A Few Useful Things to Know About Machine Learning

* A generalized notion of machine learning:

1. (Develop and Improve) Representation: The model that will relate the input to the output. E.g., K nearest neighbours, logistic regression
2. (Try) Evaluation: The objective function that distinguishes good classifiers from bad ones. E.g., accuracy, squared error
3. (Measure) Optimization: A function that finds the optimal values of the loss function by estimating the parameters and selecting the values that give the least error. E.g., gradient descent, greedy search.

* The choice of one component depends largely on the other components.
* Generalization: Set aside some data from the beginning to test the classifier. Data diversity is a good way to ensure better generalization of a model.
* What are the crucial factors in making a good model apart from having a big corpus of data?
* How increasing the dimensionality of independent feature vector space affects the amount of data needed for generating better predictions from a model.
* Two different ways of taking the input knowledge, one being Induction and the other being Deduction.
* How to choose a model based on inference gained by analysing the model and its intrinsic details.
* Discussed what Bias, Variance and Overfitting mean in the context of Machine Learning.
* How one can avoid the problem of overfitting by using Cross-Validation.
* How dimensionality of training set affect generalization
* How non uniformity gives advantage to learner
* Some common misperception about number of features to use
* How theoretical guarantees fails in real world application
* Discussed feature engineering:
  1. What matters: choosing of factors that significantly contribute to a certain prediction.
  2. Challenges faced during feature engineering are:
     1. It is time consuming
     2. It is hard to find patterns in a huge dataset
  3. It is a display of intuition, creativity and “Black Art”
  4. “Holy Grail” of feature engineering is to make it automated.
  5. Combination of irrelevant features may seem relevant
  6. Using excessive features can lead to overfitting.
* Discussed why more data is better than complex algorithm:
  + 1. Optimization:  comparison between collecting more data and improving the algorithm
    2. Types of learners: fixed vs parameterized
    3. Simple algorithm with more data has higher payoff: comparison between complex learner’s vs complex classifiers.
* Discussed about ensembling and different types like boosting, bagging and stacking.
* Discussed about simple model’s vs complex models and which model should be preferred or not-preferred.
* Just because a model is capable of learning a function, does not mean practically it can learn to approximate it. (Because we might have less data than we actually need).
* Discussed about correlation does not imply causation with an example.

Contributors: -

1. Akanksha (https://github.com/akanksha317)
2. Varun Bhardwaj (https://github.com/VarunBhardwaj03)
3. Varun Bhardwaj (https://github.com/VARUN-BHARWAJ)
4. Suhayl Mahek (https://github.com/suhaylmahek)
5. Viswanathan Vivek (https://github.com/vviveks)