# 1 Cours

Répondre aux questions suivantes

1. Qu'est-ce que le problème de la classification supervisée ?
   *A partir d'un ensemble $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, $\mathbf{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, trouver une application $f : \mathcal{X} \to \mathcal{Y}$ modélisant la relation entre une donnée $\mathbf{x}$ et son étiquette $y$.*
2. Quelles sont les hypothèses faites pour le problème de l'apprentissage supervisé à partir de $n$ données ?
   *Hypothèses :*
   *– les données de $\mathcal{S}$ sont i.i.d selon une distribution de probabilité $p$ définie sur $\mathcal{X} \times \mathcal{Y}$*
   *– $p$ est une distribution inconnue et fixe*
3. Quel est le problème de la généralisation ?
   *Capacité de pouvoir correctement affecter des étiquettes à des données qui ne faisaient pas partie de l'ensemble d'apprentissage. Ne pas faire d'erreur de classification sur l'ensemble d'apprentissage $\mathcal{S}$ ne garantit pas que l'erreur en généralisation sera faible. Le contraire peut même parfois s'observer et on parle de sur-apprentissage (over-fitting).*

# 2 Régression logistique

Dans cet exercice, on s'intéresse à la régression logistique, qui permet d'aborder des problèmes de classification supervisée. On se pose plus précisément le problème de la classification binaire. On suppose que l'espace auquel appartiennent les données est $\mathbb{R}^d$ et que l'espace des étiquettes est $\mathcal{Y} = \{0, 1\}$.

1. En régression logistique, la classification d'un exemple $\mathbf{x} \in \mathbb{R}^d$ se fait selon la valeur d'une fonction $f$ dépendant d'un vecteur $\mathbf{w} \in \mathbb{R}^d$ et d'un réel $b$ qui est définie par :

$$f(\mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x} + b))}$$

   (a) donner les valeurs minimale et maximale de $f$
   *On a*

$$\lim_{\mathbf{w}\cdot\mathbf{x}+b \to +\infty} f(\mathbf{x}) = 1 \quad (1)$$

$$\lim_{\mathbf{w}\cdot\mathbf{x}+b \to -\infty} f(\mathbf{x}) = 0 \quad (2)$$

   (b) proposer et justifier une règle d'affectation d'une classe à un exemple $\mathbf{x}$ en fonction de $f(\mathbf{x})$
   *Cf. question suivante : la classification de $\mathbf{x}$ par régression logistique se fait selon la position de $\mathbf{x}$ par rapport à l'hyperplan $\mathbf{w} \cdot \mathbf{x} + b$. Il est donc naturel de déterminer l'affectation de $\mathbf{x}$ selon que $\mathbf{w} \cdot \mathbf{x} + b$ est positif ou négatif. La frontière entre ces deux cas s'obtient lorsque $\mathbf{w} \cdot \mathbf{x} + b = 0$ soit $f(\mathbf{x}) = 0$. Une règle d'affectation simple est de d'affecter la classe +1 à $\mathbf{x}$ si $f(\mathbf{x}) \geq 0.5$ et -1 sinon.*

   (c) quelles sont les limites de la classification par régression logistique (donner un exemple de problème qui ne peut être appris)
   *Classification linéaire → problème du XOR*

   (d) dans le cas où $d = 2$, $\mathbf{w} = [-2\ 1]^\top$ et $b = 1$, représenter graphiquement la surface de décision correspondant à la règle d'affectation proposée précédemment.
   *Facile : droite d'équation $x_2 = 2x_1 - 1$*

2. Soit un ensemble d'apprentissage $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\}$ et une distribution $p(X, Y)$ sur $\mathbb{R}^d \times \mathcal{Y}$. En s'aidant de $p(X, Y) = P(Y|X)p(X)$, écrire la vraisemblance des données de $\mathcal{S}$. Rappeler les hypothèses usuellement faites en classification supervisée.
   *Cf. questions de cours plus haut.*

