le cnam

Reconnaissance des formes et méthodes neuronales

Chapitre 3 : Régression

Michel Crucianu
http://cedric.cnam.fr/~crucianm/rfmn.html

le cnam

Méthodes de régression

- Modélisation à but décisionnel (prédictif)
- Objectif : déterminer une dépendance entre variables numériques, permettant de prédire les valeurs de certaines (variables expliquées, ou dépendantes, ou prédites, ou de sortie) à partir des valeurs prises par les autres (variables explicatives, ou « indépendantes », ou prédictives, ou d'entrée)
- Pourquoi utiliser les méthodes de régression :
  - La variable expliquée est difficile à mesurer alors que les variables explicatives sont faciles à mesurer
    - Exemple : estimer la résistance à la traction d'un polymère (dont la mesure destructive) à partir de variables caractérisant le monomère et le processus de polymérisation

le cnam

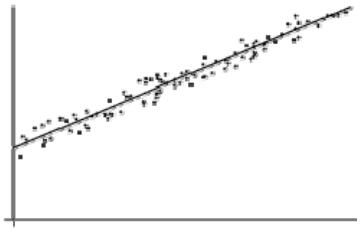
Méthodes de régression (2)

- Pourquoi utiliser les méthodes de régression (suite) :
  - Les valeurs des variables explicatives peuvent être connues avant celle de la variable expliquée et une prédiction de celle-ci est utile
    - Exemple : prédire le volume d'algues vertes sur des plages à partir de la quantité d'engrais utilisés lors de la campagne agricole précédente dans le bassin hydrographique correspondant et de la température moyenne de l'eau lors du trimestre précédent
  - Dans un ensemble de variables, on cherche à identifier celles dont dépend la variable expliquée
  - On cherche à savoir comment contrôler les variables d'entrée pour obtenir des valeurs désirées pour la variable de sortie
    - Exemple : la concentration d'un des produits d'une réaction chimique, à partir des quantités initiales de réactifs, de la température et de la pression

le cnam

Méthodes de régression (3)

- Critères d'évaluation :
  - Précision de la prédiction
  - Complexité du modèle
  - Lisibilité des dépendances



- Modèles linéaires généralisés : dépendance linéaire des paramètres
- Généralisations des modèles linéaires : modèles additifs généralisés, modèles non linéaires

le cnam

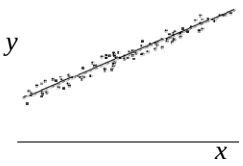
Régression linéaire

- Objectif : obtenir un modèle linéaire avec une dépendance également linéaire des variables explicatives
- Intérêts
  - Simplicité de modélisation (estimation des paramètres) et de prévision
  - Lisibilité : la forme additive rend explicite la dépendance des variables prédictives individuelles (valable pour tout modèle additif)
  - Moyens statistiquement bien fondés d'évaluation
- Forme générale du modèle pour 1 variable de sortie et *m* variables d'entrée : $Y = \omega_0 + \sum_{j=1}^{m} \omega_j X_j + \varepsilon$
  avec tirages indépendants identiquement distribués (iid)

le cnam

Estimation des paramètres

- On considère un ensemble de *n* observations,
  $D_n = \{(y_i, \mathbf{x}_i)\}_{1 \le i \le n}$
  pour lesquelles on écrit (e_i sont les résidus)
  $y_i = w_0 + \sum_{j=1}^{m} w_j x_{ji} + e_i \qquad 1 \le i \le n$



- On a aussi $\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}$, où une colonne de 1 a été introduite dans la matrice $\mathbf{X}$, permettant d'inclure w_0 dans $\mathbf{w}$
- Estimation des paramètres : méthode des moindres carrés où on cherche à minimiser $C(\mathbf{w}) = \sum_{i=1}^{n} (e_i)^2$

le cnam

Estimation des paramètres (2)

