21 juin 2021.

Pour aller à l'essentiel

- Quelques questions sont assez techniques (calculs, optimisation). Le choix vous est donné d'admettre les résultats ou de les démontrer. Pendant la PC, je vous recommande de les admettre afin de pouvoir vous concentrer sur les aspects directement liés au cours de science des données.
- Le but de cette PC est d'illustrer les principes de minimisation du risque empirique, maximisation de la vraisemblance, et régularisation avec deux algorithmes de classification : la **régression logistique** et les **machines à vecteurs de support** (ou **SVM**). Ces deux méthodes sont implémentées dans `scikit-learn`.

1 Régression logistique

Nous considérons ici un problème de classification binaire en dimension p : nous disposons d'un jeu d'apprentissage $\mathcal{D} = \{(\vec{x}^{\,i},\ y^i)\}_{i=1,\dots,n}$ composé de n individus étiquetés $(\vec{x}^{\,i},\ y^i) \in \mathbb{R}^{p+1} \times \{0,1\}$.

Nous considérons ici $\vec{x} \in \mathbb{R}^{p+1}$, après avoir ajouté un 1 à gauche d'un vecteur p-dimensionnel, afin de simplifier les notations vectorielles et matricielles comme dans la section 7.6.2 du poly : $\beta_0 + \sum_{j=1}^{p} \beta_j x_j$ peut alors être noté $\langle \vec{\beta},\ \vec{x} \rangle$.

On appelle **fonction logistique** (à ne pas confondre avec la *fonction de coût logistique* de la section 7.4.2 du poly) la fonction

$$\begin{aligned} \sigma : \mathbb{R} &\to [0,1] \\ u &\mapsto \frac{1}{1+e^{-u}}. \end{aligned}$$

Son graphe est représenté sur la figure 1. Cette fonction est dérivable et sa dérivée vérifie (vous pouvez le vérifier)

$$\sigma'(u) = \sigma(u)(1-\sigma(u)) \text{ en tout point } u \in \mathbb{R}. \tag{1}$$

1.1 Minimisation du risque empirique

1. Pourquoi un modèle paramétrique linéaire, c'est-à-dire de la forme $f : \vec{x} \mapsto \langle \vec{\beta},\ \vec{x} \rangle$, n'est-il pas approprié pour un problème de classification binaire ?

On pourrait utiliser un modèle linéaire comme *fonction de décision* : $f(\vec{x}) \geq 0$ conduit à prédire une étiquette positive, et $f(\vec{x}) < 0$ conduit à prédire une étiquette négative.

Dans le cas de la **régression logistique**, on préfère utiliser comme fonction de décision la composition d'une fonction linéaire et de la fonction logistique :

$$f(\vec{x}) = \sigma(\langle \vec{\beta},\ \vec{x} \rangle). \tag{2}$$

2. Comment peut-on alors interpréter $f(\vec{x})$? Prêtez attention à l'espace d'arrivée de σ.

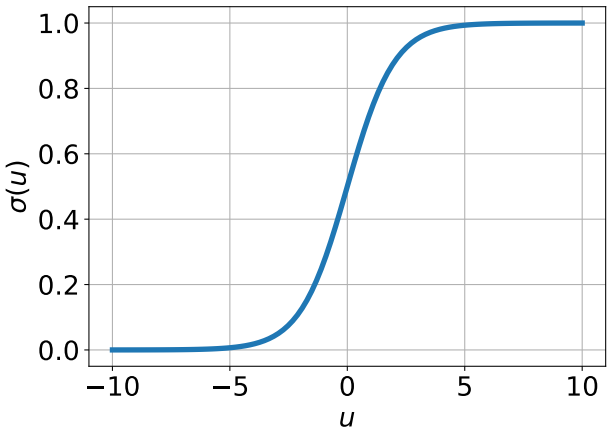


FIGURE 1 – Graphe de la fonction logistique

3. Utiliser cet espace des hypothèses et la fonction de coût logistique (définie à la section 7.4.2 du poly) pour poser l'apprentissage d'un classifieur binaire sous la forme de la minimisation d'un risque empirique.
4. Montrer ou admettre que le risque empirique est convexe. Admet-il un minimum global ?
5. Comment minimiser le risque empirique ? On pourra montrer ou admettre que le gradient du risque empirique en $\vec{\beta}$ vaut

$$\nabla_{\vec{\beta}} R_n = -\frac{1}{n}\sum_{i=1}^{n}\left(y^i - \frac{1}{1+e^{-\langle\vec{\beta},\,\vec{x}^{\,i}\rangle}}\right)\vec{x}^{\,i}.$$

Pour le calculer, on pourra poser $\sigma_i = \sigma(\langle\vec{\beta},\ \vec{x}^{\,i}\rangle)$ et commencer par exprimer $\nabla_{\vec{\beta}}\sigma_i$ en fonction de $\vec{x}^{\,i}$ et σ_i.

