**Discriminative models include:**

* Logistic regression
* Linear regression
* Support vector machines
* Random forest

**Generative models include:**

* Naïve Bayes
* Latent Dirichlet Allocation
* Generative Adversarial Network
  + There is a generative model and a discriminative model in GAN
  + Ultimate aim is to use generative model to improve the discriminative model
  + Generative model will produce fake cats and discriminative will try to identify whether is it fake a not, thus improving the discriminative model. It is iterative

**Why is Naïve Bayes naïve?**

The Naïve Bayes algorithm is called ‘naïve’ because it makes the assumption that the occurrence of a certain feature is independent of the occurrence of other features

**Multinominal vs multivariate**

Multinominal refers to y, while multivariate refers to X

For classification modeling

Numerical X variables:

* Use visualization plot to see the distribution of X variable against Y (target). If the distributions are similar, the X variable is not very good
* If the distribution is slightly different, you can use T-test to confirm

Categorical X variables:

* You can use chi-sq test to test each X category against Y
* Or you can plot a bar chart. X-axis contains all the X categories and Y-axis is the proportion of each X category resulting in 1. There will be a y = a line, where a is the baseline accuracy. If category A is 10% (or whatever threshold you set) above or below the baseline accuracy, that category is a good feature to predict your target. If category B does not pass the threshold line, you can drop that category. So you just need to do get dummies for the relevant categories that is above the baseline + threshold or baseline – threshold

Drop columns that have more than 30% null values

For NLP

LSI (Latent Semantic Indexing):

* Informational retrieval using Latent Semantic Analysis
* Distributional hypothesis: words that are similar in meaning occurs in similar pieces of text

Word2Vec

Doc2Vec

Global Vector (GLOVE)

Material

Stanford cs 224n – Deep learning for NLP

Book – lean analytics

[yapjiaqing@gmail.com](mailto:yapjiaqing@gmail.com)

**Interview questions (need to know)**

* Point estimate
* Interval estimate
* Frequentist vs Bayesian
* Parametric models vs non-parametric models
  + Parametric models can see equation because the parameters (variables) are finite
    - E.g Linear regression
  + Non-parametric models cannot see equation because the parameters are infinite
  + Non-parametric have lesser assumptions than parametric models
  + Adding more rows will increase more parameters
  + It is usually more accurate than parametric models but it is usually unexplainable
    - E.g. SVC – rbf

**If you have a lot of features, the X matrix is very sparse calculating the statistical significance, p-value is not good. It is not reliable**

**Distribution of likelihood does not have to be the same as the distribution of the prior**

**Pymc3 package is sequential programming. It is not parallel program. Cannot run parallel on multiple instances**