10 juin 2020.

Pour aller à l'essentiel

- Quelques questions sont assez calculatoires et les calculs en question ne relèvent pas particulièrement du cours de science des données. 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 :
  - utilisation d'une classe d'hypothèse et d'une fonction de perte pour formuler un problème de minimisation du risque empirique ;
  - équivalence entre minimisation du risque empirique et maximisation de la vraisemblance ;
  - rôle de la régularisation.
- Remarquez que dans la partie 2.1, on ne construit pas la SVM à marge rigide comme un problème de minimisation du risque empirique ; la question 11 vous montre cependant comment l'interpréter ainsi, ce qui permet dans la partie 2.2 d'étendre cet algorithme au cas (plus réaliste) non-linéairement séparable.

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 ?

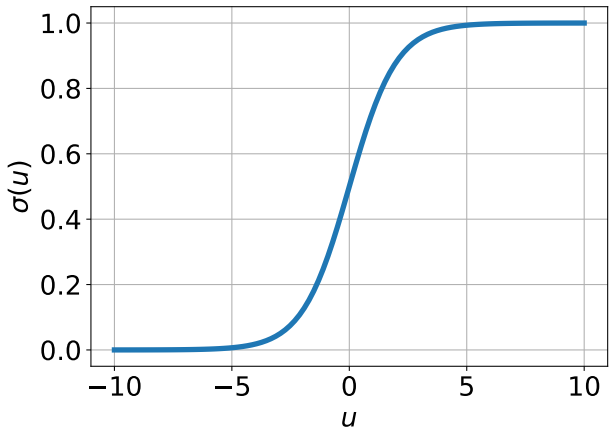


FIGURE 1 – Graphe de la fonction logistique

Solution Une telle fonction, appliquée à des variables prenant des valeurs réelles, va prendre non pas deux mais une infinité de valeurs. □

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.

Solution f modélise la *probabilité* que $\vec{x}$ appartienne à la classe positive. □

Nous considérons donc l'espace des hypothèses $\mathcal{F} = \{f : \vec{x} \mapsto \sigma(\langle \vec{\beta}, \vec{x} \rangle); \vec{\beta} \in \mathbb{R}^{p+1}\}$.

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.

Solution Nous cherchons $\vec{\beta}^*$ qui vérifie

$$\vec{\beta}^* \in \underset{\vec{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} L\left(\sigma(\langle \vec{\beta}, \vec{x}^{\,i} \rangle), y^i\right)$$

c'est-à-dire

$$\vec{\beta}^* \in \underset{\vec{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} -y^i \ln \sigma(\langle \vec{\beta}, \vec{x}^{\,i} \rangle) - (1 - y^i) \ln\left(1 - \sigma(\langle \vec{\beta}, \vec{x}^{\,i} \rangle)\right).$$

□

4. Montrer ou admettre que le risque empirique est convexe. Admet-il un minimum global ?

Solution

— Le risque empirique est convexe : on peut le réécrire comme

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

La convexité de $u \mapsto \ln(1+e^{-u})$ se montre en calculant sa dérivée seconde.

— Quand $\|\vec{\beta}\|_2 \to +\infty$, $R_n(\vec{\beta}) \to +\infty$ et R_n continue, donc R_n admet un minimum global (par le théorème 2 du poly d'optimisation). Ce minimum n'est cependant pas nécessairement atteint en un point unique (on peut avoir une infinité de solutions) . □

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.

Solution En posant $\sigma_i = \sigma(\langle \vec{\beta}, \vec{x}^{\,i}\rangle)$, et en déduisant de l'équation (1) que

$$\nabla_{\vec{\beta}}\sigma_i = \vec{x}^{\,i}\sigma_i(1-\sigma_i),$$

le gradient de R_n en $\vec{\beta}$ vaut

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

Il n'y a pas de moyen évident de trouver $\vec{\beta}$ tel que $\nabla_{\vec{\beta}} R_n = 0$. On recourra donc à un algorithme de gradient. □

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 $((\vec{u}^{\,1},\vec{v}^{\,1}),(\vec{u}^{\,1},\vec{v}^{\,2}),\dots,(\vec{u}^{\,n},\vec{v}^{\,n}))$ de Z est définie par

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

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 $\mathbb{P}(Y=1|X=\vec{x}^{\,i})$.

