Feuille de TD 5 – Correction

Bases du ML et de la classification

Exercice 1 – Régression logistique

On pose

$$\sigma(u) = \frac{e^u}{1+e^u}, \qquad p_i(\beta) = \sigma(\langle X_i, \beta\rangle),$$

de sorte que, conditionnellement à X_i, Y_i suit une loi de Bernoulli de paramètre $p_i(\beta^\star)$. La fonction ℓ_n est l'*opposée* de la log-vraisemblance conditionnelle : l'estimateur du maximum de vraisemblance, lorsqu'il existe, est donc un minimiseur de ℓ_n.

1. **Gradient, Hessienne et convexité.**

   Correction. Pour tout i,
   $$\nabla_\beta \log(1 + e^{\langle X_i,\beta\rangle}) = p_i(\beta) X_i.$$
   Par conséquent,
   $$\nabla \ell_n(\beta) = \sum_{i=1}^n (p_i(\beta) - Y_i) X_i.$$
   Comme $\sigma'(u) = \sigma(u)(1-\sigma(u))$, on obtient
   $$H_n(\beta) = \nabla^2 \ell_n(\beta) = \sum_{i=1}^n p_i(\beta)(1-p_i(\beta)) X_i X_i^\mathsf{T}.$$
   En particulier, pour tout $u \in \mathbb{R}^d$,
   $$u^\mathsf{T} H_n(\beta) u = \sum_{i=1}^n p_i(\beta)(1-p_i(\beta)) \langle X_i, u\rangle^2 \geq 0.$$
   Ainsi ℓ_n est convexe. Si $H_n(\beta)$ est non singulière, elle est définie positive et ℓ_n est strictement convexe (sur le domaine où cette hypothèse vaut). Elle admet alors au plus un minimiseur.

   Il faut distinguer unicité et existence : dans un échantillon complètement séparable, l'infimum de la perte logistique peut être approché lorsque $\|\beta\| \to \infty$ sans être atteint. Sous l'hypothèse supplémentaire d'existence de $\widehat{\beta}$, la stricte convexité assure son unicité et $\nabla \ell_n(\widehat{\beta}) = 0$.

2. **Développement du score autour de $\beta^\star$.**

   Correction. Posons $\Delta_n = \widehat{\beta} - \beta^\star$ et appliquons le théorème fondamental de l'analyse à l'application vectorielle $\nabla \ell_n$ le long du segment joignant $\beta^\star$ à $\widehat{\beta}$:
   $$0 = \nabla \ell_n(\widehat{\beta}) = \nabla \ell_n(\beta^\star) + \left\{ \int_0^1 H_n(\beta^\star + t\Delta_n)\, dt \right\} \Delta_n.$$
   En notant
   $$\overline{H}_n = \int_0^1 H_n(\beta^\star + t\Delta_n)\, dt,$$
   on a donc, dès que $\overline{H}_n$ est inversible,
   $$\widehat{\beta} - \beta^\star = -\overline{H}_n^{-1} \nabla \ell_n(\beta^\star).$$

Précaution sur l'énoncé. En dimension $d = 1$, le théorème des accroissements finis permet d'écrire $\overline{H}_n = H_n(\widetilde{\beta})$ pour un $\widetilde{\beta}$ du segment, et donc $|\widetilde{\beta} - \beta^\star| \le |\widehat{\beta} - \beta^\star|$. En dimension $d > 1$, il n'existe en général pas un unique point $\widetilde{\beta}$ réalisant simultanément cette identité matricielle. La formule avec la Hessienne moyenne $\overline{H}_n$ est la version vectorielle exacte. Elle suffit pour toute la suite.