3. Dans le cadre de la régression logistique, on fait l'hypothèse que la probabilité $P(Y = 1|X = \mathbf{x})$ qu'un exemple $\mathbf{x}$ soit de la classe $+1$ est égale à $f(\mathbf{x})$. En utilisant ce modèle statistique pour les données, et en s'inspirant de l'écriture de la vraisemblance pour une variable de Bernoulli, écrire la vraisemblance de l'échantillon $\mathcal{S}$ en fonction des $y_i$, des $p(\mathbf{x}_i) := p(X = \mathbf{x}_i)$ et de $\mathbf{w}$ et $b$.
   *Tout comme pour une variable de Bernoulli $Z$ de paramètre $\theta$ pour laquelle $P(Z = z) = (1 - \theta)^(1 - z)\theta^z$ avec $z \in \{0, 1\}$, on a*

$$P(Y = y|X = \mathbf{x}) = f(\mathbf{x})^y(1 - f(\mathbf{x}))^{(1-y)}$$

   *(cette notation permet de compacter en une seule écriture les expressions de $P(Y = 1|X = \mathbf{x})$ et $P(Y = 0|X = \mathbf{x})$)*
   *On en déduit l'expression de la vraisemblance $\mathcal{V}$ :*
   *– pour $\ell$ couples $(X_1, Y_1), \ldots, (X_\ell, Y_\ell)$ de variables aléatoires i.i.d on a*

$$\begin{aligned} P((X_1, Y_1), \ldots, (X_\ell, Y_\ell)) &= \prod_{i=1}^{\ell} P(X_i, Y_i) \textit{car les couples de v.a. sont i.i.d} \\ &= \prod_{i=1}^{\ell} [P(Y_i|X_i)p(X_i)] \textit{ évident} \end{aligned}$$

   *– d'où, la vraisemblance des données $\mathcal{S}$ sous le modèle de régression logistique*

$$\begin{aligned} \mathcal{V} &= \prod_{i=1}^{\ell} P(Y_i = y_i|X_i = \mathbf{x}_i)p(X_i = \mathbf{x}_i) \\ &= \prod_{i=1}^{\ell} \left[f(\mathbf{x}_i)^{y_i}(1 - f(\mathbf{x}_i))^{1-y_i}\right] \prod_{i=1}^{\ell} p(X_i = \mathbf{x}_i) \\ &= \prod_{i=1}^{\ell} \left[\left(\frac{1}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x}_i + b))}\right)^{y_i} \left(\frac{\exp(-(\mathbf{w} \cdot \mathbf{x}_i + b))}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x}_i + b))}\right)^{1-y_i}\right] \prod_{i=1}^{\ell} p(X_i = \mathbf{x}_i) \\ &= \prod_{i=1}^{\ell} \frac{\exp(-(1 - y_i)(\mathbf{w} \cdot \mathbf{x}_i + b))}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x}_i + b))} \prod_{i=1}^{\ell} p(\mathbf{x}_i) \end{aligned}$$

4. Exprimer la log-vraisemblance $\mathcal{L}$ de l'échantillon $\mathcal{S}$ sous le modèle (de la régression) logisitique. Ne pas faire figurer les termes dépendant des $p(\mathbf{x}_i)$.
   *On a*

$$\begin{aligned} \mathcal{L} &= \log \mathcal{V} \\ &= \sum_{i=1}^{\ell} (y_i - 1)(\mathbf{w} \cdot \mathbf{x}_i + b)) - \sum_{i=1}^{\ell} \log\left(1 + \exp(-(\mathbf{w} \cdot \mathbf{x}_i + b))\right) \end{aligned}$$

   *Note : le modèle que l'on cherche à obtenir est celui de $P(Y|X)$ en estimant les paramètres $\mathbf{w}$ et $b$ qui régissent cette distribution sous le modèle de la régression logistique. La loi $p(X)$ des $X$ est une loi sur les descriptions des exemples que l'on ne cherche généralement pas à expliciter dans le cas de problèmes d'apprentissage supervisé, ce pourquoi on peut négliger ce terme dans l'expression de $\mathcal{L}$.*

5. Quels sont les paramètres à estimer ?
   $\mathbf{w}$ *et $b$ : ces paramètres définissent complètement la distribution de probabilité P(Y|X) sous le modèle logistique.*

