Formulas

Wed, 08 Jun 2022 at 02:48

Text Analysis

#idf #search #tf-idf #recall #percision

TERM FREQUENCY

$$tf(t,d) = count(t,d)$$

INVERSE DOCUMENT FREQUENCY

$$idf(t) = log_{10}(\frac{N}{df(t)})$$

> *Where $df(t)$ is the number of documents that contain the term t*

TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY

$$tfidf(t,d) = tf(t,d) \times idf(t)$$

DOCUMENT RELEVANCE

$$relevance(d) = \sum_{i=0} tfidf(t_i, d)$$

Logistic Regression

#logit #link #function #odds-ratio

$$P(event) = \frac{outcome\, of\, interest}{all\, possible\, outcomes}$$

$$odds(event) = \frac{P}{1-P}$$

$$odds\, ratio = \frac{odds(event_1)}{odds(event_2)}$$

LOG ODDS - LOGIT FUNCTION

$$logit(p) = \ln(\frac{P}{1-P}) = \ln(p) - \ln(1-p)$$

SIGMOID FUNCTION - INVERSE LOGIT FUNCTION - ANTILOG

$$logit^{-1}(\alpha) = \frac{e^\alpha}{1+e^\alpha} = \frac{1}{1+e^{-\alpha}}$$

ESTIMATED PROBABILITY THROUGH THE LOGIT LINK FUNCTION

$$\ln(\frac{P}{1-P}) = \beta_0 + \beta_1 x_1$$

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$$

> *When we invert the logit expression we reach the odds-ratio expression used in interpreting the results Logistic Regression > Interpetation*

$$(\frac{P(y=1)}{1-P(y=1)}) = e^{\sum_{j=0}^{k}\beta_j x_j}$$

> *And per the properties of exponentials can be expressed as*

$$(\frac{P(y=1)}{1-P(y=1)}) = \prod_{j=0}^{k} e^{\beta_j x_j}$$

PSUEDO-R^2

$$psuedo = R^2 = 1 - \frac{deviance}{null\,deviance}$$

1. deviance is your model's variance
2. null deviance is calculated as $\frac{positive\,class}{total\,records}$
   - another way is by taking the $log(odds)$ of the positive class and projecting it back to a probability

Decision Trees

#entropy #information #gain #classification

BASE ENTROPY

$$H_C = -\sum_{i=0}^{k} P(C_i)\log_2 P(C_i)$$

1. The minimum value of entropy is 0 which means the node is 100% pure
2. The maximum value of entropy is $\log_2 k \;\; ; \;\; k = no.\,of\,classes$

CONDITIONAL ENTROPY

$$H_{C|X} = -\sum_{x\in X} P(x)\sum_{i=0}^{k} P(C_i|x)\log_2 P(Ci|x)$$

INFORMATION GAIN

$$InfoGain = H_C - H_{C|X}$$

GINI INDEX

$$Gini_x = 1 - \sum_{\forall x\in X} P(x)^2$$

Naïve Bayes

#bayes #theorem #classification

POSERIORI PROBABLITY

$$P(C_k|X) = \frac{P(X|C_k) \times P(C_k)}{P(X)} \propto P(X|C_k) \times P(C_k) \quad ; \forall\ k \in k\ classes$$

1. $P(X|C_k)$ is the likelihood
2. $P(X)$ is the prior probability for the predictor
3. $P(C_k)$ is the initial or prior probability of the class

$$P(X|C_k) = \prod_{i=0}^{n} P(x_i|C_k)$$

Model Evaluation and Selection

#accuracy #recall #percision #tpr #fpr

ACCURACY - RECOGNETION RATE

$$accuracy = \frac{TP + TN}{P + N}$$

ERROR RATE

$$error\,rate = 1 - acccuracy = \frac{FP + FN}{P + N}$$

RECALL - SENSITIVITY - TPR

$$reacll = \frac{TP}{P} = \frac{TP}{TP + FN}$$

SPECIFICITY - TNR

$$specificity = \frac{TN}{N}$$

PRECISION

$$percision = \frac{TP}{P'} = \frac{TP}{TP + FP}$$

$$percision \propto \frac{1}{recall}$$

FPR

$$fpr = 1 - specifity = \frac{FP}{FP + TN}$$

F-SCORE - F_1 - HARMONIC MEAN

$$F_1 = 2 \times \frac{percision \times recall}{percision + recall}$$

FBETA - F_B

> *Value of β is chosen such that recall is considered β times more important than precision.*

$$F_\beta = (1 + \beta^2) \times \frac{percision \times reacll}{(\beta^2 \times percision) + reacll}$$