Solution Posons $((\vec{x}^{\,1},y^1),(\vec{x}^{\,2},y^2),\dots,(\vec{x}^{\,n},y^n))$ un échantillon de (X,Y). La vraisemblance de cet échantillon dépend de $\vec{\beta}$ et s'écrit

$$L((\vec{x}^{\,1},y^1),(\vec{x}^{\,2},y^2),\dots,(\vec{x}^{\,n},y^n);\vec{\beta})=\prod_{i=1}^{n}\mathbb{P}(Y=y^i|X=\vec{x}^{\,i})g_X(\vec{x}^{\,i}).$$

De plus,

$$\begin{aligned}\mathbb{P}(Y=y^i|X=\vec{x}^{\,i})&=\begin{cases}\mathbb{P}(Y=1|X=\vec{x}^{\,i}) & \text{si } y^i=1\\ 1-\mathbb{P}(Y=1|X=\vec{x}^{\,i}) & \text{si } y^i=0\end{cases}\\ &=\mathbb{P}(Y=1|X=\vec{x}^{\,i})^{y^i}(1-\mathbb{P}(Y=1|X=\vec{x}^{\,i}))^{1-y^i}.\end{aligned}$$

Donc

$$L((\vec{x}^{\,1},y^1),(\vec{x}^{\,2},y^2),\dots,(\vec{x}^{\,n},y^n);\vec{\beta})=\prod_{i=1}^{n}g_X(\vec{x}^{\,i})\mathbb{P}(Y=1|X=\vec{x}^{\,i})^{y^i}(1-\mathbb{P}(Y=1|X=\vec{x}^{\,i}))^{1-y^i}.$$

et la log-vraisemblance est

$$\ell((\vec{x}^{\,1},y^1),(\vec{x}^{\,2},y^2),\dots,(\vec{x}^{\,n},y^n);\vec{\beta})=\sum_{i=1}^{n}\ln g_X(\vec{x}^{\,i})+y^i\ln\mathbb{P}(Y=1|X=\vec{x}^{\,i})+(1-y^i)\ln(1-\mathbb{P}(Y=1|X=\vec{x}^{\,i})).$$

□

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 ?

Solution Voir la question 2 de la section 1.1 :

$$\mathbb{P}(Y=1|X=\vec{x}^{\,i})=\sigma(\langle\vec{\beta},\vec{x}^{\,i}\rangle).$$

La log-vraisemblance vaut :

$$\begin{aligned}&\ell((\vec{x}^{\,1},y^1),(\vec{x}^{\,2},y^2),\dots,(\vec{x}^{\,n},y^n);\vec{\beta})=\sum_{i=1}^{n}\ln g_X(\vec{x}^{\,i})+\sum_{i=1}^{n}\ln\sigma(\langle\vec{\beta},\vec{x}^{\,i}\rangle)^{y^i}(1-\sigma(\langle\vec{\beta},\vec{x}^{\,i}\rangle))^{1-y^i}\\ &=\sum_{i=1}^{n}\ln g_X(\vec{x}^{\,i})+\sum_{i=1}^{n}y^i\ln\sigma(\langle\vec{\beta},\vec{x}^{\,i}\rangle)+(1-y^i)\ln(1-\sigma(\langle\vec{\beta},\vec{x}^{\,i}\rangle))\end{aligned}$$

$g_X(\vec{x}^{\,i})$ ne dépend pas de $\vec{\beta}$. Maximiser la vraisemblance revient donc à maximiser

$$\sum_{i=1}^{n} y^i \ln \sigma(\langle \vec{\beta}, \vec{x}^{\,i} \rangle) + (1 - y^i) \ln(1 - \sigma(\langle \vec{\beta}, \vec{x}^{\,i} \rangle)),$$

ce qui est équivalent à minimiser le risque empirique calculé à la question 3 de la partie 1.1. □

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 ?

