All the Annoying Assumptions

In this article I have tried to collect and map the assumptions of common data science algorithms.

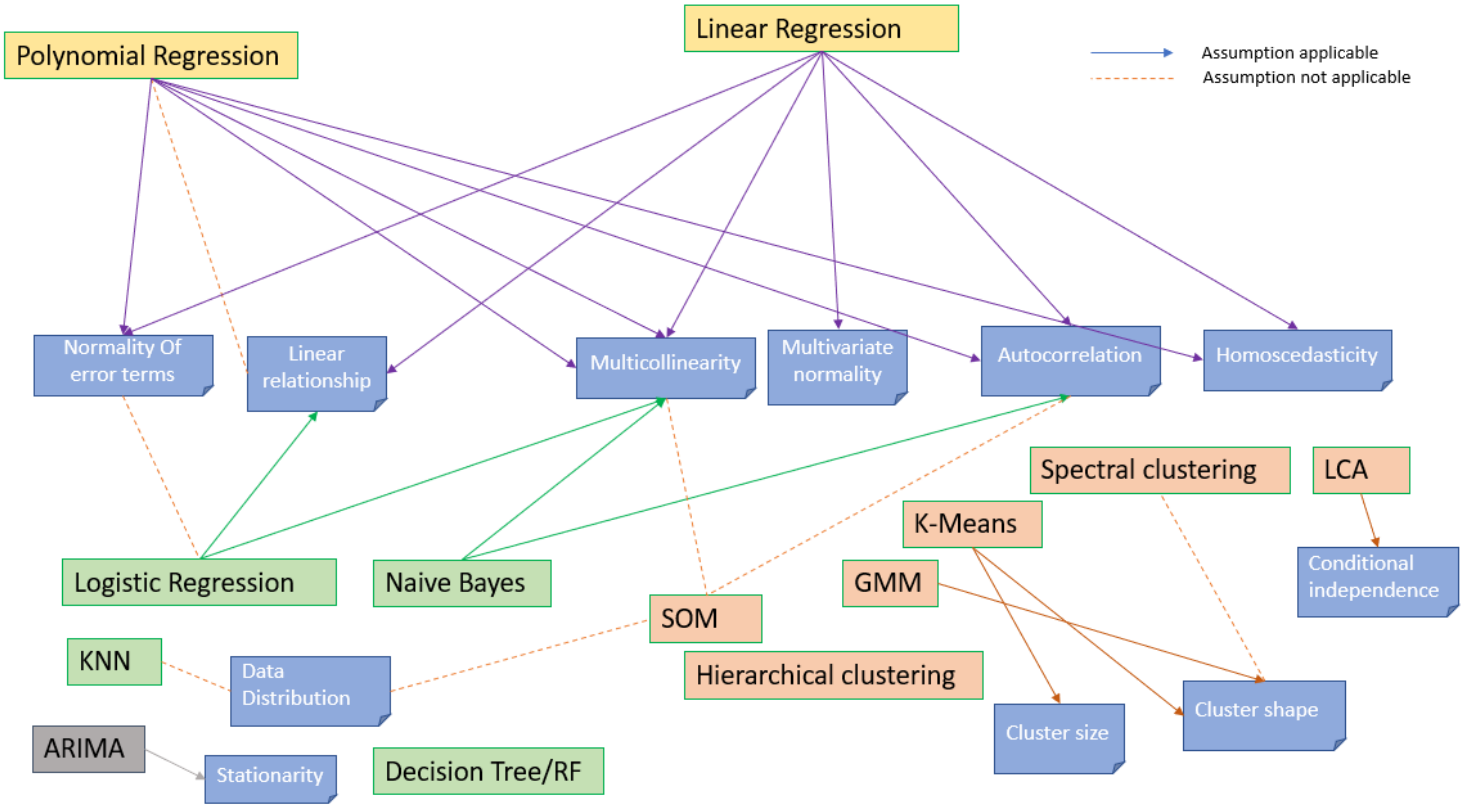
Dip Ranjan Chatterjee Follow

Aug 26, 2019 · 6 min read ★

Assumptions are annoying little things which need to be adhered if you are to successfully implement any model. Most model, however basic, has got a set of

assumptions and in this article I have tried to map them to a pattern for easy understanding.

Let me start with a picture of an assumption map which I have tried build. From there on we can easily navigate the assumptions pertaining to each of the sections.



Assumption Map

Let's start..

Regression

Both linear and polynomial regression share a common set of assumptions which need to satisfied if their implementation is to be of any good.

1. **Linear relationship :** Linear regression needs the relationship between **the independent and dependent variables to be linear.** This linearity assumption **can best be tested with scatter plots.** However as quite obvious the **linearity assumption is not valid for polynomial regression.**

2. **Multivariate normality :** Regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a **histogram or a Q-Q-Plot. Normality can be checked with a goodness of fit test, e.g., the Kolmogorov-Smirnov test. When the data is not normally distributed a non-linear transformation (e.g., log-transformation) might fix this issue.**

3. **Multicollinearity :** Regression assumes that there is little or no multicollinearity in the data. **Multicollinearity occurs when the independent variables are too highly correlated with each other.** You can use a correlation matrix to check if the variables are correlated to each other. You can remove one set of the columns which are correlated while keeping the other.

