00MachineLearning

command

StatQuest - machine learning

Bias Variance - trade off

The most important thing that how doing a SSE is, but how it performs with Testing Data

Decision Tree

Machine learning is all about making predictions and classifications

Cross validation: when we to compare different model learning methods ... we get a idea of how well they work in practice

Logistic regression, k-nearest neighbor, support vector machines, random forest

Confusion matrix:

Sensitivity: ... % of positives correctly identified

Specificity: ... % of negatives correctly identified

Linear regression & least squares

Bias: the inability for a machine learning method to capture a true relationship

Variance: difference in fits between datasets ... tightening & overfitting

Sweetspot: between a simple model and a complex model ... regularization, boosting, and bagging are methods to find sweetspot

R^2 ...

Relative & transition & Entropy: ? prediction of positive events correctly classified

RSS and MSE in R

$Surprise = \log\left(\frac{1}{probability}\right)$ ... Entropy: expected surprise of an event ...

Mutual Information: $\sum \sum$ ...

Entropy is not an average, the specific of change of information, not tied to one variable ...

Sum of Squares around the mean ... variation around the mean

$R^2 = \frac{\text{Variance}(mean) - \text{Variance}(line)}{\text{Variance}(mean)}$

degree of freedom

t-tests, p-value

Odds: ratio of ...

p-values: ...

"does nothing a "scale of evidence" so statistically significant"? Fisher's exact test, chi-square test

Logistic regression predicts True or False instead of ...

Log(odds/ratio) ...

Logistic regression R^2 ...

Log Likelihood

Saturated models and Deviance

(second page)

Logistic regression in R

Design Matrix

Regularization: Ridge Regression, Lasso Regression

Elastic-Net Regression:

Principle Component Analysis: ... Eigenvalue = ...

Linear Discriminant Analysis: like PCA, but focus on maximizing separability among known categories

MDS and PCoA ...

PCA ...

t-SNE

Hierarchical clustering, k-means clustering, DBSCAN, nearest neighbor

Naive Bayes: simple probabilities are also called Likelihood ...

Gaussian Naive Bayes ... for continuous variables

Decision and Classification Trees ... gene capacity, information gain

Feature Selection preventing data? lot of ways to guess what missing data might be

Regression Trees ... numerical ...

Cost complexity Pruning: ...

One-Hot encoding, Label encoding, Target encoding, Bayesian Mean encoding, K-fold target encoding

Classification Trees in Python

Random Forest: ... randomized data set, ...

... and then from a heat map or an MDS plot to show how samples are related by predictions

Gradient Boost: ...

Stochastic gradient descent ...

Adaboost, Gradient Boost, XgBoost

XgBoost in Python

Cosine Similarity = $\frac{\sum A_i B_i}{\sqrt{\sum A_i^2}\sqrt{\sum B_i^2}}$ ... Support Vector Machines: maximum margin classifier

... Radial kernel: $e^{-\gamma(a-b)^2}$