TD 3 - Descente de gradient, Modèles linéaires

Exercice 1 – Apéro

Q 1.1 Parmi les fonctions suivantes, lesquelles sont convexes :
$f(x) = xcos(x)$, $g(x) = -log(x) + x^2$, $h(x) = x\sqrt{x}$, $t(x) = -log(x) - log(10 - x)$?

Q 1.2 Soit une application linéaire $f \in \mathbb{R}^n \to \mathbb{R}$; rappeler ce qu'est le gradient de f : $\nabla f(\mathbf{x})$. Donner le gradient de $f(\mathbf{x}) = 2x_1 + x_2^2 + x_2x_3$

Q 1.3 Exprimer $\nabla_{\mathbf{x}}(f(\mathbf{x}) + g(\mathbf{x}))$, $\nabla_{\mathbf{x}} tf(\mathbf{x})$.
Donner l'expression de $\nabla_{\mathbf{x}} b'\mathbf{x}$ avec $b \in \mathbb{R}^d$ et $\nabla_{\mathbf{x}}\mathbf{x}'A\mathbf{x}$ pour A symétrique.

Exercice 2 – Régression linéaire

Soit un ensemble de données d'apprentissage $\mathcal{D} = \{\mathbf{x^i}, y^i\}_{i=1,\dots,N}, \mathbf{x^i} \in \mathbb{R}, y^i \in \mathbb{R}$.
Par convention que l'on suivra dans toute la suite du cours, la matrice de données sera notée X, où chaque ligne correspond à un exemple. La matrice Y des réponses est donc une matrice colonne ; la matrice W des poids également. L'erreur sur $\mathcal{D}$ sera notée $C(W)$.

Q 2.1 Résolution analytique

Q 2.1.1 Rappeler le principe de la régression linéaire. Quelle fonction d'erreur $C(W)$ est utilisée ?

Q 2.1.2 Quelles sont les dimensions des matrices X, W et Y ? Rappeler la formulation matricielle de l'erreur.

Q 2.1.3 Trouver analytiquement la matrice W solution de la régression linéaire, qui minimise $C(W)$.

Q 2.1.4 Même question si l'on considère maintenant une machine linéaire avec biais. Quelle est la valeur optimale du biais w_0 dans ce cas ?

Q 2.2 Rappeler le principe de l'algorithme de descente du gradient. Donner son application au cas de la régression linéaire.

Q 2.3 On considère dans la suite un problème à 2 dimensions.

Q 2.3.1 Tracer l'espace des paramètres en 2D. Positionner arbitrairement les points $\mathbf{w}^0$, point initial, et $\mathbf{w}^*$, solution analytique du problème. Etant donnée la nature quadratique du coût, tracer les iso-contours de la fonction de coût dans l'espace des paramètres. Quelle est la forme de la fonction de coût $C(\mathbf{w}^0)$ dans l'espace des paramètres ?

Q 2.3.2 Dessiner le vecteur $\nabla C(\mathbf{w}^0)$. A quoi correspond ce vecteur géométriquement ?

Exercice 3 – Régression logistique

Q 3.1 On considère un ensemble de données X muni d'étiquettes binaires $Y = \{0, 1\}$. En régression logistique, on considère que le log-rapport des probas conditionnelles $p(y|x)$ peut être modélisé par une application linéaire : $\log\left(\frac{p(y|x)}{(1-p(y|x))}\right) = \theta.x$.

- Quel est le but ?
- Quelle étiquette prédire pour x si $\theta.x > 0$?
- Que vaut $p(y|x)$? Tracer la fonction $p(y|x)$ en fonction de $\theta.x$.

- En déduire le type de frontière que la regression logistique permet de déterminer

Q 3.2 Pour une dimension x_i, quelle est l'influence de sa valeur pour $p(y|x)$? Quelle est la limite de la régression logistique ?

Q 3.3 Soit θ les paramètres recherchés. Quelle est l'expression de la vraissemblance conditionnelle de *theta* par rapport à un exemple (x, y) ? La log-vraissemblance ? Et sur un ensemble d'exemples $\mathcal{D}$?

Q 3.4 Proposer un algorithme pour résoudre le problème de la régression logistique.