4. **Autocorrelation : R**egression analysis requires that there is little or no autocorrelation in the data. **Autocorrelation occurs when the residuals are not independent from each other.** While a scatterplot allows you to check for autocorrelations, you can test the linear regression model for autocorrelation with the **Durbin-Watson test.**

5. **Homoscedasticity :** Homoscedasticity describes a situation in which the **error term is the same across all values of the independent variables.** The scatter plot is good way to check whether the data are homoscedastic (meaning the residuals are equal across the regression line). If homoscedasticity is present, a non-linear correction might fix the problem.

6. The error terms must be **normally distributed**.

> *Regression methods like Lasso, Ridge, Elastic Net are nothing but regularization techniques which tries to balance bias vs variance. Hence, these techniques does not have any separate assumptions of their own.*
>
> *SVR too is quite tolerant of the input data and does not have any separate assumptions of its own.*

Clustering

There are four types of clustering algorithms in widespread use: ***hierarchical clustering, k-means cluster analysis, latent class analysis,*** **and** ***self-organizing maps.***

1. **K-Means clustering** method considers two assumptions regarding the clusters **— first that the clusters are spherical and second that the clusters are of similar size.** Spherical assumption helps in separating the clusters when the algorithm works on the data and forms clusters. If this assumption is violated, the clusters formed may not be what one expects. On the other hand, assumption over the size of clusters helps in deciding the boundaries of the cluster.

2. Certain resemblance measures (e.g., Euclidean distance)assume that the variables are uncorrelated within clusters.

3. In **spectral clustering**, the data points are treated as nodes of a graph. Thus, clustering is treated as a graph partitioning problem. The nodes are then mapped to a low-dimensional space that can be easily segregated to form clusters. **An important point to note is that no assumption is made about the shape/form of the clusters.**

4. **Hierarchical clustering has got no separate assumption of its own.**

5. Similar to other clustering algorithm, the **GMM has some assumptions about the format and shape of the data.** If those criteria is not meet the performance might drop significantly.

6. LCA defines latent classes by the criterion of **"conditional independence."** This means that, within each latent class, each variable is statistically independent of every other variable. For example, within a latent class that corresponds to a distinct medical syndrome, the presence/absence of one symptom is viewed as unrelated to presence/absence of all others. For some applications, conditional independence may be an inappropriate assumption. For example, one may have two very similar items, such that responses on them are probably always associated. For this and certain related situations, extensions of the latent class model exist.

7. In comparison to multivariate techniques the **nonparametric SOM** procedures have a number of advantages. **First, they do not make assumptions regarding the distributions of variables and nor do they require independence among variables.** Second, they are easier to implement and are able to solve nonlinear problems of very high complexity. Last, they more effectively cope with noisy and missing data, very small dimensionality and samples of unlimited size.

Classification

The common types of classifiers are as follows: ***Logistic Regression, Naive Bayes Classifier, KNN, Random Forest***

1. **Logistic regression** is used for predicting dependent variables that take membership in one of a limited number of categories (treating the dependent variable in the binomial case as the outcome of a Bernoulli trial) rather than a continuous outcome. **Given this difference, the assumptions of linear regression are violated. In particular, the residuals cannot be normally distributed.**

2. There is a **linear relationship between the logit of the outcome and each predictor variables**. Recall that the logit function is `logit(p) = log(p/(1-p))`, where p is the probabilities of the outcome. This can be checked by visually inspecting the scatter plot between each predictor and the logit values.

3. There is no high intercorrelations (i.e. **multicollinearity**) among the predictors. This can be checked by visualizing the Cook's distance values.

4. In case of **Naive Bayes Classifier** an assumption is that the **predictors/features are independent**. Another assumption made here is that all the **predictors have an equal effect on the outcome.**

5. **KNN is an *non parametric lazy learning* algorithm.** That is a pretty concise statement. When you say a technique is non parametric , it means that it **does not make any assumptions on the underlying data distribution.**

6. In **Decision Tree** as we have no probabilistic model, but just binary split, we don't need to make **any assumption at all**. That was about Decision Tree, but it also applies for **Random Forest.** The difference is that for Random Forest we use Bootstrap Aggregation. It has no model underneath, and the only assumption that it relies is that **sampling is representative**. But this is usually a common assumption.

> *Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to **decrease variance** (bagging), **bias** (boosting), or **improve predictions** (stacking). Hence these too does not have any separate assumptions of their own.*

Time series analysis

ARIMA models work on the assumption of **stationarity** (i.e. they must have a constant variance and mean). If your model is non-stationary, you'll need to transform it before you can use ARIMA.

I will keep updating this article the more I learn about data science, till then cheers.

References:

1. https://www.researchgate.net/publication/263084866_An_Introduction_to_Self-Organizing_Maps

2. https://cdn1.sph.harvard.edu/wp-content/uploads/sites/59/2016/10/harvard-lecture-series-session-5_LCA.pdf

3. https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/

4. Wikipedia

5. https://www.statisticssolutions.com/assumptions-of-linear-regression/

6. Numerous Stack overflow questions

7. https://cdn1.sph.harvard.edu/wp-content/uploads/sites/59/2016/10/harvard-lecture-series-session-5_LCA.pdf

8. https://towardsdatascience.com/spectral-clustering-aba2640c0d5b

9. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

10. https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/

11. https://www.statisticssolutions.com/assumptions-of-logistic-regression/