6. Donner l'expression de $\frac{\partial \mathcal{L}}{\partial w_j}$, pour $j = 1, \ldots, d$ ainsi que l'expression de $\frac{\partial \mathcal{L}}{\partial b}$.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w_j} &= \sum_{i=1}^{\ell} (1 - y_i)x_{ij} + \sum_{i=1}^{\ell} \frac{x_{ij} \exp(-(\mathbf{w} \cdot \mathbf{x}_i + b))}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x}_i + b)} \\ \frac{\partial \mathcal{L}}{\partial b} &= \sum_{i=1}^{\ell} (1 - y_i) + \sum_{i=1}^{\ell} \frac{\exp(-(\mathbf{w} \cdot \mathbf{x}_i + b))}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x}_i + b)} \end{aligned}$$

*où $x_{ij}$ est la $j^{eme}$ coordonnée du vecteur $\mathbf{x}_i$*

7. Le gradient de $\mathcal{L}$ peut-il s'annuler directement ? Proposer une équation de mise à jour (type gradient) permettant de produire une suite de paramètres menant à un maximum (local) de vraisemblance.

   *Non le gradient ne peut s'annuler directement (remarque : si l'on veut annuler directement le gradient, on est amené à considérer un système de $d+1$ équations* non linéaires *à $d+1$ inconnues).*

   *Une solution simple est d'utiliser une méthode de gradient. Ici, on a les équations de mise à jour :*

$$w_j^{t+1} \leftarrow w_j^t + \eta_j \frac{\partial \mathcal{L}}{\partial w_j}$$
$$b^{t+1} \leftarrow b^t + \eta_b \frac{\partial \mathcal{L}}{\partial b}$$

   *avec $\eta_j > 0$ et $\eta_b > 0$. Note : on peut choisir $\eta_1 = \ldots = \eta_d = \eta_b$.*

   *On remarquera que l'on effectue ici une montée de gradient (d'où le signe $'+'$ dans les équations de mise à jour) car on veut* maximiser *la vraisemblance des données par rapport au modèle.*

8. Ecrire l'algorithme complet d'apprentissage par régression logistique.

   *Facile. Se reporter à l'algorithme vu en cours sur l'apprentissage pour les perceptrons multi-couches.*

9. Proposer une version stochastique de l'algorithme d'apprentissage proposé précédemment. Dans le cas d'un pas d'apprentissage adaptatif, rappeler les conditions nécessaires et suffisantes sur le pas adaptatif pour être assuré de la convergence de l'algorithme. Rappeler les avantages de la version stochastique par rapport à la version non stochastique.

   *Il suffit d'enlever les $\sum_{i=1}^{\ell}$ dans les expressions du gradient obtenues précédemment pour obtenir les équations permettant d'évaluer le gradient stochastique pour un exemple $\mathbf{x}_j$.*

   *Conditions sur le pas d'apprentissage $\eta_t$ : $\sum_t \eta_t = +\infty$ et $\sum_t \eta_t^2 < +\infty$*

   *Avantages de la version stochastique*
   - *on n'a pas besoin de toute la base d'apprentissage au début de l'apprentissage mais on peut acquérir les exemples au d'apprentissage fur et à mesure et les apprendre au fur et à mesure (en utilisant les formule de gradient stochastique) – cf. apprentissage incrémental et apprentissage en ligne*
   - *permet de s'extraire plus facilement de minima locaux*
   - *gradients plus faciles à évaluer*

10. En considérant l'ensemble d'apprentissage

$$\mathcal{S} = \left\{ \left( \begin{bmatrix} 1 \\ 0 \end{bmatrix}, 1 \right), \left( \begin{bmatrix} 0 \\ 1 \end{bmatrix}, 0 \right) \right\}$$

    les valeurs initiales $\mathbf{w}^0 = [0\ 1]^\top$ et $b^0 = -1$ et un pas d'apprentissage fixe $\eta = 0.3$, faire deux itérations des algorithmes d'apprentissage proposés.