DSS
@Berkeley

Intro to Machine Learning

Workshop Series #10

Description

The Data Science Society is proud to present its ninth workshop of the year in an Introduction to Machine Learning. In this workshop, we will be covering the entire life cycle of how data scientists approach some of industry’s difficult questions by giving a high-level explanation into each step of the process. Whether you are an aspiring data scientist, or just plain curious as to how the magic happens, this workshop is for you!

Topics of this workshop include the following:

- Data Cleaning
- Deciding on a model based on data and circumstance
- Regressions and Classification
- Feature Engineering
- Model Evaluation

Setting up Anaconda

Mac: Python 3.6

https://www.anaconda.com/download/#macos

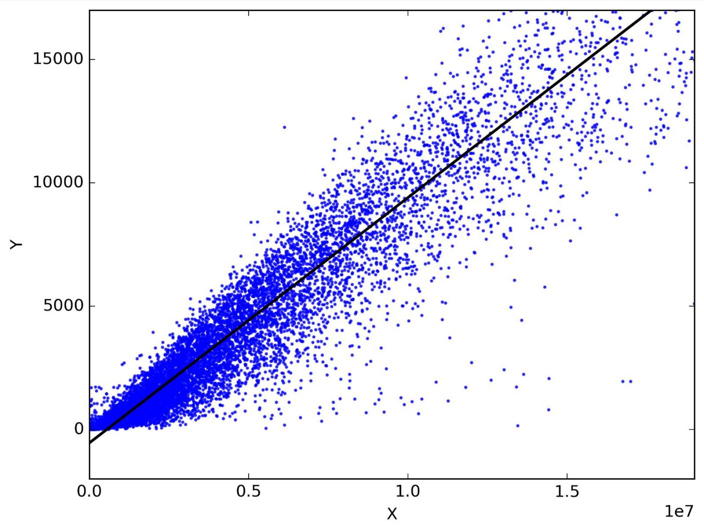
Windows: https://medium.com/@GalarnykMichael/install-python-on-windows-anaconda-c63c7c3d1444

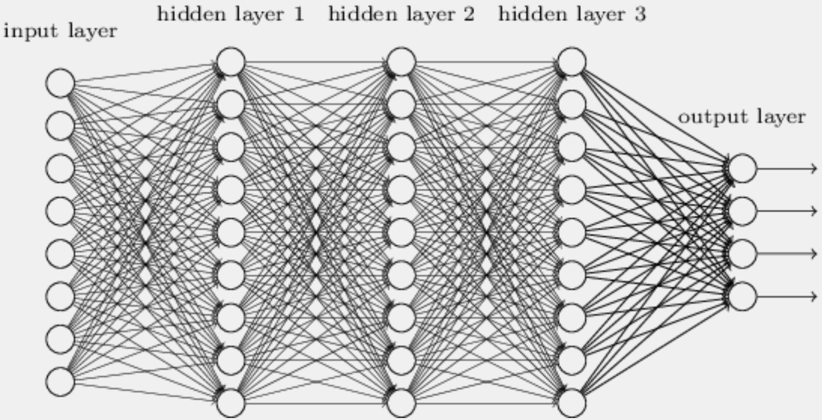
Agenda

- **[7:10 pm]** - Introduction, Setup
- **[7:15 pm]** - Linear Regression Example
- **[8:00 pm]** - Break
- **[8:05 pm]** - Case Study: Kickstarter Data
- **[8:50 pm]** - Q&A
- **[8:50-9:00 pm]** - Final Thoughts + Feedback

What is Machine Learning?

- Machine Learning is a field of Computer Science which uses algorithms and statistical techniques to “learn” actions or behavior.
- Origins in early 1950’s





Understanding Your Problem

- What questions are you trying to answer?
- What type of data do you have access to?
- How will your models be used?

