Perceptron et régression logistique

1 Cours

Répondre aux questions suivantes

1. Qu'est-ce que le problème de la classification supervisée ?
2. Quelles sont les hypothèses faites pour le problème de l'apprentissage supervisé à partir de n données ?
3. Quel est le problème de la généralisation ?

2 Apprentissage d'un perceptron

Soit l'ensemble d'apprentissage

$$\mathcal{S} = \left\{ \left(\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, +1 \right), \left(\mathbf{x}_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, -1 \right), \left(\mathbf{x}_3 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}, +1 \right), \left(\mathbf{x}_4 = \begin{bmatrix} -2 \\ -3 \end{bmatrix}, -1 \right), \left(\mathbf{x}_5 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, +1 \right) \right\}.$$

On va dérouler l'algorithme d'apprentissage présenté en cours et rappelé ici [**?**] :

- Classification binaire $y_i \in \{-1, +1\}$
- Initialisation $\mathbf{w} = \mathbf{0}$
- Répéter jusqu'à convergence ou bien atteinte d'un nombre max d'itérations
  - pour tous les exemples $(\mathbf{x}_p, y_p)$ faire
    - si $\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}}_p) = y_p$
      ne rien faire
    - sinon
      $\mathbf{w} \leftarrow \mathbf{w} + y_p \tilde{\mathbf{x}}_p$

1. Représenter sur le plan les points d'apprentissage $\mathbf{x}_1, \ldots, \mathbf{x}_5$
2. L'hyperplan cible, qui a permis d'étiqueter les exemples de $\mathcal{S}$ a pour equation $y = x$. Représenter cet hyperplan cible.
3. Donner les coordonnées des points $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_5$. A quoi sert cette transformation. Représenter graphiquement son utilité.
4. Dérouler l'algorithme du perceptron sur l'ensemble de données $\mathcal{S}$. Quelle est l'équation de l'hyperplan obtenu ? On utilisera la fonction sign telle que $\text{sign}(x) = 1$ si $x > 0$ et $\text{sign}(x) = -1$ sinon. Représenter graphiquement cet hyperplan. Comment faire pour que l'hyperplan appris s'approche de l'hyperplan cible ?
5. Selon vous, de quoi dépend le nombre d'itérations qu'il faut mettre en œuvre pour que l'algorithme s'arrête ? Avez-vous une idée de preuve ?

3 Perceptron linéaire

3.1 Convergence de l'algorithme

Dans cet exercice, nous montrons que l'algorithme d'apprentissage I du perceptron s'arrête en un maximum de R^2/γ^2 itérations si R est la norme maximale des $\mathbf{x}_i$ et $\gamma > 0$ est tel qu'il existe un vecteur $\mathbf{w}^*$ tel que $\mathbf{w}^* \cdot \mathbf{x}_i$.

1. Montrer qu'à chaque mise à jour $\mathbf{w}^{k+1}$ du vecteur courant $\mathbf{w}^k$ on a :

   $$\mathbf{w}^* \cdot \mathbf{w}^{k+1} \geq \mathbf{w}^* \cdot \mathbf{w}^k + \gamma.$$

   En déduire une borne inférieure sur $\mathbf{w}^* \cdot \mathbf{w}^k$ ne faisant intervenir que k et γ.
2. Monter également qu'à chaque mise à jour $\mathbf{w}^{k+1}$ du vecteur courant $\mathbf{w}^k$ on a :

   $$\|\mathbf{w}^{k+1}\|^2 \leq \|\mathbf{w}^k\|^2 + R^2$$

   En déduire une borne supérieure sur $\|\mathbf{w}^k\|^2$ ne faisant intervenir que k et R.
3. En déduire que le nombre maximal d'itérations de l'algorithme du perceptron est R^2/γ^2. (Indice : utiliser l'inégalité de Cauchy-Schwarz $\mathbf{u} \cdot \mathbf{v} \leq \|\mathbf{u}\|\|\mathbf{v}\|$.)

