

QCM ML COSSE

automne 2023

Partie 1 (14pts)

1. [5pts] Indiquer si les affirmations suivantes sont vraies ou fausses

Vrai / Faux L'itération de descente de gradient pour la fonction de coût donnée par la somme des carrés des résidus peut s'écrire

$$\beta_j \leftarrow \beta_j + \eta \sum_{i=1}^N \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}) \right) (-\tilde{x}_j^{(i)}) \quad \text{où } \tilde{\mathbf{x}} = [1, \mathbf{x}]$$

Vrai / Faux La différence entre la régression Ridge et la régression LASSO provient du fait que dans le cas de la régression Ridge, la pénalité sur le vecteur des coefficients de régression est donnée par le carré de la norme ℓ_2 alors que dans le cas de la régression LASSO, cette pénalité est donnée par la norme ℓ_1

Vrai / Faux Dans l'approche de sélection du meilleur sous-ensemble, le nombre de modèles à tester pour des données de dimension D est $\sum_{k=1}^{D+1} \binom{D+1}{k}$

Vrai / Faux La matrice $\mathbf{X}^T \mathbf{X}$ dans les équations normales doit être inversible pour assurer l'unicité du vecteur des coefficients de régression

Vrai / Faux L'estimateur de Maximum de Vraisemblance est équivalent à un estimateur de Maximum à Posteriori défini avec un "a priori" uniforme

Vrai / Faux La régression Lasso est efficace pour la sélection de caractéristiques car elle a tendance à annuler exactement certains coefficients, excluant ainsi ces caractéristiques du modèle

Vrai / Faux Dans le cadre de la classification linéaire à K classes, l'approche "un contre un" requiert l'entraînement de $K(K+1)/2$ discriminants

1. [5pts] Indiquer si les affirmations suivantes sont vraies ou fausses

- Vrai / Faux L'Analyse Discriminante Gaussienne (GDA) est un modèle génératif, ce qui signifie qu'elle modélise la distribution jointe des caractéristiques et des valeurs cibles
- Vrai / Faux Si un modèle est particulièrement performant sur les données d'entraînement mais peu précis sur les données de test, c'est que le modèle est vraisemblablement associé à un terme de biais élevé
- Vrai / Faux Lors de l'étape d'apprentissage, ajouter un terme de régularisation à la fonction de coût conduira probablement à une augmentation du biais et à une diminution de la variance
- Vrai / Faux La fonction d'entropie binaire croisée est symétrique par rapport aux valeurs cibles et aux prédictions. I.e. intervertir les cibles et les prédictions ne change pas la valeur de la fonction.
- Vrai / Faux Une fois entraîné, un réseau de neurones multi-couches dont toutes les fonctions d'activation, hormis pour l'unité de sortie, sont des fonctions linéaires est équivalent à un modèle de régression logistique si la fonction d'activation de sortie est une fonction sigmoïde
- Vrai / Faux La dérivée de la fonction sigmoïde $\sigma(a)$ par rapport à la valeur de préactivation "a" est donnée par $\sigma(a)(1 - \sigma(a))$
- Vrai / Faux Si l'on suppose des entrées binaires $x_1, x_2 \in \{0, 1\}$, un simple perceptron de la forme $y(x) = \sigma(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ où σ est la fonction de Heaviside $\sigma(x) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{sinon} \end{cases}$ peut représenter n'importe laquelle des primitives booléennes AND, OR, NAND et NOR
NOR = "not" OR, NAND = "not" AND

Decembre 2022

1. [5pts] Indiquer si les affirmations suivantes sont vraies ou fausses

- Vrai / Faux Minimiser la fonction d'entropie binaire croisée correspond à rechercher un estimateur de maximum de vraisemblance
- Vrai / Faux Les modèles de régression logistique est un exemple de modèle de classification de type génératif
- Vrai / Faux Dans le cadre de la régularisation de type Ridge et pour un coefficient de régularisation $\lambda > 0$, augmenter la valeur de λ résulte en une augmentation de la variance de la famille de modèles correspondants
- Vrai / Faux Dans le cadre de la régression linéaire, ajouter une pénalité de type Ridge, de coefficient associé $\lambda > 0$, correspond à traduire les valeurs propres de la matrice $\mathbf{X}^T \mathbf{X}$ de λ
- Vrai / Faux Les équations normales admettent toujours au moins une solution
- Vrai / Faux Dans le cas d'un modèle de classification linéaire, initialiser l'ensemble des poids et des biais à zéro aura pour conséquence de n'entraîner aucune mise à jour de ces paramètres dans le cadre d'une descente de gradient
- Vrai / Faux Une descente de gradient appliquée à un modèle linéaire et à une fonction de coût donnée par la somme des carrés des erreurs, retourne toujours le minimum global.