Solution Pour la régularisation ℓ_2 : Le problème d'optimisation devient

$$\vec{\beta}^* \in \underset{\vec{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} -y^i \ln \sigma(\langle \vec{\beta}, \vec{x}^{\,i} \rangle) - (1 - y^i) \ln \Big(1 - \sigma(\langle \vec{\beta}, \vec{x}^{\,i} \rangle)\Big) + \lambda \|\vec{\beta}\|_2^2.$$

L'effet de la régularisation ℓ_2 est le même que pour la régression linéaire : on contraint les coefficients à appartenir à une boule ℓ_2 et on évite le sur-apprentissage. □

2. Même question pour la régularisation ℓ_1.

Solution Le problème d'optimisation devient

$$\vec{\beta}^* \in \underset{\vec{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} -y^i \ln \sigma(\langle \vec{\beta}, \vec{x}^{\,i} \rangle) - (1 - y^i) \ln \Big(1 - \sigma(\langle \vec{\beta}, \vec{x}^{\,i} \rangle)\Big) + \lambda \|\vec{\beta}\|_1.$$

L'effet de la régularisation ℓ_1 est le même que pour la régression linéaire : on obtient une solution parcimonieuse. □

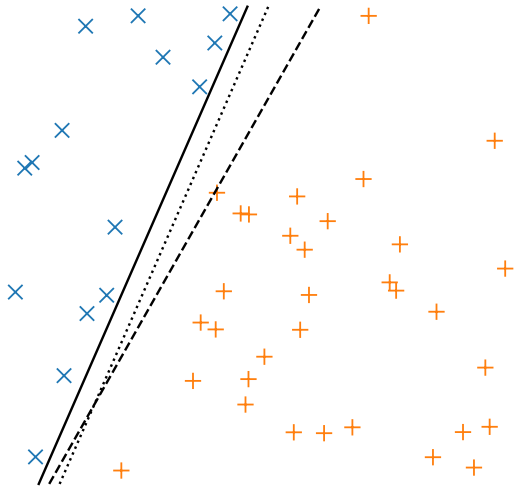
2 Machine à vecteurs de support

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 $(\vec{x}^{\,i}, y^i) \in \mathbb{R}^p \times \{-1,1\}$.

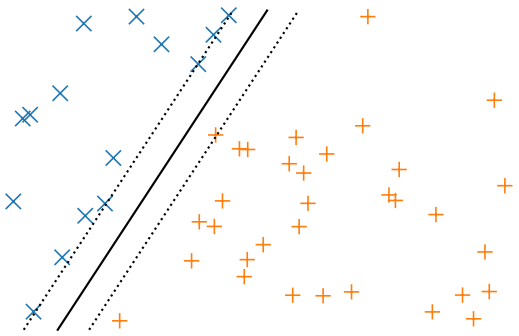
2.1 SVM à marge rigide

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.

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 $i = 1, \dots, n$?



(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 $\langle \vec{w}, \vec{x} \rangle + b = \pm 1$.

Solution D'un même côté de H, tous les y^i ont le même signe, de même que tous les $\langle \vec{w}, \vec{x}^{\,i} \rangle + b > 0$, donc leur produit est toujours de même signe.

On peut décider (à un signe près sur $\vec{w}$ et b) de fixer $\langle \vec{w}, \vec{x}^{\,i} \rangle + b > 0$ pour $y^i = 1$ et $\langle \vec{w}, \vec{x}^{\,i} \rangle + b < 0$ pour $y^i = -1$, donc $y^i \left(\langle \vec{w}, \vec{x}^{\,i} \rangle + b \right) > 0$. □

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 ?

Solution

— La fonction de décision est $\vec{x} \mapsto \langle \vec{w}, \vec{x} \rangle + b$.

— Le modèle est $\vec{x} \mapsto \operatorname{sgn}\left(\langle \vec{w}, \vec{x} \rangle + b \right)$. □

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 ?

Solution L'hyperplan représenté par la droite en traits interrompus (- - -) a la plus petite marge, puisqu'elle vaut 0 (un des + orange est dessus). L'hyperplan représenté par la droite en pointillés ($\cdots$) a la plus grande marge des 3. □

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 ?