3.2 Fonction parité

Etant donné n entrées booléennes, $x_1, \ldots, x_n$ et un entier i entre 0 et n, construire un perceptron à fonction d'activation linéaire à seuil qui retourne 1 si et seulement si le nombre d'entrées égales à 1 est supérieur ou égal à i (il y aura bien sûr une cellule biais). Déduire de cette construction un perceptron à fonction d'activation linéaire à seuil possédant une couche cachée de n neurones et calculant la fonction *parité*, égale à 1 si le nombre d'entrées égales à 1 est pair et égale à 0 sinon.

4 Régression logistique

Dans cet exercice, on s'intéresse à la régression logistique, qui permet d'aborder des problèmes de classification supervisée. On se pose plus précisément le problème de la classification binaire. On suppose que l'espace auquel appartiennent les données est $\mathbb{R}^d$ et que l'espace des étiquettes est $\mathcal{Y} = \{0, 1\}$.

1. En régression logistique, la classification d'un exemple $\mathbf{x} \in \mathbb{R}^d$ se fait selon la valeur d'une fonction f dépendant d'un vecteur $\mathbf{w} \in \mathbb{R}^d$ et d'un réel b qui est définie par :

$$f(\mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x} + b))}$$

   (a) donner les valeurs minimale et maximale de f

   (b) proposer et justifier une règle d'affectation d'une classe à un exemple $\mathbf{x}$ en fonction de $f(\mathbf{x})$

   (c) quelles sont les limites de la classification par régression logistique (donner un exemple de problème qui ne peut être appris)

   (d) dans le cas où $d = 2$, $\mathbf{w} = [-2\ 1]^\top$ et $b = 1$, représenter graphiquement la surface de décision correspondant à la règle d'affectation proposée précédemment.

2. Soit un ensemble d'apprentissage $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\}$ et une distribution $p(X, Y)$ sur $\mathbb{R}^d \times \mathcal{Y}$. En s'aidant de $p(X, Y) = P(Y|X)p(X)$, écrire la vraisemblance des données de $\mathcal{S}$. Rappeler les hypothèses usuellement faites en classification supervisée.
3. Dans le cadre de la régression logistique, on fait l'hypothèse que la probabilité $P(Y = 1|X = \mathbf{x})$ qu'un exemple $\mathbf{x}$ soit de la classe $+1$ est égale à $f(\mathbf{x})$. En utilisant ce modèle statistique pour les données, et en s'inspirant de l'écriture de la vraisemblance pour une variable de Bernoulli, écrire la vraisemblance de l'échantillon $\mathcal{S}$ en fonction des y_i, des $p(\mathbf{x}_i) := p(X = \mathbf{x}_i)$ et de $\mathbf{w}$ et b.
4. Exprimer la log-vraisemblance $\mathcal{L}$ de l'échantillon $\mathcal{S}$ sous le modèle (de la régression) logisitique. Ne pas faire figurer les termes dépendant des $p(\mathbf{x}_i)$.
5. Quels sont les paramètres à estimer ?
6. Donner l'expression de $\frac{\partial \mathcal{L}}{\partial w_i}$, pour $i = 1, \ldots, d$ ainsi que l'expression de $\frac{\partial \mathcal{L}}{\partial b}$.
7. Le gradient de $\mathcal{L}$ peut-il s'annuler directement ? Proposer une équation de mise à jour (type gradient) permettant de produire une suite de paramètres menant à un maximum (local) de vraisemblance.
8. Ecrire l'algorithme complet d'apprentissage par régression logistique.
9. Proposer une version stochastique de l'algorithme d'apprentissage proposé précédemment. Dans le cas d'un pas d'apprentissage adaptatif, rappeler les conditions nécessaires et suffisantes sur le pas adaptatif pour être assuré de la convergence de l'algorithme. Rappeler les avantages de la version stochastique par rapport à la version stochastique.
10. En considérant l'ensemble d'apprentissage

$$\mathcal{S} = \left\{ \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, 0 \right) \right\}$$

    les valeurs initiales $\mathbf{w}^0 = [0\ 1]^\top$ et $b^0 = -1$ et un pas d'apprentissage fixe $\eta = 0.3$, faire deux itérations des algorithmes d'apprentissage proposés.