1.2 Formulation probabiliste

Nous considérons maintenant que notre jeu d'apprentissage est la réalisation de l'échantillon aléatoire $((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$, constitué de n copies i.i.d. de (X, Y). Ici X est un vecteur aléatoire à valeurs dans $\mathbb{R}^{p+1}$ et Y une variable aléatoire discrète à valeurs dans $\{0, 1\}$. $\vec{\beta} \in \mathbb{R}^{p+1}$ est maintenant un paramètre à estimer.

Vraisemblance Nous avons jusqu'à présent défini la vraisemblance uniquement pour une variable aléatoire à densité ou pour une variable aléatoire discrète (voir `erratum_estimation.pdf`). Cette définition peut être étendue à un vecteur aléatoire réel Z dont certaines composantes, notées U, sont à densité et les autres, notées V, sont discrètes, de la façon suivante. On note g la densité du vecteur aléatoire à densité U. Une réalisation $\vec{z}$ de Z peut être décomposée comme $(\vec{u},\ \vec{v})$, avec $\vec{u}$ la composante à densité et $\vec{v}$ la composante discrète. Alors la vraisemblance d'un échantillon $((\vec{u}^{\,1}, \vec{v}^{\,1}), (\vec{u}^{\,1}, \vec{v}^{\,2}), \ldots, (\vec{u}^{\,n}, \vec{v}^{\,n}))$ de Z est définie par

$$L(\vec{z}^{\,1}, \vec{z}^{\,2}, \ldots, \vec{z}^{\,n}; \theta) = \prod_{i=1}^{n} \mathbb{P}(V = \vec{v}^{\,i} | U = \vec{u}^{\,i}) g(\vec{u}^{\,i}), \tag{3}$$

où g et $\mathbb{P}_{V|U=\vec{u}}$ peuvent toutes deux être paramétrées par θ.

1. Posons g_X la densité de X. Écrire la log-vraisemblance du jeu d'apprentissage $\mathcal{D}$ en fonction de $\mathbb{P}(Y = 1|X = \vec{x}^{\,i})$.

2. Dans cette log-vraisemblance, remplacer $\mathbb{P}(Y = 1|X = \vec{x}^{\,i})$ par sa valeur telle que modélisée dans la section 1.1. Qu'en conclure sur l'estimateur par maximum de vraisemblance ?

1.3 Régularisation

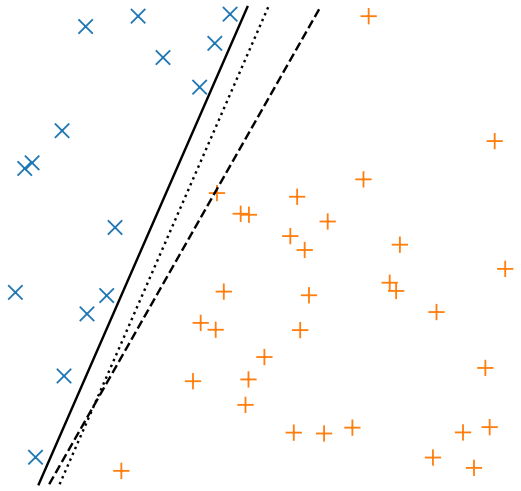
1. Écrire la version régularisée ℓ_2 de la minimisation du risque empirique proposée plus haut. Quel est l'effet de ce régulariseur sur le modèle appris ?
2. Même question pour la régularisation ℓ_1.

2 Machine à vecteurs de support

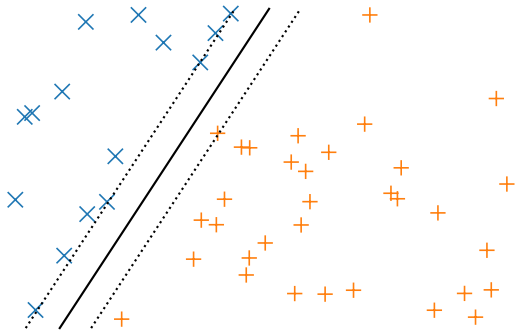
Nous considérons ici toujours un problème de classification binaire en dimension p, mais allons utiliser $\{-1, 1\}$ pour les étiquettes. Nous disposons d'un jeu d'apprentissage $\mathcal{D} = \{(\vec{x}^{\,i}, y^i)\}_{i=1,\dots,n}$ composé de n individus étiquetés $(\vec{x}^{\,i}, y^i) \in \mathbb{R}^p \times \{-1, 1\}$.

2.1 SVM à marge rigide

Nous supposons ici que les données sont linéairement séparables : il existe un hyperplan de $\mathbb{R}^p$ tel que tous les individus de la classe positive (étiquetés $+1$) soient d'un côté de cet hyperplan et tous les individus de la classe négative (étiquetés -1) de l'autre. Un tel exemple est illustré sur la figure 2a.



