LELEC2870: séance d'exercices 1
Méthodes linéaires pour la régression

25 novembre 2013

1 Objectifs

Les méthodes linéaires permettent souvent d'obtenir de bon résultats en régression, tant que la relation entre les features (ou entrées) et la valeur à prédire (ou sortie) n'est pas trop non-linéaire. Au cours de cette première séance d'exercices, vous serez amenés à implémenter une régression linéaire pour un ensemble de features que vous sélectionnerez en utilisant des méthodes linéaires simples.

2 Dataset et notations

Les données de cette séance sont disponibles au format SciLab. Après avoir chargé les données, vous pourrez travailler sur un ensemble de 442 instances provenant d'une étude sur le diabète. Pour chacune de ces instances, les valeurs de 10 features sont fournies, ainsi que la valeur à prédire. Pour cette séance, le learning set est $\{(\mathbf{x}_p, t_p)|p = 1...P\}$ où P est le nombre d'instances, $\mathbf{x}_p$ est un vecteur (rangée) de dimension D

$$\mathbf{x}_p = \begin{pmatrix} x_p^1 & x_p^2 & \cdots & x_p^D \end{pmatrix} \tag{1}$$

et t_p est la valeur cible. Pour les calculs, les entrées sont placées dans la matrice

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_P \end{pmatrix} = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^D \\ x_2^1 & x_2^2 & \cdots & x_2^D \\ \vdots & \vdots & \ddots & \vdots \\ x_P^1 & x_P^2 & \cdots & x_P^D \end{pmatrix} \tag{2}$$

et les valeurs cible sont placées dans le vecteur (colonne)

$$\mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_P \end{pmatrix}. \tag{3}$$

Les features incluent l'âge (feature 1), le sexe (feature 2), le body mass index (feature 3) et la pression sanguine (feature 4) de chaque patient, ainsi que le résultat de différents test sanguins (features 5 à 10). Durant cette séance d'exercices, il peut être intéressant de lier vos résultats à l'interprétation des features.

3 Feature selection avec la corrélation

Dans la première partie de la séance d'exercices, vous devez implémenter une régression linéaire univariée. Puisque 10 features sont disponibles, nous vous demandons de sélectionner l'une d'entre elles avant de réaliser la régression. Un exemple de critère adéquat pour choisir cette feature est la corrélation entre la feature elle-même et la valeur à prédire.

Pour chaque feature, visualisez la dépendance entre la feature et la valeur à prédire. Calculez la corrélation. Qu'observez-vous ? Comment pouvez-vous utiliser la corrélation pour sélectionner une bonne feature pour faire de la régression ? Pensez-vous que votre choix est optimal à la fois pour des méthodes de régression linéaires et non-linéaires ?

4 Régression linéaire univariée

Une fois que vous avez sélectionné une feature avec la corrélation, vous pouvez faire une régression linéaire, qui peut être vue comme une réseau neuronal simple à une couche (see Fig. 1). L'expression analytique de la sortie est dans ce cas

$$y = \mathbf{x}\mathbf{w} + w_0 \tag{4}$$

où $\mathbf{w}$ est un vecteur (colonne) de poids, w_0 est le biais, σ est une fonction d'activation et y est la sortie du réseau. Afin de simplifier les notations, le biais est généralement inclus dans les vecteurs d'entrées et de poids, i.e. $\mathbf{x}$ devient

$$\begin{pmatrix} 1 & \mathbf{x} \end{pmatrix} \tag{5}$$

et $\mathbf{w}$ devient

$$\begin{pmatrix} w_0 \\ \mathbf{w} \end{pmatrix}. \tag{6}$$

Cette convention est utilisée dans le reste de cette séance d'exercices. Puisque sa fonction d'activation est linéaire, la sortie du réseau neuronal à une couche devient donc

$$y = \mathbf{x}\mathbf{w}. \tag{7}$$

Notez que $\mathbf{x}$ et $\mathbf{w}$ sont des vecteurs (colonne) de longueur $(D + 1)$.

En utilisant la méthode de la pseudo-inverse, entraînez une régression linéaire et visualisez le résultat de votre régression. En particulier, comparez vos prédictions avec les valeurs à prédire. Êtes-vous satisfait du résultat ? Comment pourriez-vous l'améliorer ?

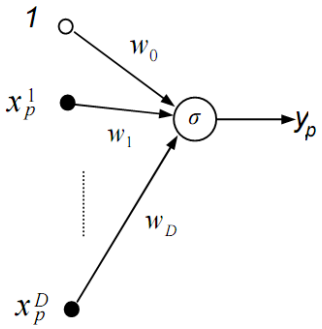

FIGURE 1 – Réseau neuronal à une couche.

5 L'apprentissage en tant que minimisation de l'erreur d'entraînement

En régression, la mean square error (MSE)

$$E(w|X,T) = \frac{1}{P}\sum_{p=1}^{P}(y_p - t_p)^2 = \frac{1}{P}\|\mathbf{y} - \mathbf{t}\|^2 = \frac{1}{P}\left[\sigma(\mathbf{X}\mathbf{w}) - \mathbf{t}\right]^T\left[\sigma(\mathbf{X}\mathbf{w}) - \mathbf{t}\right] \tag{8}$$

est généralement utilisée pour quantifier la qualité d'un modèle. Elle estime l'espérance du carré de l'erreur faite par le modèle. Quand cette erreur est calculée en utilisant le training set, elle devient une *erreur d'entraînement* et peut être utilisée pour apprendre des modèles.

Par design, la solution donnée par la méthode de la pseudo-inverse minimise la MSE d'entraînement. Vérifiez cette propriété (i) en visualisant la valeur de cette MSE d'entraînement en fonction des deux paramètres du modèle et (ii) en localisant la solution sur cette surface. Quelles sont les propriétés de cette surface d'erreur ?

6 Régression linéaire bivariée

Choisissez les deux meilleures features en utilisant la corrélation et entraînez une autre régression linéaire. Visualisez les résultats et comparez les prédictions aux valeurs à prédire. Trouvez un critère pour montrer que votre régression linéaire est capable de mieux approximer les valeurs à prédire en utilisant deux features qu'en utilisant une seule feature (par rapport aux données d'entraînement).