La régression logistique

Votre premier modèle de classification

Rafik LACHAAL

La régression logistique.

- Algorithme de classement
- Bon compromis entre performance du modèle et pouvoir explicatif
- Peut se généraliser à des valeurs de sortie multiclasses.
- Admet en entrée des variables qualitative et quantitative

La régression logistique

- Pour rappelle un modèle linéaire s'écrit :

$$h(X) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_n X_n$$

- On peut aussi l'écrire sous forme matriciel :

$$h(X) = \theta^T X + b$$

La régression logistique.

- L'objectif de la régression logistiques est de prédire la probabilité d'appartenir à une classe plus qu'une autre.
- Pour se faire il suffit seulement de ramener la valeur donné par notre modèle linéaire entre 0 et 1.
- Pour se faire on utilise la fonction logistique :

$$f(x) = \frac{1}{1 + \exp(-x)}$$

La régression logistique.

- Fonction hypothèse :

$$logisticRegression(x) = \begin{cases} 0, h(x) < 0.5 \\ 1, h(x) > 0.5 \end{cases}$$

$$0 < h(x) < 1$$

$$h(x) = P(y = 1|x, \theta) = \frac{1}{1 + \exp(-(\theta^T x \ + b))}$$

La fonction coût ce minimise aussi grâce à la descente de gradient.

Indicateurs pour les problèmes de classification

- Indicateur basique : **la matrice de confusion**

Elle met en regard les données prédites et les données observées.

		Observations		
		+	**-**	Total
Prédictions	**+**	250	150	400
	-	50	550	600
	Total	300	700	1000

Taux d'erreur : (50 + 150)/1000 = 20%

Sensibilité (rappel ou recall, taux de vrais positifs): 250/(250 + 50) = 83%

Précision (Proportion de positifs parmi tout les positifs prédits): 250/(250 + 150) = 63%

Spécificité (taux de vrais négatifs): 550/(150 + 550) = 78%

Indicateurs pour les problèmes de classification

- **La courbe ROC :**

Receiver Operating Caracteristique = Fonction d'efficacité du récèpteur

Pour toute les méthodes de classifications le classement se fait en calculant un score. Si ce score est supérieur à un certain seuil s est attribué à une classe, sinon à l'autre.

La sensibilité et la spécificité sont des fonction de s.

Indicateurs pour les problèmes de classification

- **La courbe ROC :**

Si on note a(s) la sensibilité en fonction du seuil et β(s) la spécificité en fonction du seuil, alors on peut tracer le graphique de la courbe ROC en mettant en abscisse a(s) et en ordonné 1- β(s).

Chaque point de la courbe ROC est donnés par le couple

(a(s) , 1- β(s)) avec s

Toutes les valeurs de la courbe ROC sont comprise entre (0, 0) et (1, 1).

Indicateurs pour les problèmes de classification

- **La courbe ROC :**

L'analyse de la courbe ROC va permettre de choisir un seuil de décision optimal.

http://www.navan.name/roc/

Indicateurs pour les problèmes de classification

- **L'AUC** (Area Under the Curve) :

Représente la surface sous la courbe :

$$AUC = \int_{s=+\infty}^{s=-\infty} \big(1 - \beta(s)\big) d\alpha(s)$$

Permet de comparer les modèles de différents types en utilisant la courbe ROC.

Indicateurs pour les problèmes de classification

- **L'AUC** (Area Under the Curve) :

Plus grande est l'AUC et meilleur est le modèle