# Logistic Regression

# Logistic Regression Assumptions

1. The logistic regression assumes that there is minimal or no multicollinearity among the independent variables.
2. The Logistic regression assumes that the independent variables are linearly related to the log of odds.
3. The logistic regression usually requires a large sample size to predict properly.
4. The Logistic regression which has two classes assumes that the dependent variable is binary and ordered logistic regression requires the dependent variable to be ordered, for example Too Little, About Right, Too Much.
5. The Logistic regression assumes the observations to be independent of each other.

# Naïve Bayes Classifier

Multinomial Naive Bayes:

This is mostly used for document classification problem, i.e whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.

The fundamental Naive Bayes assumption is that each feature makes an:

* independent
* equal

contribution to the outcome.

**The zero-frequency problem**

Add 1 to the count for every attribute value-class combination (Laplace estimator) when an attribute value (Outlook=Overcast) doesn’t occur with every class value (Play Golf=no).

**Laplace Correction**

Tips to improve the model

1. Try transforming the variables using transformations like BoxCox or YeoJohnson to make the features near Normal.
2. Try applying Laplace correction to handle records with zeros values in X variables.
3. Check for correlated features and try removing the highly correlated ones. Naive Bayes is based on the assumption that the features are independent.
4. Feature engineering. Combining features (a product) to form new ones that makes intuitive sense might help.
5. Try providing more realistic prior probabilities to the algorithm based on knowledge from business, instead of letting the algo calculate the priors based on the training sample.

For this case, ensemble methods like bagging, boosting will help a lot by reducing the variance.

# Assumptions in KNN

Before using KNN, let us revisit some of the assumptions in KNN.

KNN assumes that the data is in a feature space. More exactly, the data points are in a metric space. The data can be scalars or possibly even multidimensional vectors. Since the points are in feature space, they have a notion of distance – This need not necessarily be Euclidean distance although it is the one commonly used.

Each of the training data consists of a set of vectors and class label associated with each vector. In the simplest case , it will be either + or – (for positive or negative classes). But KNN , can work equally well with arbitrary number of classes.

We are also given a single number "k" . This number decides how many neighbors (where neighbors is defined based on the distance metric) influence the classification. This is usually a odd number if the number of classes is 2. If k=1 , then the algorithm is simply called the nearest neighbor algorithm.

# Multiple linear regression analysis makes several key assumptions:

There must be a linear relationship between the outcome variable and the independent variables. Scatterplots can show whether there is a linear or curvilinear relationship.

Multivariate Normality–Multiple regression assumes that the residuals are normally distributed.

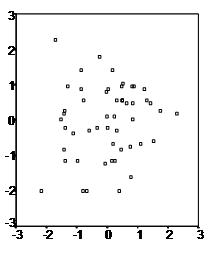
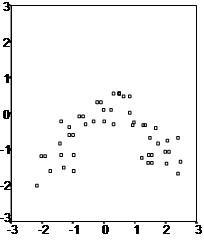
No Multicollinearity—Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values.

Homoscedasticity–This assumption states that the variance of error terms are similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.

Multiple linear regression requires at least two independent variables, which can be nominal, ordinal, or interval/ratio level variables. A rule of thumb for the sample size is that regression analysis requires at least 20 cases per independent variable in the analysis. Learn more about sample size here.

# **Multiple Linear Regression Assumptions**

First, multiple linear regression requires the relationship between the independent and dependent variables to be linear.  The linearity assumption can best be tested with scatterplots. The following two examples depict a curvilinear relationship (left) and a linear relationship (right).



Second, the multiple linear regression analysis requires that the errors between observed and predicted values (i.e., the residuals of the regression) should be normally distributed. This assumption may be checked by looking at a histogram or a Q-Q-Plot.  Normality can also be checked with a goodness of fit test (e.g., the Kolmogorov-Smirnov test), though this test must be conducted on the residuals themselves.

Third, multiple linear regression assumes that there is no multicollinearity in the data.  Multicollinearity occurs when the independent variables are too highly correlated with each other.

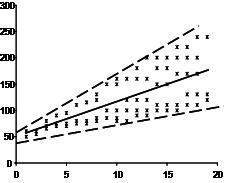
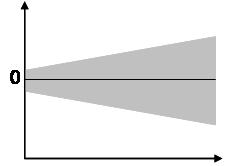
Multicollinearity may be checked multiple ways:

1) Correlation matrix – When computing a matrix of Pearson’s bivariate correlations among all independent variables, the magnitude of the correlation coefficients should be less than .80.

2) Variance Inflation Factor (VIF) – The VIFs of the linear regression indicate the degree that the variances in the regression estimates are increased due to multicollinearity. VIF values higher than 10 indicate that multicollinearity is a problem.

If multicollinearity is found in the data, one possible solution is to center the data.  To center the data, subtract the mean score from each observation for each independent variable. However, the simplest solution is to identify the variables causing multicollinearity issues (i.e., through correlations or VIF values) and removing those variables from the regression.

The last assumption of multiple linear regression is homoscedasticity.  A scatterplot of residuals versus predicted values is good way to check for homoscedasticity.  There should be no clear pattern in the distribution; if there is a cone-shaped pattern (as shown below), the data is heteroscedastic.



If the data are heteroscedastic, a non-linear data transformation or addition of a quadratic term might fix the problem.