- Solution : $\mathbf{w}^* = \mathbf{X}^+ \mathbf{y}$, $\mathbf{X}^+$ étant la pseudo-inverse Moore-Penrose de la matrice $\mathbf{X}$ (en général non carrée)
- Si ${}^T\mathbf{X}\,\mathbf{X}$ est inversible, alors $\mathbf{X}^+ = \left({}^T\mathbf{X}\,\mathbf{X}\right)^{-1}\,{}^T\mathbf{X}$
- Problèmes : singularité ou mauvais conditionnement de ${}^T\mathbf{X}\,\mathbf{X}$

  → réduction nécessaire du nombre *m* de variables explicatives, utilisant par exemple l'ACP, en conservant uniquement les composantes correspondant aux valeurs propres > 0 et assurant un bon conditionnement ($\lambda_{\max}/\lambda_{\min} \leq$ seuil numéro de condition)

25 septembre 2013 RCP208 7

le cnam

Interprétation probabiliste

- On considère $Y = \omega_0 + \sum_{j=1}^{m} \omega_j X_j + \varepsilon$, avec ε suivant $\mathcal{N}\left(0, \sigma^2\right)$
- Étant données les observations D_n et un modèle $\mathbf{w}$, la vraisemblance (*likelihood*) est définie par $l\left(\mathbf{w} \middle| D_n\right) = p\left(D_n \middle| \mathbf{w}\right)$
- Si les observations sont iid, alors $p\left(D_n \middle| \mathbf{w}\right) = \prod_{i=1}^{n} p\left(y_i, \mathbf{x}_i \middle| \mathbf{w}\right)$
- Question : quel modèle maximise la vraisemblance ?
- Avec les hypothèses ci-dessus, on peut facilement montrer (voir page suivante) que la solution $\mathbf{w}^*$ des moindres carrés est également celle du maximum de vraisemblance

25 septembre 2013 RCP208 8

le cnam

Interprétation probabiliste (2)

- On suppose que ε suit une loi $\mathcal{N}(0,\sigma^2)$
- En explicitant la loi normale et en remplaçant ε par son expression on obtient

$$p(y_i,\mathbf{x}_i\,|\,\mathbf{w}) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma^2}\left(y_i - w_0 - \sum_{j=1}^{m} w_j x_{ji}\right)^2}$$

et donc

$$p(D_n\,|\,\mathbf{w}) = \prod_{i=1}^{n} p(y_i,\mathbf{x}_i\,|\,\mathbf{w}) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - w_0 - \sum_{j=1}^{m} w_j x_{ji}\right)^2}$$

- Cette expression étant strictement monotone et décroissante dans l'argument

$$C(\mathbf{w}) = \sum_{i=1}^{n}\left(y_i - w_0 - \sum_{j=1}^{m} w_j x_{ji}\right)^2$$

ses maxima par rapport à $\mathbf{w}$ correspondent aux minima de $C(\mathbf{w})$

⇒ La solution du maximum de vraisemblance est donc celle des moindres carrés

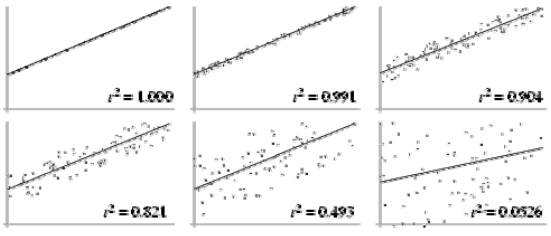
le cnam

Validité du modèle

- Somme des carrés des résidus : $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- Somme des carrés de régression : $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
- Somme totale des carrés : $\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

(équation d'analyse de la variance)

- Coefficient de corrélation multiple (coefficient de détermination) :



$$r^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

le cnam

Validité du modèle (2)

