

Question 1 (sur 1 point)

Nous avons accès à un jeu de données d'images dermatoscopiques de lésions cutanées pigmentées. Pour chaque image, un dermatologue a annoté l'image pour indiquer le type de lésion, il y a 7 types possibles: carcinome de type 1, carcinome de type 2, kératose actinique, dermatofibrome, mélanome, naevus (grain de beauté), lésion vasculaire.

Notre but est de prédire le type de lésion à partir de son image.

- Quel est le type de la variable à prédire "type de lésion" ? Expliquez brièvement pourquoi.
- Quel est le type de problème (apprentissage supervisé, apprentissage non supervisé) ? Expliquez brièvement pourquoi.

Question 2 (sur 3 points)

Nous considérons à présent un autre jeu de données; pour 128 patients atteints de leucémie lymphoblastique aiguë, nous connaissons le type et le stade de tumeur (B1,B2,B3,B4,T1,T2,T3 ou T4; B pour tumeur de type B-cell, T pour tumeur de type T-cell). Pour chaque patient, nous avons également les données d'expression génétique; au total, cela donne 12 625 caractéristiques.

- Dans ce contexte, quelle est la variable à prédire et quelles sont les variables explicatives ?
- Est-ce qu'utiliser une régression linéaire ou une régression logistique est adaptée dans ce cas ? Expliquez pourquoi dans les deux cas.
- Quel est le défi posé par ces données ? Proposer un type de méthode (vu en cours) que l'on pourrait utiliser.

Question 3 (sur 2 points)

Nous considérons de nouveau le jeu de données d'images dermatoscopiques de lésions cutanées pigmentées, où nous avons accès au type de lésion de chaque image. Pour simplifier, je regroupe dans la même classe toutes les lésions malignes et dans l'autre classe toutes les lésions bénignes. Le but est donc maintenant de prédire à partir d'une image si la lésion est maligne ou bénigne.

J'entraîne sur ces données d'entraînement un réseau de neurones et j'obtiens sur ces données une accuracy globale de 98%, c'est-à-dire que dans 98% des cas, j'ai bien su déterminer si la lésion était bénigne ou non.

Est-ce un résultat que je peux présenter au dermatologue en lui disant que mon algorithme est performant ? Expliquer pourquoi et quelle démarche avoir si ce n'est pas la bonne.

Question 4 (sur 2 points)

Nous considérons le jeu de données ozone (vu en classe), dans le but prédire le maximum d'ozone à partir de la nébulosité et de la température (à 9h, 12h, 15h).

Nous appliquons une régression linéaire et utilisons l'algorithme de descente de gradient sans réfléchir aux hyperparamètres choisis. Le résultat est très mauvais et nous voyons que les valeurs du maximum d'ozone prédites sont très loins des valeurs réelles.

Proposer une explication.

Question 5 (QCM sur 2 point)

J'entraîne un réseau de neurones sur un jeu de données d'images (50 000 images) pour prédire leur classe (oiseau, camion, voiture, ...).

Cochez les cases qui vous semblent vraies. Vous pouvez cocher un ou plusieurs cases.

R1: Je relie forcément toutes mes entrées (pixels de l'image) à chacun des neurones.

R2: Pour optimiser les paramètres du réseaux de neurones, l'algorithme très simple du perceptron devrait marcher.

R3: Dans le réseau, j'utilise plusieurs couches de plusieurs neurones. Je réfléchis à l'architecture du modèle pour obtenir un résultat performant.