mapsto Vv$ forment des bijections sur $\{u \in \mathbb{R}^{nn} \| \|u\| = 1\}$ et $\{v \in \mathbb{R}^{pp} \| \|v\| = 1\}$. Donc

$$\max_{\|u\|=1, \|v\|=1} u^T X v = \max_{\|u\|=1, \|v\|=1} u^T \Sigma v = \max_{\|u\|=1, \|v\|=1} \sum_i u_i v_i \sigma_i.$$

Si les valeurs propres σ_i sont ordonnées par ordre décroissant, on voit qu'en prenant $u_1 = v_1 = 1$ et $u_i = v_i = 0$ ailleurs, on obtient $\sum_i u_i v_i \sigma_i = \sigma_1 = \max_i \sigma_i$. D'autre part $\|u\| = \|v\| = 1$ avec ces définitions.

Il est clair qu'on obtiendra jamais une valeur plus élevée. Pour le justifier rigoureusement, on peut par exemple remarquer que si $\|u\| = \|v\| = 1$,

$$\sum_i u_i v_i \sigma_i \leq \sum_i |u_i v_i| \sigma_i \leq (\max_i \sigma_i) \sum_i |u_i v_i| \leq (\max_i \sigma_i) \|u\| \|v\| = (\max_i \sigma_i).$$

8) **Let $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$ be real numbers. Is the following function convex or concave?**

$$(\theta_0, \theta_1) \mapsto \frac{1}{2} \sum_{i=1}^{n} \left(y_i + 3\theta_0 - \theta_1 x_i\right)^2 .$$

En écrivant ce problème de façon matricielle, on a $f(x) = \|Y - X\theta\|_2^2$, avec $X = \begin{pmatrix} -3 & x_1 \\ \ldots & \ldots \\ -3 & x_n \end{pmatrix}$.
Comme f est différentiable (forme quadratique), on peut la dériver pour déterminer si elle est convexe.

$$\begin{aligned} f(\theta) &= \|Y - X\theta\|_2^2 = (Y - X\theta)^T (Y - X\theta) \qquad \text{(par définition)} \\ &= Y^T Y - (X\theta)^T Y - Y^T X\theta + (X\theta)^T (X\theta) \qquad \text{(car } (A+B)^T = A^T + B^T\text{)} \\ &= Y^T - \theta^T X^T Y - Y^T X\theta + \theta^T X^T X\theta \qquad \text{(car } (AB)^T = B^T A^T\text{)} \\ &= Y^T Y - 2(Y^T X)\theta + \theta^T (X^T X)\theta \qquad \text{(car } Y^T X\theta = (Y^T X\theta)^T = \theta^T X^T \theta \in \mathbb{R}\text{)} \end{aligned}$$

$$\begin{aligned} \nabla_\theta f(\theta) &= -2(Y^T X)^T + 2(X^T X)\theta \qquad \text{(car } \nabla_x (a^T x) = a \text{ et } \nabla_x (x^T S x) = 2S \text{ si } S \text{ est symétrique)} \\ &= -2X^T Y + 2X^T X\theta \end{aligned}$$

$$\nabla_\theta \nabla_\theta f(\theta) = 2X^T X$$

Or $X^T X$ est une matrice symétrique positive. On en déduit que f est convexe (cf. rappel sur la convexité).

Vérifions que $X^T X$ est bien une matrice symétrique. Il faut montrer qu'elle est symétrique et que ses valeurs propres sont positives. La transposition est involutive, donc

$$(X^T X)^T = X^T (X^T)^T = X^T X$$

D'autre part si λ une valeur propre de $X^T X$ et u un vecteur propre associé. L'astuce est d'introduire des normes (dont on sait qu'elles sont positives). On a par définition

$$X^T X u = \lambda u \Rightarrow u^T X^T X u = \lambda u^T u \Rightarrow (Xu)^T (Xu) = \lambda u^T u \Rightarrow \|Xu\|^2 = \lambda \|u\|^2$$

$\|Xu\|$ et $\|u\|$ sont strictement positifs (car u est non nul par définition des valeurs propres) donc λ est aussi positif. Ceci pour tout λ valeur propre de $X^T X$.

10) **For any $X \in \mathbb{R}^{n \times p}$ express $\mathrm{Ker}\left(X^\top X\right)$ in terms of $\mathrm{Ker}(X)$.**

On va montrer ce résultat par double inclusion.

$$u \in Ker(X) \Rightarrow Xu = 0 \Rightarrow X^T Xu = 0 \Rightarrow u \in Ker(X^T X)$$

Donc $Ker(X) \subset Ker(X^T X)$. Pour l'inclusion inverse, l'astuce est à nouveau d'introduire une norme (car elle vérifie $\|a\| = 0 \Rightarrow a = 0$).

$$u \in Ker(X^T X) \Rightarrow X^T Xu = 0 \Rightarrow u^T X^T Xu = 0 \Rightarrow \|Xu\|^2 = 0 \Rightarrow u \in Ker(X)$$

Donc $Ker(X^T X) \subset Ker(X)$.