3. **Fonction génératrice du score (optionnel).**

Correction. Il est utile de préciser le conditionnement. On note $\mathcal{X}_n = \sigma(X_1, \dots, X_n)$. Conditionnellement à $\mathcal{X}_n$, les Y_i sont indépendantes et $Y_i \sim \text{Bernoulli}(p_i^\star)$, où $p_i^\star = p_i(\beta^\star)$. Puisque

$$-\frac{1}{\sqrt{n}} \nabla \ell_n(\beta^\star) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i - p_i^\star) X_i,$$

on obtient, pour $u_i = \langle t, X_i \rangle / \sqrt{n}$,

$$\begin{aligned} &\mathbb{E}\left[\exp\left(-\frac{1}{\sqrt{n}} \langle t, \nabla \ell_n(\beta^\star) \rangle \right) \middle| \mathcal{X}_n \right] \\ &\qquad = \prod_{i=1}^{n} \left(1 - p_i^\star + p_i^\star e^{u_i}\right) e^{-p_i^\star u_i}. \end{aligned}$$

Pour u proche de 0,

$$\log(1 - p + pe^u) - pu = \frac{1}{2} p(1-p) u^2 + O(|u|^3).$$

Les X_i étant uniformément bornés, $|u_i| = O(n^{-1/2})$ uniformément en i, et la somme des restes est $nO(n^{-3/2}) = O(n^{-1/2})$. Ainsi

$$\begin{aligned} &\mathbb{E}\left[\exp\left(-\frac{1}{\sqrt{n}} \langle t, \nabla \ell_n(\beta^\star) \rangle \right) \middle| \mathcal{X}_n \right] \\ &\qquad = \exp\left(\frac{1}{2} t^\mathsf{T} \left(\frac{1}{n} H_n(\beta^\star) \right) t + O(n^{-1/2}) \right). \end{aligned}$$

Le conditionnement sur les covariables explique pourquoi les $p_i^\star$ et les X_i peuvent être traités comme des constantes dans le produit.

4. **Lois asymptotiques.**

Correction. La question précédente et la convergence $n^{-1} H_n(\beta^\star) \to H(\beta^\star)$ donnent

$$-\frac{1}{\sqrt{n}} \nabla \ell_n(\beta^\star) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, H(\beta^\star)).$$

Dans le cas de covariables aléatoires i.i.d.,

$$H(\beta^\star) = \mathbb{E}\left[p_{\beta^\star}(X)(1 - p_{\beta^\star}(X)) X X^\mathsf{T} \right].$$

Comme $\widehat{\beta} \to \beta^\star$ presque sûrement, tout le segment entre $\widehat{\beta}$ et $\beta^\star$ finit par appartenir à la boule où la convergence de $n^{-1} H_n$ est uniforme. Par continuité de H,

$$\frac{1}{n} \overline{H}_n \longrightarrow H(\beta^\star) \qquad \text{p.s.}$$

La matrice limite étant non singulière, $\overline{H}_n$ est inversible pour n assez grand. La formule de la question 2 et le théorème de Slutsky donnent alors

$$\sqrt{n}(\widehat{\beta} - \beta^\star) = \left(\frac{1}{n}\overline{H}_n\right)^{-1}\left(-\frac{1}{\sqrt{n}}\nabla\ell_n(\beta^\star)\right) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, H(\beta^\star)^{-1}).$$

Dans ce modèle correctement spécifié, la covariance du score est égale à l'information de Fisher, d'où la forme simple $H(\beta^\star)^{-1}$.

5. **Intervalle de confiance pour une coordonnée.**

Correction. L'information observée $H_n(\widehat{\beta})$ est d'ordre n et $H_n(\widehat{\beta})^{-1}$ estime donc directement la covariance asymptotique de $\widehat{\beta}$. Si $z_{1-\alpha/2}$ désigne le quantile d'ordre $1-\alpha/2$ de $\mathcal{N}(0,1)$, un intervalle de confiance asymptotique pour $\beta_j^\star$ est

$$\mathcal{I}_{n,\alpha}^{(j)} = \left[\widehat{\beta}_j - z_{1-\alpha/2}\sqrt{[H_n(\widehat{\beta})^{-1}]_{jj}},\ \widehat{\beta}_j + z_{1-\alpha/2}\sqrt{[H_n(\widehat{\beta})^{-1}]_{jj}}\right].$$

Sa probabilité de couverture tend vers $1-\alpha$.