Solution Comme nous n'avons aucune information dans la zone entre les $\times$ bleues et les $+$ orange, il n'y a aucune raison de préférer une fonction de décision qui favorise les uns ou les autres. Un hyperplan qui maximise la marge est ainsi en quelque sorte « au milieu » entre les $\times$ bleues et les $+$ orange. □

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 ?

Solution La marge vaut $\min\left(\text{dist}(H_-,H), \text{dist}(H_+,H)\right)$. Si H_- est plus près de H que H_+, la marge vaut $\text{dist}(H_-,H)$ et on peut déplacer H vers H_+ de sorte à diminuer $\text{dist}(H_+,H)$ et augmenter $\text{dist}(H_-,H)$. La marge est bien maximisée quand $\text{dist}(H_-,H) = \text{dist}(H_+,H)$. □

6. La zone entre H_+ et H_- est parfois appelée « zone d'indécision ». Pourquoi ?

Solution Cela revient à ce que nous avons observé question 4 : nous n'avons aucune information sur cette zone. □

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_-).

Solution Déplacer de ϵ un vecteur de support peut affecter la position de H ; déplacer de ϵ une observation qu iest loin de la zone d'indécision ne change rien à la solution. □

8. Quelle est la valeur de la marge ?

Solution La marge vaut $\frac{1}{\|\vec{w}\|_2}$.

Plus précisément, posons $\vec{u}$ un vecteur sur H et $\vec{v}$ sa projection sur H_-. La marge vaut $\|\vec{u} - \vec{v}\|_2$. $\vec{u}$ vérifie $\langle \vec{w},\vec{u} \rangle + b = 0$ et $\vec{v}$ vérifie $\langle \vec{w},\vec{v} \rangle + b = -1$. On a donc $\langle \vec{w}, \vec{u} - \vec{v} \rangle = 1$. De plus, $\vec{w}$ et $(\vec{u} - \vec{v})$ sont colinéaires ($\vec{w}$ étant un vecteur normal de H). Ainsi, $\|\vec{w}\|_2 \|\vec{u} - \vec{v}\|_2 = 1$. □

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 $i = 1, \ldots, n$?

Solution $y^i \langle \vec{w}, \vec{x}^{\,i} \rangle + b \geq 1$. □

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

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

Solution Maximiser $\frac{1}{\|\vec{w}\|_2}$ est équivalent à minimiser $\frac{1}{2}\|\vec{w}\|_2^2$. □

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 ?

Solution

— L'espace des hypothèses est l'ensemble des hyperplans de $\mathbb{R}^p$.

— Le risque empirique est forcé d'être nul ; la fonction de perte est, par exemple, la fonction de perte 0/1, et on a contraint toutes les observations de $\mathcal{D}$ à être correctement classifiées.

— $\frac{1}{2}\|\vec{w}\|_2^2$ est un régulariseur, comme dans le cas de la régression ridge. □

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 :

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

Solution Il s'agit de la formulation duale d'un problème d'optimisation convexe sous contraintes. Introduisons n multiplicateurs de Lagrange $\{\alpha_i\}_{i=1,\ldots,n}$, un pour chaque contrainte. Le lagrangien est donc la fonction

$$\begin{array}{rcl} L : \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}_+^n & \to & \mathbb{R} \\ \vec{w},b,\vec{\alpha} & \mapsto & \dfrac{1}{2}\|\vec{w}\|_2^2 - \displaystyle\sum_{i=1}^{n} \alpha_i \left(y^i \left(\langle \vec{w}, \vec{x}^{\,i}\rangle + b \right) - 1 \right). \end{array}$$

Le problème dual est donc

$$\max_{\vec{\alpha}\in\mathbb{R}_+^n} \inf_{\vec{w}\in\mathbb{R}^p, b\in\mathbb{R}} \frac{1}{2}\|\vec{w}\|_2^2 - \sum_{i=1}^{n} \alpha_i \left(y^i \left(\langle \vec{w}, \vec{x}^{\,i}\rangle + b \right) - 1 \right).$$

Le lagrangien est convexe en $\vec{w}$ et est donc minimal quand son gradient en $\vec{w}$ est nul, à savoir quand

$$\vec{w} = \sum_{i=1}^{n} \alpha_i y^i \vec{x}^{\,i}. \tag{5}$$

De plus, il est affine en b. Son infimum est donc $-\infty$, sauf si son gradient en b est nul (auquel cas la fonction affine est « plate »), à savoir si

$$\sum_{i=1}^{n} \alpha_i y^i = 0. \tag{6}$$

La fonction duale est donc maximisée dans ce deuxième cas.

