

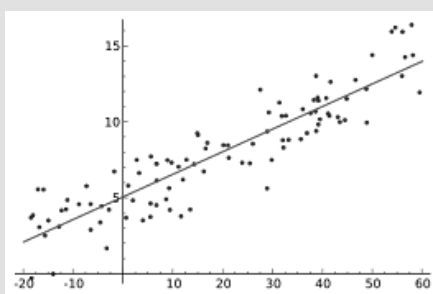
### Three Common Types of Problem

**Regression** To find the relationship between a dependent variable and many independent variables

**Classification** To classify an observation to one of the several known categories

**Clustering** To group a set of objects into several unknown clusters

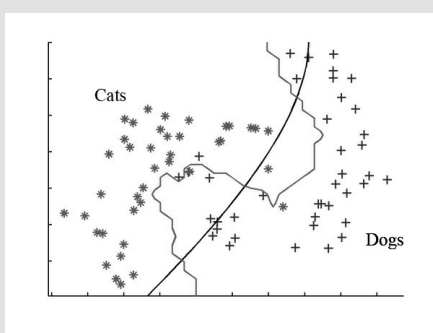
### Regression



#### 🔗 Model evaluation methods:

$R^2$ , Adjusted  $R^2$ , MAE (*mean absolute error*), MSE (*mean square error*), RMSE (*root mean square error*), AIC (*Akaike information criterion*), BIC (*Bayesian information criterion*), Residual analysis, Goodness-of-fit test, Cross validation

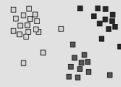
### Classification



#### 🔗 Model evaluation methods

Accuracy, Confusion matrix, Sensitivity and specificity, ROC (*receiver operating characteristic*), AUC (*area under the curve*), Cross validation

### Clustering



#### 🔗 Model evaluation methods

Models can be externally evaluated using data that are not used for clustering but with known class labels

#### 🔍 General steps to build a model

1. Collecting the data.
2. Preparing the data and fixing issues such as missing values and outliers.
3. Use exploratory analysis to help study the content of your data and select a proper algorithm that suits your need.
4. Training a model using the algorithm you just selected. Start with a simple model that only uses the most important variables/features.
5. Check model performance using the evaluation methods.
6. If the model is not satisfactory, choose another algorithm or introduce different variables into the existing model.

#### 🔍 Popular tools of implementation

R ML libraries including *stats*, *glmnet*, *caret*

**Python** popular packages for ML including *scikit-learn*, *statsmodels*

**Alteryx Designer** 'drag-n-drop' and requires minimum coding

**Microsoft Azure Machine Learning Studio** 'drag-n-drop' and requires minimum coding

#### 🔗 Linear Regression

**Learning style** Supervised

**Problem** Regression

**Use case** Revenue prediction

Widely used for predicting numeric values (or quantities). It trains and predicts fast, but can be prone to overfitting so proper feature selection is often needed.

#### 🔗 Logistic regression

**Learning style** Supervised

**Problem** Classification

**Use case** Customer churn prediction

A generalized linear model with dependent variable being binary (0-1). Mostly used to predict whether an event is going to occur based on the dependent variables.

#### 🔗 Decision Tree

**Learning style** Supervised

**Problem** Classification/Regression

**Use case** Targeted advertising

It requires little data preparation and can handle both numeric and categorical data. Easy to interpret and visualize but susceptible to overfitting.

#### 🔗 Random Forest

**Learning style** Supervised

**Problem** Classification/Regression

**Use case** Credit card fraud detection

An ensemble method that combines many decision trees together. It has all pros that a basic decision tree has, can handle many features and usually has high accuracy.

#### 🔗 K-means

**Learning style** Unsupervised

**Problem** Clustering

**Use case** Customer segmentation

This method groups objects into  $k$  clusters. The goal is to have the objects in one cluster more similar to each other than to any object in other clusters. When  $k$  is not pre-determined, many methods can be used to find a good value of  $k$ , such as the elbow method and silhouette method.

### ✦ Naïve Bayes

<b>Learning style</b>	Supervised
<b>Problem</b>	Classification
<b>Use case</b>	Email spam filtering

A conditional probability model that assumes all features are conditionally independent on each other. Trains and predicts fast but the precision is low for small datasets and can suffer from 'zero-frequency' problem.

### ✦ K-nearest Neighbors (KNN)

<b>Learning style</b>	Supervised
<b>Problem</b>	Classification
<b>Use case</b>	Bank credit risk analysis

A lazy learning algorithm that doesn't require much in training, but can be slow in prediction if you have a large data set.

### ✦ Support Vector Machine (SVM)

<b>Learning style</b>	Supervised
<b>Problem</b>	Classification/Regression
<b>Use case</b>	Text classification

It uses some kernel function to map data points to a higher dimensional space and find a hyperplane to divide these points in that space. Ideal for very large data set with high dimensions, or if you know the decision boundary is not linear.

### References

Basics of machine learning  
<https://www.analyticsvidhya.com/blog/2015/06/machine-learning-basics/>

A practical guide to exploratory analysis  
<https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/>

A cheat sheet of the libraries/modules of each algorithm in Python/R  
<http://www.dummies.com/programming/big-data/data-science/machine-learning-dummies-cheat-sheet/>

A cheat sheet for using Microsoft Azure Machine Learning Studio  
<https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-algorithm-cheat-sheet>

Tool sheet of Alteryx Designer  
[http://www.alteryx.com/sites/default/files/alteryx-designer-tools-sheet\\_0.pdf](http://www.alteryx.com/sites/default/files/alteryx-designer-tools-sheet_0.pdf)



By **lulu\_0012**  
[cheatography.com/lulu-0012/](https://cheatography.com/lulu-0012/)

Published 27th March, 2017.  
Last updated 30th April, 2017.  
Page 2 of 2.

Sponsored by **Readable.com**  
Measure your website readability!  
<https://readable.com>