6. **Ellipsoïde de confiance.**

Correction. La forme quadratique associée à la loi normale limite converge vers une loi du χ^2 à d degrés de liberté :

$$(\widehat{\beta} - \beta^\star)^\top H_n(\widehat{\beta})(\widehat{\beta} - \beta^\star) \xrightarrow{\mathcal{L}} \chi^2_d.$$

En notant $\chi^2_{d,1-\alpha}$ son quantile d'ordre $1-\alpha$, on peut donc prendre

$$\mathcal{E}_{n,\alpha} = \left\{\beta \in \mathbb{R}^d :\ (\beta - \widehat{\beta})^\top H_n(\widehat{\beta})(\beta - \widehat{\beta}) \le \chi^2_{d,1-\alpha}\right\}.$$

Alors $\mathbb{P}(\beta^\star \in \mathcal{E}_{n,\alpha}) \to 1-\alpha$.

7. **Sélection de variables.**

Correction. Deux méthodes classiques sont les suivantes.

- *Tests de Wald.* Pour chaque coordonnée, on teste $H_{0,j} : \beta_j^\star = 0$ à l'aide de

$$Z_j = \frac{\widehat{\beta}_j}{\sqrt{[H_n(\widehat{\beta})^{-1}]_{jj}}}.$$

  On retient les variables dont l'intervalle de confiance n'inclut pas 0. En cas de nombreux tests simultanés, il faut corriger la multiplicité (par exemple par Bonferroni ou par une procédure FDR).

- *Régression logistique pénalisée par le lasso.* On résout

$$\widehat{\beta}_\lambda^{\text{lasso}} \in \operatorname{argmin}_{\beta\in\mathbb{R}^d}\{\ell_n(\beta) + \lambda\|\beta\|_1\}.$$

  La pénalité ℓ^1 force certaines coordonnées à être exactement nulles; λ est typiquement choisi par validation croisée.

Ces procédures répondent à des objectifs différents : les tests quantifient une incertitude sous un modèle fixé, tandis que le lasso vise surtout la prédiction et une représentation parcimonieuse.

Exercice 2 – Non-consistance du plus proche voisin

On distingue soigneusement trois sources d'aléa : l'échantillon d'apprentissage $\mathcal{D}_n$, la covariable X d'un nouvel individu et son label Y. Le couple test (X, Y) est indépendant de $\mathcal{D}_n$. Dans le modèle considéré,

$$\mathbb{P}(Y = 1 \mid X = x) = \alpha > \frac{1}{2} \qquad \text{pour tout } x \in [0,1]^d.$$

1. **Classifieur de Bayes.**

   Correction. Puisque $\eta(x) = \alpha > 1/2$ pour tout x,

   $$g^\star(x) = 1 \quad \text{pour tout } x \in [0,1]^d.$$

   Autrement dit, la covariable X ne contient ici aucune information sur la classe : la meilleure règle prédit toujours la classe la plus probable.

2. **Erreur conditionnelle d'un classifieur quelconque.**

   Correction. Fixons x. Conditionnellement à $\mathcal{D}_n$ et à $X = x$, la valeur $g_n(x)$ est connue. Comme le label test est indépendant de $\mathcal{D}_n$ sachant $X = x$,

   $$\begin{aligned}\mathbb{P}(g_n(X) \neq Y \mid \mathcal{D}_n, X = x) &= (1-\alpha)\mathbf{1}_{\{g_n(x)=1\}} + \alpha \mathbf{1}_{\{g_n(x)=0\}} \\ &= \alpha - (2\alpha - 1) g_n(x).\end{aligned}$$

   On moyenne maintenant par rapport à l'échantillon d'apprentissage :

   $$\mathbb{P}(g_n(X) \neq Y \mid X = x) = \alpha - (2\alpha - 1)\mathbb{E}\big[g_n(X) \mid X = x\big].$$

   Ici, l'espérance du membre de droite porte sur l'aléa de $\mathcal{D}_n$. Pour $g^\star(x) = 1$, l'erreur vaut

   $$\mathbb{P}(g^\star(X) \neq Y \mid X = x) = 1 - \alpha.$$

   Elle ne dépend pas de x, donc $\mathcal{R}(g^\star) = 1 - \alpha$.

