Stats for Data Science

Saumya Bhatnagar

February 12, 2020

Table of contents

Regression, Classification, Clustering

Regression

1. Linear
2. KNN
3. SVM
4. Random Forest

Classification

1. Logistic
2. KNN
3. SVM Classifier
4. Random Forest

Clustering

1. K-Means
2. Hierarchical
3. DBSCAN
4. HDBSCAN

Regression analysis is a statistical technique to assess the relationship between an predictor variable and one or more response factors.

Outcome Variable	**GLM Family**	**Link**	**Mean to Variance**
Continuous, unbounded	Normal or Standard Gaussian	Identity	
Continuous, non-negative	Gamma or inverse Gamma		
Discrete/ counts/ rate	Poisson Quassi-poisson or negative binomial	Log	Identity If not Identity
Count	Gamma		Over dispersion
Counts with multiple zero	Zero inflated poisson may be checked for fitting		
Binary	Binomial or Logistic regression		
Nominal	Multinomial regression		

Regression Model Selection Criteria

Three methods to classifier

1. model a classification rule - knn, decision tree, perceptron, svm
2. model the probability of class membership given input data - perceptron with cross-entropy cost
3. make a probabilistic model of data within each class - naive bayes 1 & 2 are discriminative classifications 3 is generative classification 2 & 3 probabilistic classification

Monothetic: Cluster members have some common property
Polythetic: Cluster members are similar to each other. Distance between elements define relationship

Life Time Value (LTV)

content...

Propensity of Cross-sell

content...

Thank You!