1. [5pts] Indiquer si les affirmations suivantes sont vraies ou fausses

- Vrai / Faux Augmenter le nombre de neurones dans la couche cachée d'un réseau à une couche augmente la variance de la famille de modèles correspondants
- Vrai / Faux Un réseau de neurones profond dont toutes les fonctions d'activation sont données par l'identité est équivalent à un modèle linéaire
- Vrai / Faux L'algorithme de descente de gradient appliqué à un réseau de neurones avec un taux d'apprentissage suffisamment faible et un nombre suffisamment grand d'itérations converge nécessairement vers le minimum global de la fonction coût
- Vrai / Faux Le nombre minimum de neurones nécessaires à l'apprentissage d'un modèle de type 'OU Exclusif' (XOR) est égal à 4
- Vrai / Faux En terme d'expressivité (i.e. capacité à capturer une distribution de données), un réseau de neurones est plus puissant qu'un modèle linéaire basé sur des caractéristiques polynomiales

automne 2023 rattrapage

1. [5pts] Indiquer si les affirmations suivantes sont vraies ou fausses

- Vrai / Faux Les réseaux de neurones à action directe (feedforward) sont constitués de connections entre neurones orientées uniquement de l'entrée vers la sortie.
- Vrai / Faux Le gradient $\partial L / \partial \mathbf{w}$ de la fonction d'entropie binaire croisée pour une seule observation $(\mathbf{x}^{(i)}, t^{(i)})$ et un réseau de neurones consistant en un unique neurone de paramètres (\mathbf{w}, b) est donné par $(\sigma(\mathbf{w}^T \mathbf{x}^{(i)} + b) - t^{(i)}) \sigma'(\mathbf{w}^T \mathbf{x}^{(i)} + b) \mathbf{x}^{(i)}$
- Vrai / Faux Lorsqu'il est utilisé pour un ensemble d'observations à K classes, le modèle de classification "un contre tous" requiert l'entraînement de K modèles de classification binaires au minimum
- Vrai / Faux L'analyse discriminante Gaussienne est un exemple de modèle discriminant
- Vrai / Faux L'apprentissage d'un réseau de neurones par backpropagation nécessite le calcul de la dérivée de la fonction d'activation
- Vrai / Faux Le biais d'une famille de modèle encode la sensibilité de ces modèles à une perturbation des données d'entraînement tandis que la variance représente la capacité du modèle à capturer la tendance des données d'entraînement.
- Vrai / Faux Dans un réseau de neurones, un neurone muni d'une fonction d'activation sigmoïdale et dont la valeur de préactivation est proche de zéro peut être approximé par un modèle linéaire

1. [5pts] Indiquer si les affirmations suivantes sont vraies ou fausses

- Vrai / Faux La solution des équations normales est donnée par $\beta = X^T(X^T X)^{-1}t$ où t est le vecteur des valeurs cibles, X la matrice des caractéristiques et β le vecteur des coefficients de régression.
- Vrai / Faux Étant donné une distribution de probabilité $p(x; \theta)$ paramétrée par θ , l'estimateur de maximum de vraisemblance pour un échantillon de m observations indépendantes est donné par $\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^m p(x_i; \theta)$
- Vrai / Faux La distribution de Laplace est donnée par $Lap(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$
- Vrai / Faux Étant donné un ensemble de données comprenant deux classes distinctes, il est toujours possible de séparer les deux classes en considérant un nombre suffisamment grand de caractéristiques polynomiales de degré suffisamment élevé
- Vrai / Faux Le modèle de régression logistique peut séparer des classes qui ne sont pas linéairement séparables pourvu qu'il soit combiné avec des caractéristiques polynomiales
- Vrai / Faux La fonction sigmoïde est donnée par $\sigma(x) = e^x / (e^x + 1)$
- Vrai / Faux La validation croisée à k blocs divise un échantillon de données en k sous-échantillons parmi lesquels l'un des sous-échantillons est utilisé comme ensemble de validation et les $k - 1$ sous-échantillons restants sont utilisés comme ensemble d'entraînement