(A) Données linéairement séparables ($p = 2$) et 3 exemples d'hyperplan séparateur.



(B) Les droites en pointillés représentent les hyperplans parallèles à l'hyperplan séparateur, d'équations $\langle \vec{w},\ \vec{x} \rangle + b = \pm 1$.

1. Si nous posons $\vec{w} \in \mathbb{R}^p$, $b \in \mathbb{R}$ tels que $\langle \vec{w},\ \vec{x} \rangle + b = 0$ soit l'équation d'un tel hyperplan, quel est le signe de $y^i \left(\langle \vec{w},\ \vec{x}^{\,i} \rangle + b \right)$ pour $i = 1, \dots, n$?
2. Cet hyperplan fait donc office de modèle de classification. Quelle est l'équation de la fonction de décision du modèle ? Quel est le modèle de classification binaire correspondant ?
3. Nous allons maintenant définir la **marge** d'un tel classifieur : c'est la distance entre l'hyperplan $\langle \vec{w},\ \vec{x} \rangle + b = 0$ et le point de $\mathcal{D}$ qui en est le plus proche. Comparez les 3 hyperplans de la figure 2a : lequel a la plus petite marge ? La plus grande marge ?

4. Le principe des classifieurs à vaste marge (*large margin classifiers* en anglais) est de choisir, parmi plusieurs classifieurs possibles, celui qui a la plus grande marge. Voyez-vous pourquoi ?

Nous allons maintenant chercher à déterminer $\vec{w} \in \mathbb{R}^p$ et $b \in \mathbb{R}$ tels que l'hyperplan H d'équation $\langle \vec{w},\ \vec{x} \rangle + b = 0$ ait la plus grande marge possible.

Pour cela, nous allons poser définir deux hyperplans parallèles à H :

$$\begin{cases} H_- : \langle \vec{w},\ \vec{x} \rangle + b = -1 \\ H_+ : \langle \vec{w},\ \vec{x} \rangle + b = +1, \end{cases}$$

de sorte à ce que le(s) point(s) positif(s) le(s) plus proche(s) de H soit sur H_+ et que le(s) point(s) négatif(s) le(s) plus proche(s) de H soit sur H_-. Les valeurs ± 1 sont choisies sans perte de généralité, utiliser une constante $c > 0$ à la place de 1 reviendrait à diviser $\vec{w}$ et b par c. Ces hyperplans sont représentés en pointillés sur la figure 2b.

5. Cela signifie que H_- et H_+ sont à la même distance de H. Pourquoi cela est-il compatible avec l'idée de chercher un hyperplan H de marge maximale ?
6. La zone entre H_+ et H_- est parfois appelée « zone d'indécision ». Pourquoi ?
7. Les points situés *sur* H_+ et H_- sont appelés **vecteurs de support** et donnent leur nom à cette méthode : **machine à vecteurs de support** en français, **support vector machine (SVM)** en anglais. Voyez-vous d'où vient leur nom ? Pour comprendre, supposez que vous déplaciez un tel point d'une distance ϵ faible ; comment cela affecterait-il H, H_+ et H_- ? Même question pour un point situé loin de H_+ (ou H_-).
8. Quelle est la valeur de la marge ?
9. Les observations $\vec{x}^{\,i}$ étant situées à l'extérieur de la zone d'indécision, quelle est l'inégalité vérifiée par $y^i \langle \vec{w},\ \vec{x}^{\,i} \rangle + b$ pour $i = 1, \ldots, n$?
10. Poser le problème d'optimisation sous contraintes correspondant à maximiser la marge tout en assurant que l'inégalité de la question précédente est vraie pour $i = 1, \ldots, n$. Montrer qu'il est équivalent à

$$\underset{\vec{w} \in \mathbb{R}^p, b \in \mathbb{R}}{\arg\min} \frac{1}{2} \|\vec{w}\|_2^2 \text{ t.q. } y^i \left(\langle \vec{w}, \vec{x}^{\,i} \rangle + b \right) \geq 1, i = 1, \ldots, n. \tag{4}$$

11. Identifier la formulation (4) avec la minimisation d'un risque empirique régularisé : quel est l'espace des hypothèses ? Quelle est la fonction de perte ? Quel est le régulariseur ?
12. Montrer (ou admettre) que cette formulation est équivalente à

$$\max_{\vec{\alpha} \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{l=1}^n \alpha_i \alpha_l y^i y^l \langle \vec{x}^{\,i},\ \vec{x}^{\,l} \rangle$$
$$\text{t. q. } \sum_{i=1}^n \alpha_i y^i = 0;\ \alpha_i \geq 0, i = 1, \ldots, n,$$

et que si on appelle $(\vec{w}^{\,*}, b^*)$ un minimiseur du problème d'optimisation posé à la question précedente, et α^* un maximiseur du problème ci-dessus, alors :

$$\begin{cases} \vec{w}^{\,*} = \sum_{i=1}^n \alpha_i^* y^i \vec{x}^{\,i} \\ \alpha_i^* \left(y^i \left(\langle \vec{w}^{\,*},\ \vec{x}^{\,i} \rangle + b^* \right) - 1 \right) = 0 \quad \text{pour tout } i = 1, \ldots, n. \end{cases}$$

