Fundamentals reference

Friday, September 28, 2018 10:04 AM

Problems on Multicollinearity
The idea is that you can change the value of one independent variable and not the others. However, when **independent variables are correlated, it indicates that changes in one variable are associated with shifts in another variable**. The stronger the correlation, the more difficult it is to change one variable without changing another. It **becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable *independently* because the independent variables tend to change in unison.**
Multicollinearity reduces the precision of the estimate coefficients,
Multicollinearity affects the coefficients and p-values, but it does not influence the predictions, precision of the predictions, and the goodness-of-fit statistics. If your primary goal is to make predictions, and you don't need to understand the role of each independent variable, you don't need to reduce severe multicollinearity

From <http://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/>

Testing for Multicollinearity with Variance Inflation Factors (VIF)

LR concepts
https://hackernoon.com/an-intuitive-perspective-to-linear-regression-7dc566b2c14c

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No auto-correlation
- Homoscedasticity

From <http://www.statisticssolutions.com/assumptions-of-linear-regression/>

Log R funda
https://hackernoon.com/introduction-to-machine-learning-algorithms-logistic-regression-cbdd82d81a36

SVM :-
https://www.ibm.com/support/knowledgecenter/de/SS3RA7_15.0.0/com.ibm.spss.modeler.help/svm_howwork.htm
https://machinelearningmastery.com/support-vector-machines-for-machine-learning/

https://machinelearningmastery.com/naive-bayes-for-machine-learning/
Bootstrap Aggregation is a general procedure that can be used to reduce the variance for those algorithm that have high variance
Decision trees are sensitive to the specific data on which they are trained. If the training data is changed (e.g. a tree is trained on a subset of the training data) the resulting decision tree can be quite different and in turn the predictions can be quite different.

From <https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/>

From <https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/>

MISSING DATA:-https://medium.com/ibm-data-science-experience/missing-data-conundrum-exploration-and-imputation-techniques-9f40abe0fd87

https://machinelearningmastery.com/handle-missing-data-python/

Sklearn preprocessing imputer

```
from sklearn.preprocessing import Imputer
```

From <https://stackoverflow.com/questions/30317119/classifiers-in-scikit-learn-that-handle-nan-null>

```
# Create our imputer to replace missing values with the mean e.g.
imp = Imputer(missing_values='NaN', strategy='mean', axis=0)
imp = imp.fit(X_train)
# Impute our data, then train
X_train_imp = imp.transform(X_train)
 clf = RandomForestClassifier(n_estimators=10) clf = clf.fit(X_train_imp, Y_train)
```

KNN
Statistical Imputation

Models like XGBoost/Light GBM can handle missing values

VANISHING GRADIENT
https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-

lstm/

www.deeplearningbook.org

CNN Limitation:-

- In Machine Learning. a convolutional neural network is a class of deep,feed forward artificial neural networks that has successfully been applied to analyzing visual imagery. A Convolutional Neural Networks has some drawbacks some are listed below
- Hyperparamter tuning is non-trivial
- Need a large dataset
- The scale of a net's weights (and of the weight updates) is very important for performance. When the features are of the same type (pixels, word counts, etc), this is not a problem. However, when the features are heterogeneous--like in many Kaggle datasets--your weights and updates will all be on different scales (so you need to standardize your inputs in some way).
- cost effective
- A convolution is a significantly slower operation than, say maxpool, both forward and backward. If the network is pretty deep, each training step is going to take much longer.

Xgboost vs light gbm
https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/
he leaf-wise algorithm can reduce more loss than the level-wise algorithm

https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

http://zhanpengfang.github.io/418home.html

https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/

Xgboost
https://www.kdnuggets.com/2017/10/xgboost-top-machine-learning-method-kaggle-explained.html

https://www.slideshare.net/JaroslawSzymczak1/xgboost-the-algorithm-that-wins-every-competition

https://www.youtube.com/watch?v=s3VmuVPfu0s
Jaroslaw Szymczak - Gradient Boosting in Practice: a deep dive into xgboost

From <https://www.youtube.com/watch?v=s3VmuVPfu0s>

Pydata

Why NN is imp
ANN is nonlinear model that is easy to use and understand compared to statistical methods. ANN is non-paramateric model while most of statistical methods are parametric model that need higher background of statistic. ANN with Back propagation (BP) learning algorithm is widely used in solving various classification

Random Forest
https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

The advantages of random forest are:

- It is one of the most accurate learning algorithms available.

Why NN is imp
ANN is nonlinear model that is easy to use and understand compared to statistical methods. ANN is non-paramateric model while most of statistical methods are parametric model that need higher background of statistic. ANN with Back propagation (BP) learning algorithm is widely used in solving various classification and forecasting problems. Even though BP convergence is slow but it is guranteed. However, ANN is black box learning approach, cannot inteprete relationship between input and output and cannot deal with uncertainties. To overcome this several approches have been combined with ANN such as feature selection and etc.
Meanwhile Fuzzy is quite good in handling uncertainties and can inteprete relationship between i/o by producing rules. Therefore, to increase the capability of Fuzzy and ANN, hybridization of ANN and fuzzy is usually implemented.

algorithm-d457d499ffcd

The advantages of random forest are:

- It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier.
- It runs efficiently on large databases.
- It can handle thousands of input variables without variable deletion.
- It gives estimates of what variables are important in the classification.
- It generates an internal unbiased estimate of the generalization error as the forest building progresses.
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.
- It has methods for balancing error in class population unbalanced data sets.
- Generated forests can be saved for future use on other data.
- Prototypes are computed that give information about the relation between the variables and the classification.
- It computes proximities between pairs of cases that can be used in clustering, locating outliers, or (by scaling) give interesting views of the data.
- The capabilities of the above can be extended to unlabeled data, leading to unsupervised clustering, data views and outlier detection.
  It offers an experimental method for detecting variable interactions
  Disadvantages:-
- Random forests have been observed to overfit for some datasets with noisy classification/regression tasks.
- For data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels. Therefore, the variable importance scores from random forest are not reliable for this type of data.

Linear vs Logistic
https://stats.stackexchange.com/questions/22381/why-not-approach-classification-

through-regression