En remplaçant $\vec{w}$ par sa valeur (eq. (5)) dans le dual, on obtient

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

On obtient le résultat recherché en utilisant l'équation (6).

Pour ce qui est de la relation entre $\vec{\alpha}^{\,*}$ et $(\vec{w}^{\,*}, b^*)$, il s'agit simplement des conditions de Karush-Kuhn-Tucker. □

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

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

Solution Pour chaque observation i, deux cas sont possibles :

- $\alpha_i^* = 0$: le minimiseur de $\vec{w} \mapsto \frac{1}{2}\|\vec{w}\|_2^2$ vérifie la contrainte et $y^i\langle \vec{w}^{\,*}, \vec{x}^{\,i}\rangle + b^* > 1$, autrement dit le point $\vec{x}^{\,i}$ est *à l'extérieur* des hyperplans H_+ ou H_- ;
- $\alpha_i^* > 0$: la contrainte est vérifiée en bordure de la zone de faisabilité, autrement dit à l'égalité $y^i\langle \vec{w}^{\,*}, \vec{x}^{\,i}\rangle + b^* = 1$, et $\vec{x}^{\,i}$ est un vecteur de support. □

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

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

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

$$\underset{\vec{w}\in\mathbb{R}^p, b\in\mathbb{R}}{\arg\min} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^{n}\left[1 - y^i f(\vec{x}^{\,i})\right]_+ . \tag{7}$$

1. Identifier la formulation (7) avec la minimisation d'un risque empirique régularisé.

Solution

— L'espace des hypothèses est toujours l'ensemble des hyperplans de $\mathbb{R}^p$.

— La fonction de perte est la perte hinge.

— La régularisation est une régularisation ℓ_2.

— Le coefficient de régularisation est $\lambda = \frac{1}{2C}$. □

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 7 est équivalent à

$$\underset{\vec{w}\in\mathbb{R}^p, b\in\mathbb{R}, \vec{\xi}\in\mathbb{R}^n}{\arg\min} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^n \xi_i \tag{8}$$

$$\text{t. q. } \begin{cases} y^i\left(\langle \vec{w}, \vec{x}^{\,i}\rangle + b\right) \geq 1 - \xi_i, \\ \xi_i \geq 0 \end{cases} \qquad \text{pour tout } i = 1,\ldots,n \tag{9}$$

Montrer en suivant la même démarche que pour la question 12 de la section précédente que le problème (8) est équivalent à :

$$\max_{\vec{\alpha}\in\mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n\sum_{l=1}^n \alpha_i\alpha_l y^i y^l \langle \vec{x}^{\,i}, \vec{x}^{\,l}\rangle \tag{10}$$

$$\text{t. q. } \sum_{i=1}^n \alpha_i y^i = 0 \text{ et } 0 \leq \alpha_i \leq C, \text{ pour tout } i = 1,\ldots,n.$$

et que si on appelle $(\vec{w}^{\,*}, b^*)$ un minimiseur du problème (8), et α^* un maximiseur du problème ci-dessus, alors :

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

Solution Introduisons $2n$ multiplicateurs de Lagrange $\{\alpha_i, \beta_i\}_{i=1,\ldots,n}$ et écrivons le lagrangien :

$$\begin{aligned} L : \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}_+^n \times \mathbb{R}_+^n &\to \mathbb{R} \\ \vec{w}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta} &\mapsto \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i\left(y^i\left(\langle \vec{w}, \vec{x}^{\,i}\rangle + b\right) - 1 + \xi_i\right) \\ &\quad - \sum_{i=1}^n \beta_i \xi_i. \end{aligned}$$

