## Contents:

1. Stats 101:
   1. Probability basics
   2. Population distributions
      1. Generating few samples in excel
   3. Hypothesis testing
      1. H0,H1 🡪 how to define null and alternate hypothesis
      2. Comparing sample mean with population with population mean
      3. Comparing two samples means
      4. p-value
      5. z,t,f, chi-square statistic
      6. Type1, type 2 errors
2. Data ( in excel)
   1. Structured, Unstructured
   2. Data cleaning
   3. Imputations
   4. Descriptive statistics, ANOVA
3. Linear algebra overview
   1. Matrices and vectors
   2. Addition and scalar multiplication
   3. Matrix vector multiplication and properties
   4. Inverse, Transpose
4. Linear Regression with one variable ( mathematical explanations)
   1. Model representation
   2. Cost function 🡪 intuition
   3. Gradient descent 🡪 Intuition
   4. Gradient Descent for linear Regression
5. Linear Regression with Multi-variable
   1. Expand basic math for multi-variables
   2. Implementing Linear Regression in Excel ( generate random population)
6. Logistic Regression – expansion from Linear Regression
   1. Classification
   2. Hypothesis Representation
   3. Decision boundary
   4. Cost function & gradient descent
   5. Multi- Class Classification – one vs all
7. Machine Learning
   1. Supervised Learning
      1. Rule based classification
      2. Naïve Bayes
      3. Markov models – markov chain, hidden markov model
      4. Support vector machines
      5. *Heuristics ( not explaining)* 
         1. *Genetic algorithm*
         2. *Particle-swarm optimizations*
         3. *Support vector Machines (The parameters of the maximum-margin hyperplane are derived by solving the optimization. There exist several specialized algorithms for quickly solving the quadratic programming (QP) problem that arises from SVMs, mostly relying on heuristics for breaking the problem down into smaller, more manageable chunks)*
         4. *Artificial neural networks ( meta-heurisitc algorithms are used in optimizing NN)*
   2. Unsupervised learning ( explaining only what this is, not going into the sublevels and explaining each algorithm)
      1. PCA ( explain this)
      2. Clustering ( explain this)
         1. Hierarchical clustering
         2. K-means clustering
         3. Gaussian mixture models
      3. *Self- organizing maps*
      4. *Hidden Markov model is bit both supervised and unsupervised ml model*
   3. *Reinforcement Learning ( out of the scope of this doc)*
      1. *Monte-carlo methods*
      2. *Dynamic programming intro to solve sequential decision problems ( SDP)*
8. Tree based methods
   1. Decision trees – CHAID and CART
      1. Gini coefficient
      2. Model lift
   2. Random forest
   3. Boosting and Bagging
      1. Gradient boosting
      2. Ada boost
      3. XG boost
9. Forecasting
   1. Time series – Components
   2. Time series Techniques – applications ( intuition )
      1. Moving average
      2. Exponential smoothing
      3. Extrapolation
      4. Trend estimation
      5. *Auto-regression – ARIMA ( can skip this)*
10. Neural Network - Representation[[1]](#endnote-1)
    1. Non-linear hypotheses
    2. Neurons and the brain – intuition
    3. Perceptron
    4. NN Model Representation
    5. Intuition behind NN
    6. Multi-class Classification NN
11. Neural Networks – Learning ( math heavy)
    1. Cost function
    2. Backpropagation algorithm & Intuition
    3. Gradient checking
    4. Random initialization
    5. Activation functions
    6. Regularization
    7. Normalization
    8. Over-fitting – and how to reduce that
    9. Intro to tensors
12. ML vs Deep Learning
13. Major-architectures in Neural Networks
    1. Convolutional NN
    2. Recurrent NN
       1. LSTM
       2. Back propagation through time ( BPTT), truncated BPTT
       3. GRUs
    3. Recursive NN
    4. Encoder-decoder models ( seq2seq models)
    5. Attention Mechanism
       1. Transformer models ( self-attention)
    6. *Deep belief systems*
    7. *GANs, VAEs – just touch upon to give an intuition of what they do, not explaining in detail.*
14. *Recommender systems ( mostly wont explain, just to give an understanding)*
    1. *Content-based recommendations*
       1. *Associative rule mining*
          1. *Support, confidence, lift*
    2. *Collaborative filtering models*
       1. *Similarity/ distance measures*
          1. *Pearson correlation*
          2. *Cosine similarity*
          3. *Minkowski distance measures*
          4. *Jaccard co-efficient*
       2. *Page- ranking algorithm*
    3. *Knowledge-based recommendations*
15. General modelling fair practices
    1. Confusion matrices
    2. Tp, fp, precision , recall, f1 score, roc, auc ( these maybe covered in hypothesis testing)
    3. Ks stat
    4. Cross-validation
16. Technologies and setup
    1. Things need to run the codes mentioned in this tut, and how to setup them!
    2. Hadoop file system
       1. Map, reduce examples for code
    3. Spark
    4. Python basics ( If needed)
       1. Popular packages list and their descriptions
          1. Numpy
          2. Pandas
          3. Sklearn
          4. Keras
             1. Backend libraries – Tensorflow , Theano
       2. Well known packages for different domains( because we are trying to continue to maintain this repository)
          1. NLP – nltk, spacy, Glove, word2vec, BERT, Elmo, Latent Dirichlet allocation
          2. Image processing – Opencv, OpenGl

## Division of work

|  |  |
| --- | --- |
| 1. Stats 101 2. Supervised ml models |  |
| 1. Data 2. Unsupervised ml models |  |
| 1. Tree based methods 2. Forecasting |  |
| 1. Regression 2. Neural networks |  |
| 1. NN architectures |  |

1. [↑](#endnote-ref-1)