TD1

Exercice 1 (Classification de XOR)**.** *On considère un ensemble de quatre données* $\{(0,0),(0,1),(1,0),(1,1)\}$*, constitué de deux classes :*

- $C_0 = \{(0,0),(1,1)\}$
- $C_1 = \{(0,1),(1,0)\}$.

1. *Représenter cet ensemble de données et faire apparaître les classes.*
2. *Cet ensemble peut-il être classé à l'aide d'un perceptron ?*
3. *Ecrire un réseau à deux couches permettant de classer cet ensemble.*

Exercice 2 (Régression logistique multivariée)**.** *On considère un ensemble de données* $\{\varphi_n\}_{1\leq n\leq N}$ *de* $\mathbb{R}^D$*, découpée en* K *classes* $C_1,\ldots,C_K$ *avec* $K\geq 2$*. On considère l'ensemble de données* $\mathcal{T} = \{(\varphi_n, t_n)\}_{1\leq n\leq N}$ *tel que* $\varphi_n \in C_{t_n}$*. On note* $y_k(\varphi) = \mathbb{P}(\varphi\in C_k)$*. On fait l'hypothèse suivante: Il existe* $(\omega_k)_{1\leq k\leq K}$ *vecteurs de* $\mathbb{R}^D$ *tels que*

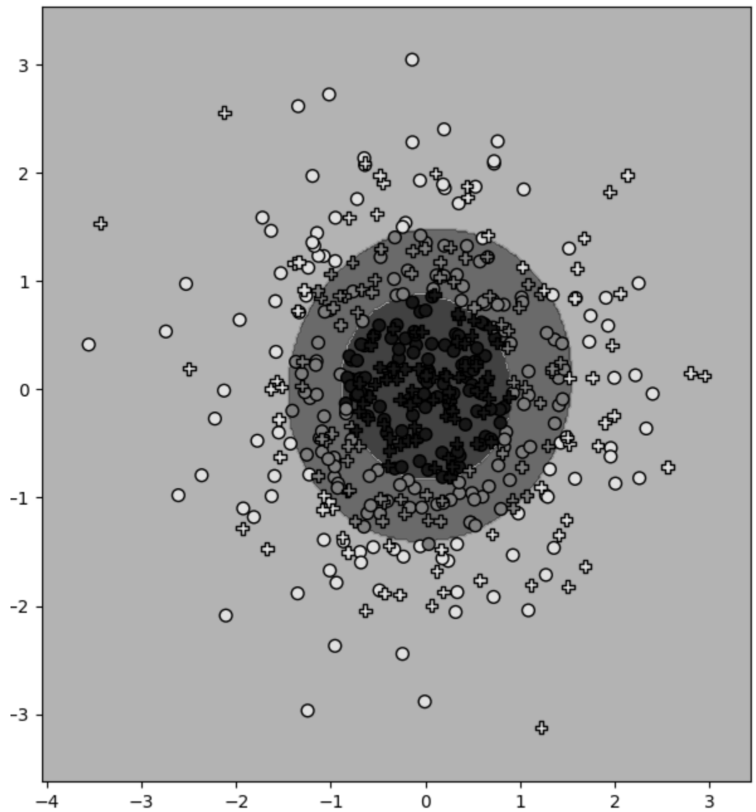
$$y_k(\varphi) = \frac{\exp(\omega_k^T\varphi)}{\sum_{j=1}^K \exp(\omega_j^T\varphi)} \tag{1}$$

On cherche à déterminer ces vecteurs ou, autrement dit, la matrice $W = (\omega_1 \mid \omega_2 \mid \ldots \mid \omega_K) \in \mathbb{R}^{D\times K}$*. Pour cela, on va chercher à maximiser par rapport à* W *la log-vraisemblance* $\log\mathbb{P}\left(\bigcap_{n=1}^N \varphi_n\in C_{t_n}\middle|W\right)$ *(ou minimiser son opposé).*

1. *Vérifier que* $\sum_{k=1}^K y_k(\varphi) = 1$.
2. *En supposant que les données sont indépendantes, écrire la log-vraisemblance :*
$$\log\mathbb{P}\left(\bigcap_{n=1}^N \varphi_n\in C_{t_n}\middle|W\right) \tag{2}$$
3. *Écrire* $\nabla_{\omega_k}\log\mathbb{P}\left(\varphi_n\in C_{t_n}\middle|W\right)$ *pour* $1\leq k\leq K$.
4. *En déduire le gradient de la log vraisemblance.*
5. *Ecrire en pseudo-code un algorithme de gradient pour obtenir la matrice* W *qui maximise la vraisemblance.*

Exercice 3 (Régression logistique multivariée : un exemple.)**.** *En général, un ensemble de données ne vérifie pas l'hypothèse de l'existence de* $(\omega_k)_{1\leq k\leq K}$*, on considère alors une transformation de l'ensemble de données pour arriver à la valider. Plus précisément, on part de* $\{(x_n,t_n), x_n\in C_{t_n}\}\subset\mathbb{R}^d\times\{1,\ldots,K\}$ *et on considère* $\varphi:\mathbb{R}^d\mapsto\mathbb{R}^D$ *puis* $\{(\varphi(x_n),t_n), x_n\in C_{t_n}\}\subset\mathbb{R}^D\times\{1,\ldots,K\}$ *pour appliquer l'exercice précédent.* φ *est appelée "feature transform".*

On considère l'ensemble suivant :



1. *Proposer une transformation de l'ensemble très simple pour qu'il vérifier l'hypothèse* $(D = 1)$.
2. *Expliciter des* $(\omega_k)_{1 \le k \le K}$ *qui conviennent.*