TD 4

Exercice 1 – Évaluation(s) de l'erreur

Q 1.1 Rappelez la fonction coût au sens des moindres carrés sur un problème d'apprentissage binaire. Proposer quelques exemples pour montrer que les échantillons correctement classés participent à la fonction coût.

Q 1.2 En faisant appel à vos connaissances sur le perceptron, proposez une nouvelle fonction coût ne pénalisant que les points mals classés.

Q 1.3 En imaginant une fonction f de complexité infinie (capable de modéliser n'importe quelle frontière de décision), tracez à la main la frontière de décision optimale au sens des coûts définis précédemment pour le deux problèmes jouets de la figure 1. Ces frontières sont-elles *intéressantes* ? Quels problèmes se posent ?

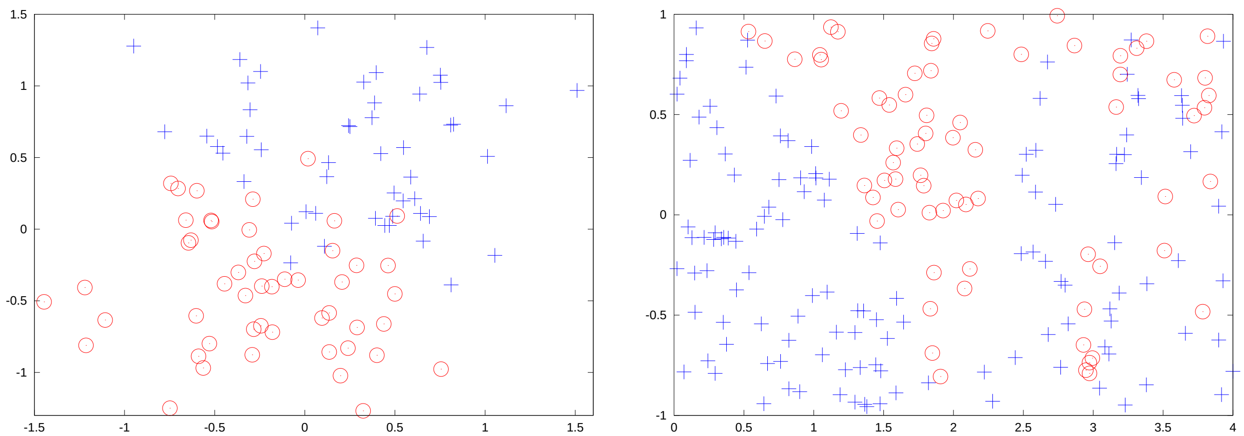

Figure 1 – Gaussiennes non séparables linéairement

Exercice 2 – Perceptron

Q 2.1 Soit $\mathbf{w} = (2, 1, 1)$ le vecteur de poids d'une séparatrice linéaire. Dessinez en 2D cette séparatrice dans le plan. Précisez sur le dessin les quantités $< \mathbf{w}, \mathbf{x} >$ par rapport à un exemple $\mathbf{x}$ bien classé et mal classé. Que se passe-t-il pour le produit scalaire dans le cas d'un exemple mal classé avec la mise-à-jour $\mathbf{w} \leftarrow \mathbf{w} + y\mathbf{x}$?

Q 2.2 Comment sont les classifieurs suivants par rapport à celui de la question précédente : $w^1 = (1, 0.5, 0.5)$, $w^2 = (200, 100, 100)$, $w^3 = (-2, -1, -1)$?

Q 2.3 Toujours en 2D, donnez un perceptron qui permet de réaliser le `AND` logique entre les entrées x_1 et x_2 (positif si les deux sont à 1, négatif sinon) et un autre pour le `ET` logique.

Q 2.4 Montrez que l'algorithme du perceptron correspond à une descente de gradient. La solution est-elle unique ?