Le problème dual de celui présenté par l'équation 8 est donc

$$\max_{\vec{\alpha}\in\mathbb{R}_+^n, \vec{\beta}\in\mathbb{R}_+^n} \inf_{\vec{w}\in\mathbb{R}^p, b\in\mathbb{R}, \vec{\xi}\in\mathbb{R}^n} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i\left(y^i\left(\langle \vec{w}, \vec{x}^{\,i}\rangle + b\right) - 1 + \xi_i\right) - \sum_{i=1}^n \beta_i\xi_i. \tag{11}$$

Comme dans le cas de la SVM à marge rigide, le lagrangien est minimal quand son gradient en $\vec{w}$ est nul, à savoir quand

$$\vec{w} = \sum_{i=1}^n \alpha_i y^i \vec{x}^{\,i}. \tag{12}$$

Toujours comme précédemment, il est affine en b et son infimum est donc $-\infty$, sauf si son gradient en b est nul, à savoir si

$$\sum_{i=1}^{n} \alpha_i y^i = 0. \tag{13}$$

De plus, il est affine en $\vec{\xi}$ et son infimum est donc $-\infty$, sauf si son gradient en $\vec{\xi}$ est nul, à savoir si

$$\beta_i = C - \alpha_i, i = 1, \ldots, n. \tag{14}$$

La fonction duale est donc maximisée quand les équations (13) et (14) sont vérifiés.

En remplaçant $\vec{w}$ par son expression (équation (12)) dans l'expression de la fonction duale, l'équation (11) peut donc être reformulée comme

$$\begin{aligned} &\max_{\vec{\alpha} \in \mathbb{R}^n} -\frac{1}{2} \sum_{i=1}^{n} \sum_{l=1}^{n} \alpha_i \alpha_l y^i y^l \langle \vec{x}^{\,i}, \vec{x}^{\,l} \rangle + \sum_{i=1}^{n} \alpha_i + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} (C - \alpha_i) \xi_i - \sum_{i=1}^{n} \alpha_i \xi_i \\ &\text{t. q. } \sum_{i=1}^{n} \alpha_i y^i = 0;\ \alpha_i \geq 0, i = 1, \ldots, n;\ C - \alpha_i \geq 0, i = 1, \ldots, n. \end{aligned}$$

Comme précédemment, les conditions de Karush-Kuhn-Tucker nous permettent de caractériser plus précisément la relation entre $\vec{\alpha}^*$ et $(\vec{w}^*, b^*)$. Pour chaque observation i, nous avons maintenant :

$$\begin{cases} \alpha_i^* g_i(\vec{w}^*, b^*) = 0 \\ \beta_i^* h_i(\vec{w}^*, b^*) = 0 \quad \text{pour tout } i = 1, \ldots, n. \end{cases}$$

avec $g_i(\vec{w}^*, b^*) = \langle \vec{w}^*, \vec{x}^{\,i} \rangle + b^*$ et $h_i(\vec{w}^*, b^*) = \left[1 - y^i \left(\langle \vec{w}, \vec{x}^{\,i} \rangle + b\right)\right]_+$. □

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

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

Solution Nous avons ainsi, pour chaque observation i, trois possibilités :

- $\alpha_i^* = 0$: le minimiseur de $\frac{1}{2}\|\vec{w}\|_2^2$ vérifie la contrainte et $y^i \left(\langle \vec{w}, \vec{x}^{\,i} \rangle + b\right) > 1$. L'observation $\vec{x}^{\,i}$ est, encore une fois, *à l'extérieur* de la zone d'indécision.
- $0 < \alpha_i^* < C$: comme précédemment, $\vec{x}^{\,i}$ est un vecteur de support situé sur la bordure de la zone d'indécision.
- $\beta_i^* = 0 : \alpha_i^* = C$, auquel cas $\left[1 - y^i \left(\langle \vec{w}, \vec{x}^{\,i} \rangle + b\right)\right]_+ > 0$. Dans ce cas, $\vec{x}^{\,i}$ est *du mauvais côté* de la frontière de la zone d'indécision.

□