13. Que dire de la valeur de α_i^* pour un vecteur de support, par opposition à un autre point du jeu d'entraînement ? On partira de

$$\alpha_i^* \left(y^i \left(\langle \vec{w}^{\,*},\ \vec{x}^{\,i} \rangle + b^* \right) - 1 \right) = 0 \text{ pour tout } i = 1, \ldots, n.$$

2.2 Pour aller plus loin : SVM à marge souple

Dans le cas non-séparable, on utilise la fonction de perte dite *hinge*, définie par

$$L_{\text{hinge}} : \{-1, 1\} \times \mathbb{R} \to \mathbb{R}$$
$$y, f(\vec{x}) \mapsto \begin{cases} 0 & \text{si } yf(\vec{x}) \geq 1 \\ 1 - yf(\vec{x}) & \text{sinon.} \end{cases}$$

De manière plus compacte, la perte hinge peut aussi s'écrire

$$L_{\text{hinge}}(f(\vec{x}), y) = \max(0, 1 - yf(\vec{x})) = [1 - yf(\vec{x})]_+ .$$

La perte hinge est positive quand un point est situé du mauvais côté non pas de l'hyperplan séparateur H, mais de H_+ pour un point d'étiquette positive (respectivement, de H_- pour un point d'étiquette négative).

La SVM à marge souple est la solution du problème d'optimisation

$$\underset{\vec{w}\in\mathbb{R}^p, b\in\mathbb{R}}{\arg\min} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^{n} \left[1 - y^i f(\vec{x}^{\,i})\right]_+ . \tag{5}$$

1. Identifier la formulation (5) avec la minimisation d'un risque empirique régularisé.
2. En introduisant une variable d'ajustement (ou variable d'écart ; on parle de *slack variable* en anglais) $\xi_i = \left[1 - y^i f(\vec{x}^{\,i})\right]_+$ pour chaque observation du jeu d'entraînement, le problème d'optimisation 5 est équivalent à

$$\underset{\vec{w}\in\mathbb{R}^p, b\in\mathbb{R}, \vec{\xi}\in\mathbb{R}^n}{\arg\min} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^{n} \xi_i \tag{6}$$
$$\text{t. q. } \begin{cases} y^i\left(\langle \vec{w},\ \vec{x}^{\,i}\rangle + b\right) \geq 1 - \xi_i, \\ \xi_i \geq 0 \end{cases} \quad \text{pour tout } i = 1, \ldots, n \tag{7}$$

   Montrer en suivant la même démarche que pour la question 12 de la section précédente que le problème (6) est équivalent à :

$$\max_{\vec{\alpha}\in\mathbb{R}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{n} \alpha_i\alpha_l y^i y^l \langle \vec{x}^{\,i},\ \vec{x}^{\,l}\rangle \tag{8}$$
$$\text{t. q. } \sum_{i=1}^{n} \alpha_i y^i = 0 \text{ et } 0 \leq \alpha_i \leq C, \text{ pour tout } i = 1, \ldots, n.$$

   et que si on appelle $(\vec{w}^{\,*}, b^*)$ un minimiseur du problème (6), et α^* un maximiseur du problème ci-dessus, alors :

$$\begin{cases} \vec{w}^{\,*} = \sum_{i=1}^{n} \alpha_i^* y^i \vec{x}^{\,i} \\ \alpha_i^*\left(\langle \vec{w}^{\,*},\ \vec{x}^{\,i}\rangle + b^*\right) = 0 \\ (C - \alpha_i^*)\left[1 - y^i\left(\langle \vec{w}, \vec{x}^{\,i}\rangle + b\right)\right]_+ = 0 \quad \text{pour tout } i = 1, \ldots, n. \end{cases}$$

3. Que dire maintenant de la valeur de α_i^* pour un vecteur de support, par opposition à un autre point du jeu d'entraînement ? On partira de

$$\begin{cases} \alpha_i^*\left(\langle \vec{w}^{\,*},\ \vec{x}^{\,i}\rangle + b^*\right) = 0 \\ (C - \alpha_i^*)\left[1 - y^i\left(\langle \vec{w}, \vec{x}^{\,i}\rangle + b\right)\right]_+ = 0 \quad \text{pour tout } i = 1, \ldots, n. \end{cases}$$