3. **Écriture du classifieur 1-NN.**

   Correction. Soit $I_n(x)$ l'indice du point d'apprentissage le plus proche de x. Comme la loi des X_i est continue, les ex aequo ont probabilité nulle; on peut aussi fixer une règle déterministe pour les départager. Alors

   $$B_i(x) = \mathbf{1}_{\{I_n(x)=i\}}, \qquad g_{1,n}(x) = \sum_{i=1}^n B_i(x) Y_i.$$

   Exactement un indice est sélectionné, d'où

   $$\sum_{i=1}^n B_i(x) = 1 \quad \text{p.s.}$$

4. **Espérance et risque du 1-NN.**

Correction. Notons encore $\mathcal{X}_n = \sigma(X_1, \dots, X_n)$. Pour x fixé, les $B_i(x)$ sont $\mathcal{X}_n$-mesurables. De plus, puisque $\mathbb{P}(Y_i = 1 \mid X_i) = \alpha$ est constante, on a

$$\mathbb{E}(Y_i \mid \mathcal{X}_n) = \mathbb{E}(Y_i \mid X_i) = \alpha.$$

La propriété de sortie des quantités mesurables dans l'espérance conditionnelle donne

$$\begin{aligned}\mathbb{E}[g_{1,n}(x) \mid \mathcal{X}_n] &= \sum_{i=1}^{n} B_i(x)\mathbb{E}(Y_i \mid \mathcal{X}_n) \\ &= \alpha \sum_{i=1}^{n} B_i(x) = \alpha.\end{aligned}$$

En utilisant ensuite la tour des espérances,

$$\mathbb{E}[g_{1,n}(x)] = \alpha.$$

La formule de la question 2 entraîne, pour tout x,

$$\begin{aligned}\mathbb{P}(g_{1,n}(X) \neq Y \mid X = x) &= \alpha - (2\alpha - 1)\alpha \\ &= 2\alpha(1 - \alpha).\end{aligned}$$

En intégrant par rapport à la loi de X,

$$\mathcal{R}(g_{1,n}) = 2\alpha(1 - \alpha).$$

Ce risque ne dépend ni de n ni de la dimension d.

5. **Non-consistance.**

Correction. Le risque du 1-NN reste strictement supérieur au risque de Bayes :

$$\begin{aligned}\mathcal{R}(g_{1,n}) - \mathcal{R}(g^\star) &= 2\alpha(1 - \alpha) - (1 - \alpha) \\ &= (1 - \alpha)(2\alpha - 1) > 0.\end{aligned}$$

Ainsi, pour tout n,

$$\mathcal{R}(g_{1,n}) = \mathcal{R}(g^\star) + (1 - \alpha)(2\alpha - 1),$$

et cet écart ne tend pas vers zéro. Par conséquent, la suite $(g_{1,n})_n$ n'est pas consistante.

6. **Pourquoi la localisation ne suffit-elle pas ?**

Correction. Lorsque n augmente, le plus proche voisin de x se rapproche bien de x. Cela réduit une éventuelle erreur de *localisation* : le voisin possède une covariable de plus en plus semblable à celle du point à classer. Mais son label reste une seule réalisation de Bernoulli(α). Même si le voisin était exactement en x, son label serait encore erroné avec probabilité $1 - \alpha$; ce bruit ne se réduit donc pas avec n.

Pour le réduire, il faut moyenner les labels de plusieurs voisins et classifier par vote majoritaire. Les deux conditions

$$k_n \longrightarrow \infty \qquad \text{et} \qquad \frac{k_n}{n} \longrightarrow 0$$

réalisent le compromis nécessaire : $k_n \to \infty$ permet de moyenner le bruit des labels, tandis que $k_n/n \to 0$ garantit que le voisinage utilisé reste local. Sous ces conditions, le classifieur k_n-NN est universellement consistant.