- Degrés de liberté : nombre de composantes indépendantes qui contribuent à chaque somme
  - $n-1$ pour $\sum_{i=1}^{n}(y_i-\bar{y})^2$, m pour $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$
  - $n-m-1$ pour $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$
- Test global du modèle : les observations peuvent-elles être le fruit du hasard ? Revient à tester l'hypothèse H_0 : tous les ω sont 0
  - On peut montrer que le rapport $\dfrac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \Big/ m}{\sum_{i=1}^{n}(y_i-\bar{y})^2 \Big/ (n-m-1)}$ suit une distribution $F(m, n-m-1)$
  - Si la valeur de ce rapport dépasse la valeur limite de la distribution pour un niveau de confiance recherché, on rejette l'hypothèse nulle H_0 et on peut considérer que la relation linéaire trouvée est significative

le cnam

Validité du modèle (3)

- La matrice des covariances pour l'estimation $\hat{\omega} = \mathbf{w}^*$ est $\left({}^T\mathbf{X}\,\mathbf{X}\right)^{-1}\sigma^2$ et permet de tester les paramètres individuellement :
  - Le rapport $w_j \Big/ \sqrt{v_j}$ (v_j étant l'élément j de la diagonale de $\left({}^T\mathbf{X}\,\mathbf{X}\right)^{-1}\sigma^2$ c'est à dire la variance de l'estimation de w_j) suit une distribution $t(n-m-1)$, on peut donc tester chaque hypothèse « w_j est 0 »
- Choix entre 2 modèles, $\mathcal{M}$ avec m paramètres, $\mathcal{M}^*$ avec $p > m$ paramètres :
  - Le rapport $\dfrac{\left[\sum_{i=1}^{n}\left(\hat{y}_i^*-\bar{y}\right)^2-\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2\right] \Big/ (p-m)}{\left[\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}\left(\hat{y}_i^*-\bar{y}\right)^2\right] \Big/ (n-p-1)}$ suit une distribution $F(p-m, n-p-1)$

  ⇒ On peut tester si la différence entre les 2 modèles est significative

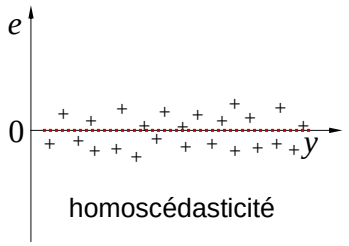
le cnam

Construction et inspection du modèle

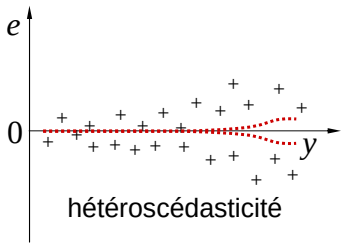
- Approches :
  - Incrémentale : ajout de nouvelles variables une par une (à chaque pas, celle qui réduit le plus la somme des carrés des résidus) et nouveau calcul des paramètres de régression
  - Par élimination : élimination progressive de variables (à chaque pas, celle dont l'exclusion fait augmenter le moins la somme des carrés des résidus) et nouveau calcul des paramètres de régression
- Inspection pour évaluer l'adéquation des hypothèses initiales :
  - Vérifier la constance de la variance des résidus (homoscédasticité)
  - Vérifier le caractère linéaire de la dépendance entre les variables
  - → Différents graphiques, par exemple celui des résidus par rapport à la variable expliquée (on reste en 2 dimensions !)

le cnam

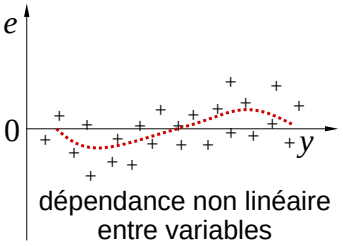
Exemples d'inspection



homoscédasticité



hétéroscédasticité



dépendance non linéaire entre variables

le cnam

Généralisations de la régression

- Modèle linéaire généralisé : Y peut ne pas être normale et

  $g(E[Y]) = \omega_0 + \sum_{j=1}^{m} \omega_j X_j$, $g(\cdot)$ étant monotone et différentiable

- Modèle additif généralisé :

  $g(E[Y]) = \omega_0 + \sum_{j=1}^{m} f_j(X_j)$, $f_j(\cdot)$ étant continues, arbitraires

- Modèle non linéaire :

  $Y = f(X_1, \ldots, X_m, \varepsilon)$

le cnam

Choix de modèle en régression