15) **For $X_1, \dots, X_n$ i.i.d. with values in $\{0, 1\}$, propose a procedure to test the hypothesis $p = P(X_1 = 1) = 1/2$.**

C'est pour tester ce genre d'hypothèses que la p-value est un outil intéressant. Par définition, on cherche à construire une variable aléatoire T dont on connait la loi, et que l'on peut évaluer sur nos données. La p-value est alors définie par $\mathbb{P}(|T| \geq |t|)$.

On peut par exemple prendre

$$T = \frac{\sum_{i=1}^n X_i - \mathbb{E}X}{\sqrt{\mathbb{V}X}} = \frac{\sum_{i=1}^n X_i - \frac{1}{2}n}{\sqrt{n * \frac{1}{2} * (1 - \frac{1}{2})}} \approx \mathcal{N}(0, 1)$$

d'après le théorème central limite.

16) **In the regression model, assuming that X is deterministic and that $\varepsilon = \mathbf{y} - X\boldsymbol{\theta}^\star$ is a Gaussian, centered, with covariance matrix $\sigma^2 \mathrm{Id}_n$ where σ^2 is known, what is the distribution of $\hat{\theta}_n$ (one could assume that X is full column rank here). Based on this, provide a confidence interval for $(1, \dots, 1)\hat{\theta}_n$.**

Le raisonnement pour construire un intervalle de confiance est toujours le même. On cherche A, B tel que $\mathbb{P}\left(A \leqslant 1^\top \theta^* \leqslant B\right) = 1$

Idée 1: $\hat{\theta} = \theta^* + (X^T X)^{-1} X^T \varepsilon \Rightarrow 1_n^T \hat{\theta}$ suit une loi normale (combinaison linéaire de loi normales i.i.d). On aura ensuite

$$\frac{1^T \hat{\theta} - \mathbb{E}1^T \hat{\theta}}{\sqrt{\mathbb{V}1^T \hat{\theta}}} \sim \mathcal{N}(0, 1)$$

Calculons l'espérance de $\hat{\theta}$.

$$\mathbb{E}\left(1^\top \hat{\theta}\right) = 1^\top \theta^*$$

Calculons maintenant sa variance.

$$
\begin{aligned}
1^T\hat{\theta} - \mathbb{E}1^\top\hat{\theta} &= 1^T(X^TX)^{-1}X^T\varepsilon \\
\mathbb{V}\hat{\theta} &= \mathbb{E}\left(1^\top\hat{\theta} - \mathbb{E}1^\top\hat{\theta}\right)\left(1^\top\hat{\theta} - \mathbb{E}1^\top\hat{\theta}\right)^\top \quad \text{(par définition)} \\
&= \mathbb{E}1^\top\left(X^\top X\right)^{-1}X^\top\varepsilon\varepsilon^\top X\left(X^\top X\right)^{-1}1 \quad \text{(par substitution)} \\
&= 1^\top\left(X^\top X\right)^{-1}X^\top\mathbb{E}(\varepsilon\varepsilon^\top)X\left(X^\top X\right)^{-1}1 \quad \text{(linéarité de } \mathbb{E}) \\
&= \sigma^2 1^T\left(X^\top X\right)^{-1}1 \quad (\text{car } \mathbb{E}\varepsilon\varepsilon^T = \sigma^2 I)
\end{aligned}
$$

Idée 2 : Si $Z \sim \mathcal{N}(0,1)$ alors en notant $F : x \mapsto \mathbb{P}(Z \leq x)$

$$
\begin{aligned}
&\mathbb{P}(-a \leqslant Z \leqslant a) = 1-\alpha \\
&\Leftrightarrow 1 - \mathbb{P}(|Z| \geq a) = 1-\alpha \\
&\Leftrightarrow 1 - 2\mathbb{P}(Z \geq a) = 1-\alpha \quad \text{(car Z a une distribution symétrique)} \\
&\Leftrightarrow \mathbb{P}(Z \leq -a) = 1 - \frac{1}{2}\alpha \\
&\Leftrightarrow a = F^{-1}\left(1-\frac{1}{2}\alpha\right)
\end{aligned}
$$

Idée 3 : Remarquons d'autre part

$$
\begin{aligned}
&-a \leqslant \frac{1^\top\hat{\theta} - 1^\top\theta^*}{\sqrt{\sigma^2 1^T\left(X^\top X\right)^{-1}1}} \leqslant a \\
&\Leftrightarrow -a\sigma\sqrt{\cdots} \leqslant 1^\top\hat{\theta} - 1^\top\theta^* \leqslant a\sigma\sqrt{\cdots} \\
&\Leftrightarrow 1^\top\hat{\theta} - a\sigma\sqrt{\cdots} \leqslant 1^\top\theta^* \leqslant 1^\top\hat{\theta} + a\sigma\sqrt{\cdots}
\end{aligned}
$$

Donc en substituant, on a bien $\mathbb{P}(A \leq 1^T\hat{\theta} \leq B) = 1-\alpha$ si on choisit

$$
\begin{aligned}
A &= 1^\top\hat{\theta} - F^{-1}\left(1-\frac{1}{2}\alpha\right)\sigma\sqrt{1^T\left(X^\top X\right)^{-1}1}, \\
B &= 1^\top\hat{\theta} + F^{-1}\left(1-\frac{1}{2}\alpha\right)\sigma\sqrt{1^T\left(X^\top X\right)^{-1}1}.
\end{aligned}
$$