Data Science Concepts

Datakick Analytics
http://datakickanalytics.com

A list of concepts, techniques and technologies for data science practitioners

Statistics

- Population, Sample
- Probability
- Distribution
- Dependence, Independence
- Correlation, Covariance
- Expected values
- Mean, Median, Mode
- Standard deviation, Variance
- Outlier identification
- Hypothesis testing
- Type I and Type II errors
- Degrees of freedom
- Generalized likelihood
- Maximum likelihood
- Bayes Theorem
- Mutual information

Numerical Analysis

- Difference equations
- Linear systems of equations
- Nonlinear systems of equations
- Monte Carlo methods
- Numerical integration

Optimization

- Linear programming
- Gradient Descent
- Evolutionary methods
- Genetic algorithms
- Simulated annealing
- Stochastic methods
- Exploration vs. Exploitation

Relational Databases

- SQL
- Tables
- Schema
- Insert, update, select, join
- Views

Linear Regression

- Linear regression
- Multiple linear regression
- Least squares
- Confidence intervals
- t-statistics
- p-value
- Residuals
- Multicollinearity
- Heteroskedasticity
- Autocorrelation

Extended Regression

- Generalized linear models
- Polynomial regression
- Logistic regression
- Exponential regression
- Regularization
- Lasso regression
- Ridge regression
- Basis functions
- Step functions

Classification

- Binary classification
- Multi-class classification
- Class probability
- One versus one
- One versus the rest

Clustering

- Hard clustering
- Soft clustering
- Centroid models
- Connectivity models
- K-means clustering
- Hierarchical clustering
- Dendrograms
- Expectation maximization
- Silhouette coefficient

Neural Networks

- Perceptron
- Backpropagation
- Activation function
- Multi-layer perceptron
- Feed-forward networks
- Radial basis networks
- Extreme Learning Machines
- Recurrent Neural Network
- Convolutional Neural Network
- Autoencoders
- LSTM's
- Markov Chain
- Generative Adversarial Network

Timeseries analysis

- Autoregressive models (AR)
- Moving average models (MA)
- ARMA/ ARIMA
- Timeseries decomposition
- GARCH Models
- ARMAX Models
- Lagged regression models
- Stationarity
- Autocorrelation
- Partial Autocorrelation
- Cross-correlation

Natural Language Processing

- Entity recognition
- Sentiment analysis
- Bag-of-words
- Topic modelling
- n-grams

Reinforcement Learning

- Reward function
- Action set
- Policy function
- State space

Non-relational Databases

- NoSQL
- Document databases
- Key-value databases
- Wide-column stores
- Graph databases

Big Data/ Data Systems

- Data lake
- Data warehouse
- Distributed file systems
- Map Reduce
- Batch processing
- Real-time processing
- Stream processing

Programming

- Data types
- Arrays
- Operators
- Logic Statements
- Loops
- Functions, Definitions

Version Control

- Git
- Branch
- Clone
- Commit
- Merge
- Push
- Pull

Visualizations

- Line plots, Scatter plots, Pie charts
- Distributions and Box-whisker plots
- Tree diagrams
- Network diagrams
- Word clouds
- Flow maps
- Heat maps

Dimensionality Reduction

- Variance/ correlation filters
- Principle Component Analysis (PCA)
- Independent Component Analysis (ICA)
- Explained variance
- Component/ factor analysis
- Non-linear component analysis

Machine Learning Concepts/ Models

- Supervised vs. Unsupervised Learning
- Semi-Supervised Learning
- Deep Learning
- Reinforcement Learning
- Lazy vs. Greedy Learners
- Regression vs. Classification
- Ensemble Learning
- Transfer Learning
- Naïve Bayes
- Tree-based Methods
- Boosting Methods
- Support Vector Machines
- Neural Networks
- k-Nearest Neighbors

Tree-based Methods

- Branch
- Leaf
- Node
- Depth
- Decision Trees
- Extra Trees
- Boosted Trees
- Random Forest
- Entropy, Information gain

Support Vector Machines

- Kernel functions
- Margin classifier
- Hyperplanes
- Polynomial kernel
- Radial kernel
- Support vector classifier

- Actor critics
- Q-learning, Q-value, Q-matrix

Ensemble Methods

- Bagging
- Boosting
- AdaBoost
- Stacking

Model Selection and Optimization

- Auto Machine Learning
- Model pipelines
- Forward selection
- Backward selection
- Hyperparameter optimization
- Cross-validation
- K-fold cross validation
- Leave one out cross validation
- Grid search
- Randomized search
- Bayesian optimization
- Gradient-based optimization
- Evolutionary optimization
- Racing algorithms

Model Evaluation

- Train/ test/ validation split
- Bias, Variance
- Error metrics
- Residual analysis
- R-squared statistic
- Adjusted R-squared
- Akaike Information Criterion
- Bayesian Information Criterion
- Variance Influence Factors
- MAE, MSE, RMSE, MAPE, MPE
- Confusion matrix
- TP, FP, FN, TN
- Accuracy, Precision, Recall
- F1 score
- Receiver Operating (ROC) curve
- Area Under Curve (AUC)