Q 2.5 Soit $w_0 = 0$. Exprimer $\| \sum y^i \mathbf{x}^i \|$ en fonction de $w_{t+1} - w_t$ (on ne considérera que les itérations où il y a effectivement une mise-à-jour) ; puis en fonction de w_{T+1} la dernière itération ; enfin en fonction de $\sqrt{\sum \|\mathbf{w}_t + y_t\mathbf{x}_t\|^2 - \|\mathbf{w}_t\|^2}$. Montrer que cette dernière expression est inférieure à $\sqrt{\sum \|\mathbf{x}_t\|^2}$

Q 2.6 Montrer le théorème de Novikoff : Soit $\mathbf{x}_1, \ldots, \mathbf{x}_T$ une séquence de T points, et r tel que $\|\mathbf{x}_t\| \leq r$; s'il existe $\rho > 0$ et $\mathbf{w}$ tels que $\rho \leq \frac{y_i < \mathbf{w}.\mathbf{x}_t >}{\|w\|}$, alors le nombre de mise-à-jour effectuées par l'algorithme du perceptron est borné par r^2/ρ^2. Vous partirez de $\frac{\mathbf{w} \sum y_t \mathbf{x}_t}{\|\mathbf{w}\|}$.

Exercice 3 – Expressivité des séparateurs linéaires

On se place dans l'espace des séparateurs linéaires : $f(\mathbf{x}_i) = \sum_j x_{ij} w_j$.

Q 3.1 Quelle est la dimension du vecteur $\mathbf{w}$? Rappeler l'écriture matricielle de $f(\mathbf{x}_i)$. Tracer approximativement les frontières optimales en utilisant un modèle linéaire basique sur la figure de l'exercice 1.

Q 3.2 Nous allons augmenter l'expressivité du modèle en étendant l'espace de représentation initial dans le cas 2D : $\mathbf{x} = [x_1, x_2]$. Soit la transformation ϕ suivante : $\phi(\mathbf{x}) = [x_1, x_2, x_1^2, x_2^2, x_1x_2]$, considérons le modèle linéaire $f(\mathbf{x}_i) = \sum_j \phi_j(\mathbf{x}_i) w_j$.

- Quelle est la dimension du vecteur $\mathbf{w}$ dans ce cas ?
- Retracer les frontières de décision optimales sur la figure en utilisant cette nouvelle représentation.
- Pouvons nous retrouver les frontières linéaires de la question précédente dans ce nouvel espace ? Dans l'affirmative, donner les coefficients w_j associés.

Q 3.3 Les frontières sont-elles plus *intéressantes* en utilisant la première ou la seconde représentation des données ? Pouvez vous comparer grossièrement l'amplitude de la fonction coût (au sens des moindres carrés par exemple) dans les cas linéaires et quadratiques ? Qu'en déduire ? Sur quel élément vous basez vous pour mesurer la qualité du modèle créé ?

Q 3.4 Afin d'augmenter l'expressivité de notre classe de séparateur, nous nous tournons vers les représentations gaussiennes.Nous créons une grille de points $\mathbf{p}^{i,j}$ sur l'espace 2d, puis nous mesurons la similarité gaussienne du point $\mathbf{x}$ par rapport à chaque point de la grille : $s(\mathbf{x}, \mathbf{p}^{i,j}) = Ke^{\frac{||\mathbf{x}-\mathbf{p}^{\mathbf{i},\mathbf{j}}||^2}{\sigma}}$. La nouvelle représentation de l'exemple est le vecteur avec contenant pour chaque dimension la similarité de l'exemple à un point de la grille.

- Quelle est la dimension du vecteur $\mathbf{w}$?
- Donnez l'expression littérale de la fonction de décision.
- Quelle est la différence avec l'algorithme des fenêtres de Parzen ?
- Quel rôle joue le paramètre σ ?

Q 3.5 Introduction (très) pragmatique aux noyaux

- Que se passe-t-il en dimension 3 si nous souhaitons conserver la résolution spatiale du maillage ?
- Afin de palier ce problème, nous proposons d'utiliser la base d'apprentissage à la place de la grille : les points servant de support à la projection seront ceux de l'ensemble d'apprentissage. Exprimer la forme littérale de la fonction de décision dans ce nouveau cadre. Quelle est la nouvelle dimension du paramètre $\mathbf{w}$?
- Soit K la matrice regroupant l'ensemble des produits scalaires, donner les dimensions de K. Exprimer la fonction de décision en fonction de K.
- Que se passe-t-il lorsque σ tend vers 0 ? vers l'infini ? A-t-on besoin de toutes les dimensions de w ou est-il possible de retrouver la même frontière de decision en limitant le nombre de données d'apprentissage ? A quoi cela correspond-il pour $||\mathbf{w}||$?