Memo pour l'année

LAURENT Thomas

Master 2 informatique 2018

Contents

Chapter 1

Fouille de donnée

1.1 Pré traitement des données

1.1.1 Nettoyage des données

Caractéristiques descriptives

Objectifs: Résumer, décrire certains aspects (tendances, variation, dispersion...) des données en utilisant certaines mesures :

Moyenne (espérance) : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Ecart moyen : $\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$

Variance : $v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Ecart type : $\alpha x := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n}(\sum_{i=1}^{n} x_i^2) - \bar{x}^2}$

Médiane : Valeur se trouvant au milieu d'une série de données ordonnées

Mode :Valeur la plus fréquente

Amplitude :min, max

1.1.2 Normalisation

Min-max : $v_n = \frac{v - v_{min}}{v_{max} - v_{min}}$

Min-max dans l'intervalle [A,B]: $v_n = \frac{v - v_{min}}{v_{max} - v_{min}} * (B - A) + A$

Z-Score : $v_n = \frac{v - moyenne}{ecart_t ype}$

Decimal scaling : $v_n = \frac{v}{100^j}$

1.2 Classification

1.2.1 Évaluation des classifieurs

Matrice de confusion

Percent of correct classification :

PCC(%) $:= \frac{N_c}{N_t} * 100$

N_c : nombre d'instances correctement classées

N_t : nombre d'instances testées ($N_t = |D_{test}|$)

Exemple:

$$: \begin{pmatrix} \_ & c1 & c2 & c3 & c4 \\ c1 & 0 & 1 & 0 & 0 \\ c2 & 1 & 60 & 0 & 1 \\ c3 & 0 & 1 & 23 & 0 \\ c4 & 1 & 0 & 7 & 5 \end{pmatrix}$$

Taux d'erreurs : 100-PCC

PCC(%) $= \frac{0+60+23+5}{100} * 100 = 88\%$

1.3 Arbre de décision

1.3.1 critères de sélection C4.5

Construction d'un arbre de décision C4.5 La construction d'un arbre de décision avec C4.5 passe par deux phases:

Phase d'expansion : La construction se fait selon l'approche descendante et laisse croître l'arbre jusqu'à sa taille maximale.

Phase d'élagage : Pour optimiser la taille l'arbre et son pouvoir de généralisation, C4.5 procède à l'élagage (pour supprimer les sous-arbres qui ne minimisent pas le taux d'erreurs)

Approche de construction d'un AD : Partitionner récursivement les données en sous-ensembles plus homogènes ... jusqu'à obtenir des partitions qui contiennent des objets qui appartiennent majoritairement à la même classe.
=> Théorie de l'information pour caractériser le degré de mélange, homogénéité, impureté, incertitude...

Théorie de l'information : Théorie mathématique ayant pour objet l'étude du contenu informationnel d'un message.
Applications en codage, compression, sécurité...

Entropie : Mesure la quantité d'incertitude dans une distribution de probabilités.

Rappel sur les probabilisées

Quelques rappels de probabilités : Soient X et Y deux variables aléatoires discrètes prenant leurs valeurs dans DX=x1,..,xn et DY=y1,..,ym respectivement.

$P(x_i) = \frac{|x_i|}{\sum_{j=1}^{n}|x_j|}$

$\sum_{i=1}^{n} P(x_i) = 1$

$P(x_i|y_i) = \frac{P(x_i,y_i)}{p(y_i)}$

$P(x_i, y_i) = p(x_i) * p(y_i)$ Si X et Y sont indépendantes

Exemple:

$$:\begin{pmatrix} Anne & Sexe & \# & \% \\ M1 & M & 25 & 25/55 \\ M1 & F & 4 & 4/55 \\ M2 & M & 25 & 25/55 \\ M2 & F & 1 & 1/55 \end{pmatrix}$$

$P(sexe = M) = P(Sexe = MetAnne = M1) + P(Sexe = MetAnne = M2) = 50/55$

$P(Anne = M2|sexe = M) = P(Sexe = MetAnne = M2)/P(Sexe = M) = \frac{25}{55}/\frac{50}{55} = \frac{25}{50} = \frac{1}{2}$

Entropie

Entropie : Mesure la quantité d'incertitude (manque d'information) dans une distribution de probabilités. Soit X une variable aléatoire discrète prenant ses valeurs dans $DX = x1, .., xn$. Soit P la distribution de probabilités associée à X.

$H(X) = -\sum_{i=1}^{n} p(x_i) * log_2(p(x_i))$

Par convention, quand $p(x) = 0, 0 * log(0) = 0$

Exemple:

X	P(X)
x_1	1/3
x_2	1/3
x_3	1/3

$H(X) = -p(x_1) * log_2(p(x_1)) - p(x_2) * log_2(p(x_2)) - p(x_3) * log_2(p(x_3))$

$H(X) = -3(\frac{1}{3} * log_2(\frac{1}{3})) = log_2(3) = 1.58$

Autre exemples:

$[\frac{1}{2}, \frac{1}{4}, \frac{1}{4}] : H(X) = 1.5$

$[1, 0, 0] : H(X) = 0$

$[\frac{1}{2}, \frac{1}{2}] : H(X) = 1$

Propriétés:

$H(X) >= 0$

$H(X)$ est maximale pour une distribution uniforme (toutes les valeurs sont équiprobables).

Entropie conjointe : L'entropie conjointe de deux variables aléatoires X et Y est l'incertitude relative à ces deux variables conjointement.

$H(X, Y) = -\sum_{i,j=1}^{n} p(x_i, y_i) * log_2(p(x_i, y_i))$

Exemple : $[0.2, 0.1, 0.3, 0.4] : H(X, Y) = 1.85$

Critère de sélection: Gain d'information:

$GAIN(T, A) = Info(T) - Info(T|A)$

Avec $Info(T)$: Entropie au niveau de T (avant de partitionner)

$Info(T) = -\sum_{c_i} freq(c_i, T) * log_2(freq(c_i, T))$

Avec $freq(c_i, T) = p(c_i) = \frac{|c_i|}{|T|}$

Avec $Info(T|A)$ l'entropie conditionnelle de T une fois partitionné selon les valeurs de l'attribut A.

$Info(T|A) = \sum_{a_j \in A} freq(a_j, T) * Info(T|a_j)$

Critère de sélection: Gain Ration:

Le gain d'information favorise les attributs ayant de larges domaines.

Le ratio de gain utilise le gain d'information avec un facteur pénalisant les attributs ayant des domaines trop larges.

$GainRatio(T, A) = \frac{Gain(T,A)}{Split_Info(T,A)}$

Avec $Split_Info(T, A) = -\sum_{a_j \in A} freq(a_j, T) * log_2(freq(a_j, T)) = EntropiedeA$

Chapter 2

Recherche Opérationnel

gggg

Chapter 3

Apprentissage par le pratique

gggg

Chapter 4

Représentation des connaissances et raisonnement

ggggg

Chapter 5

Outils formel

yyyyyy

Chapter 6

XML

uuuuu

Chapter 7

Anglais

ffffffff