Exercice 4 – Optimisation d'un modèle gaussien par descente de gradient

Nous disposons ici d'un jeu de données non-étiquetées : $\mathcal{D} = \{\mathbf{x}_i\}_{i=1,\dots,N}, \mathbf{x}_i \in \mathbb{R}^d$.
Nous souhaitons apprendre en mode non supervisé un modèle gaussien correspondant aux données de $\mathcal{D}$. Le modèle gaussien est défini par un ensemble de paramètres $\{\mu, \Sigma\}$

Q 4.1 Exprimez la log-vraisemblance en supposant les exemples de $\mathcal{D}$ statistiquement indépendants.

Q 4.2 Solution analytique

Q 4.2.1 Que vérifie la solution $W^\star$ du maximum de vraisemblance ? Montrez que la solution $W^\star$ du maximum de vraisemblance correspond à la moyenne et la covariance empirique des données $\mathcal{D}$ dans le cas où Σ est une matrice diagonale.

Q 4.3 Méthode de gradient

Q 4.3.1 Déterminez le gradient de la vraisemblance en un point W_0.

Q 4.3.2 Ecrire deux algorithmes de gradient batch et stochastique permettant d'apprendre une loi gaussienne à partir de $\mathcal{D}$.

Exercice 5 – Évaluation(s) de l'erreur

Q 5.1 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 5.2 En faisant appel à vos connaissances sur le perceptron, proposez une nouvelle fonction coût ne pénalisant que les points mals classés.

Q 5.3 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 ?

Exercice 6 – Perceptron

Q 6.1 Soit $\mathbf{w} = (2, 1)$ le vecteur de poids d'une séparatrice linéaire. Dessinez 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 6.2 Comment sont les classifieurs suivants par rapport à celui de la question précédente : $w^1 = (1, 0.5)$, $w^2 = (200, 100)$, $w^3 = (-2, -1)$?

Q 6.3 Montrez que l'algorithme du perceptron correspond à une descente de gradient. La solution est-elle unique ?

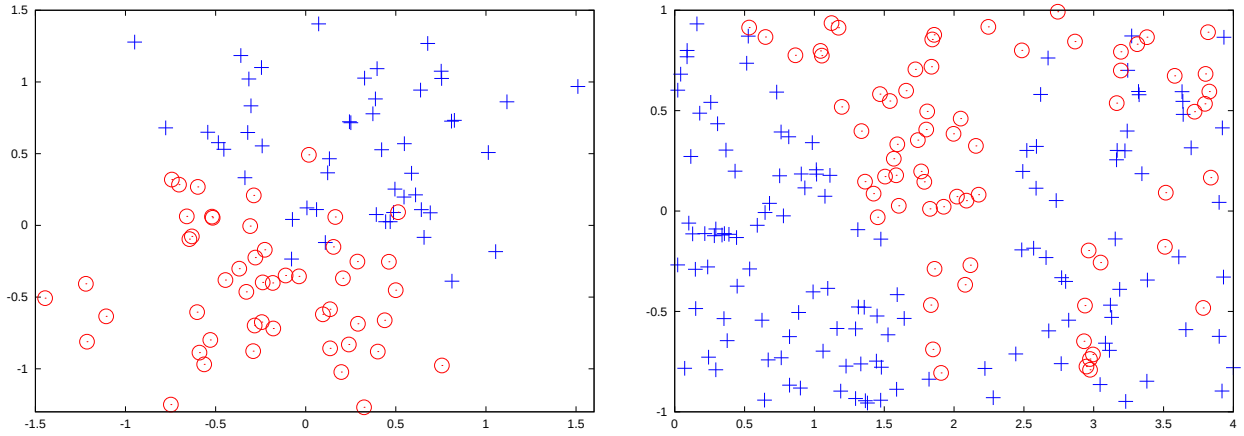

FIGURE 1 – Gaussiennes non séparables linéairement

Q 6.4 Quel problème peut-il se poser pour certaines valeurs de w ? Comment y rémédier ?

Q 6.5 Donner un perceptron qui permet de réaliser le `AND` logique entre les entrées binaires x_1 et x_2 (positif si les deux sont à 1, négatif sinon) et un autre pour le `OR` logique.

Q 6.6 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 : $\phi(\mathbf{x}) = [1, x_1, x_2, x_1^2, x_2^2, x_1 x_2]$, 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 ?
- A quoi correspond la projection *phi* ?
- 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 6.7 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 6.8 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 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.
- Quel rôle joue le paramètre σ ?

Q 6.9 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 $\mathbf{w}$?
- 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 $||\mathbf{w}||$?