ARF Examen

Exercice 1 (4 points) – Questions indépendantes

Répondre par vrai ou faux `en justifiant` votre réponse en quelques lignes.

1. Sur des données séparables linéairement, l'algorithme du perceptron converge vers un minima local du fait de l'initialisation aléatoire des poids.
2. L'estimation de densité est un préalable à tout algorithme de classification.
3. La régression logistique apprend des frontières non-linéaires car la fonction logistique est non linéaire.
4. Il est préférable d'utiliser une estimation de densité par fenêtre de Parzen en très grande dimension plutôt qu'une estimation par histogramme.
5. Il est préférable d'utiliser la régression aux moindres carrés que la régression logistique pour apprendre la probabilité d'un événement.
6. Il est préférable d'utiliser une descente de gradient pour le calcul de la régression aux moindres carrés que la solution analytique.
7. L'algorithme du perceptron converge toujours.
8. Le classifieur naïf bayésien ne peut pas classifier correctement le problème du *ou disjonctif* (*xor*, l'échiquier à 4 cases).

Exercice 2 (5 points) – SVM déséquilibré

Soit un ensemble de données $\mathcal{D} = \{(x^i, y^i) \in \mathbb{R}^d \times \{-1, 1\}\}_{i=1}^N$, soit N^+ le nombre d'exemples positifs et N^- le nombre d'exemples négatifs, avec $N^+ << N^-$, les données négatives étant très majoritaires. On veut utiliser un SVM linéaire $f(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + b$ avec $\mathbf{w} \in \mathbb{R}^d$ pour classer les données.

Rappel : Pour un problème d'optimisation $min_{\theta\in\mathbb{R}^d} J(\theta)$ sous les m contraintes $g_j(\theta) \leq 0$ pour $j = 1 \ldots m$, le Lagrangien associé est $\mathcal{L}(\theta, \mu) = J(\theta) + \sum_{j=1}^m \mu_j g_j(\theta)$ avec $\mu_j \geq 0$. Les conditions KKT spécifient qu'à l'optimum, nécessairement $\nabla\mathcal{L}(\theta, \mu) = 0$, $g_j(\theta) \leq 0$, $\mu_j \geq 0$, et $\mu_j g_j(\theta) = 0$ pour $j = 1 \ldots m$.

Q 2.1 Dans le cas de la formulation SVM usuel, quel problème se pose avec un tel déséquilibre entre les deux classes ?

Q 2.2 Rappelez la formulation du problème (primal) du SVM, avec les contraintes associées. Donnez la signification des différents termes à minimiser et des contraintes.

Q 2.3 Pour gérer le déséquilibre, on propose de modifier les contraintes de la manière suivante : $y^i(\mathbf{w}^t\mathbf{x}^i + b) \geq 1 - \xi_i \; \forall i = 1, \ldots, N$ avec $\xi_i = 0$ si $y^i = 1$, $\xi_i \geq 0$ sinon . Donner une interprétation du problème en termes d'erreurs pour les points positifs et négatifs.

Q 2.4 Donnez le Lagrangien correspondant à ce problème d'optimisation. Vous pouvez utiliser la notation $i \in N^-$ (resp. $i \in N^+$) pour énumérer les exemples négatifs (resp. les exemples positifs).

Q 2.5 Donnez les conditions d'optimalité par rapport aux variables primales $\mathbf{w}$, b et ξ_i et en déduire l'expression de $\mathbf{w}$.

Q 2.6 Formulez le problème dual correspondant en l'exprimant en particulier sans $\mathbf{w}$, et les contraintes associées.

Q 2.7 Quelle est la particularité des points supports ? Par rapport aux conditions, que vérifient les points bien classés ? Les points mal classés ?

Q 2.8 Proposer une procédure pour évaluer la constante de pénalisation du SVM.

Q 2.9 On sait que la frontière de décision recherchée est une ellipsoïde. Pourquoi cet algorithme n'est-il pas performant pour ce type de problème ? Proposez une solution efficiente au niveau computationnel.

Exercice 3 (4 points) – **C'est Risqué**

Soit un problème de classification en 1D à 2 classes $y = +1$ ou $y = -1$, avec les lois conditionnelles définies pour $x \in [0, 1]$ de la manière suivante : $p(x|y) = (1 - y)/2 + yx$. On suppose de plus que le coût d'une bonne classification est de 0, celui d'une mauvaise de β.

Q 3.1 Illustrer sur un graphique les deux lois conditionnelles des classes positives et négatives et donner intuitivement la frontière de décision.

Q 3.2 Quelle est la règle de décision bayésienne associée au problème en fonction des probabilités a posteriori ? Quel est l'espérance du coût associé à cette décision en un point x ? Est-elle optimale ?

Q 3.3 Explicitez la fonction de décision en fonction de $P(y = 1)$ et $P(y = -1)$. Quelle influence de $P(y = 1)$ sur la frontière de décision ?

Q 3.4 On suppose que la densité des exemples suit une loi proportionnelle à x^2. Calculez le risque en fonction de β et $P(y = 1)$.

Exercice 4 (5 points) – **Regression tri-logistique**

On s'attaque à un problème de classification à 3 classes en utilisant un modèle adapté de la régression logistique : $P(y = k|\mathbf{x}) = \frac{e^{\mathbf{w}_k^t \mathbf{x}}}{\sum_{j=1}^{3} e^{\mathbf{w}_j^t \mathbf{x}}}$ pour $k = 1, 2, 3$.

Q 4.1 Quel rapport avec la régression logistique ?

Q 4.2 On veut estimer les paramètres par un critère de vraisemblance sur un ensemble de données $\{(\mathbf{x}^i, y^i) \in \mathbb{R}^d \times \{1, 2, 3\}\}_{i=1}^{N}$. Donnez le problème d'optimisation associé.

Q 4.3 Proposez un algorithme d'estimation pour les paramètres $\mathbf{w}_k$ et donnez le en pseudo-code.

Q 4.4 On remarque dans les expériences conduites une forte tendance au sur-apprentissage. Proposez une méthode pour contrôler le sur-apprentissage et le changement de formulation du critère d'estimation. Donner le nouvel algorithme.