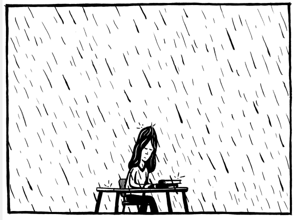
Supervised vs. Unsupervised Learning

- **Supervised Learning:** finding patterns in data to generate predictions based on a set of labeled data i.e. training set.
  - Useful to explain some dependent variables based on a set of feature variables
  - Ex: linear regression, decision trees

- **Unsupervised Learning:** inferring structure from a set of unlabeled data.
  - Useful when our data is unlabeled or we want to find general structures within our dataset.
  - Ex: clustering, anomaly detection

Regression vs. Classification

- **Regression:** examine or prediction a continuous relationship based on a set of independent variables.
  - Ex: forecasting the amount of rain tomorrow



- **Classification:** divide observations into a discrete set of categories based on a set of independent variables.
  - Ex: predicting if tumor is benign or not, determining what color the dress is

Choosing the Right Model

Some important factors to consider:

- Accuracy/Predictive Power:
- Interpretability:
- Stability:
- Ease of use:

Model Training: Test, Train, Validation

- Training Set: the dataset which will be used to to build or train your model
- Validation Set: the dataset which is used to tune your model during training
- Test Set: the dataset used to evaluate how well your model performs on “real world” data. Do not use until very end!!

Model Evaluation

How do we determine how well our model is doing?

Loss functions: a quantitative measure of model performance compared to the optimal results.

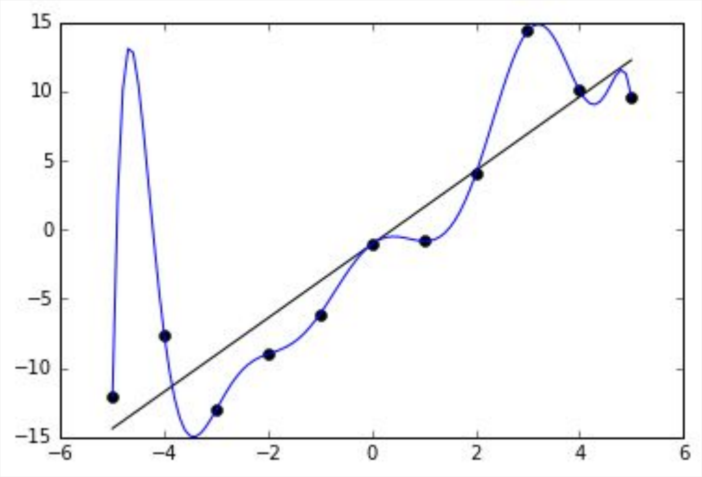
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

Examples:

- Mean squared error: measure of difference between predicted and actual values
- Regret: difference between sum of rewards of optimal strategy and applied strategy
- Accuracy: ratio of observations correctly classified under correct category

Overfitting

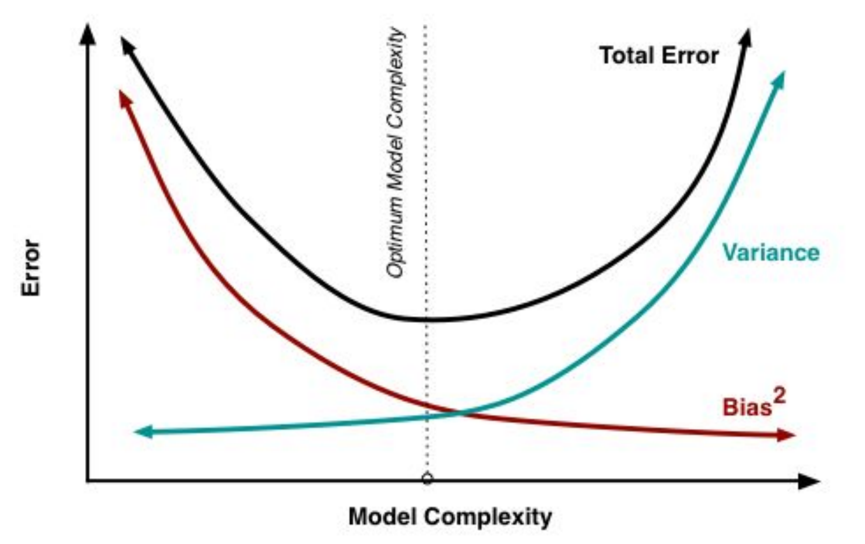
- Higher fit does always mean better “real world” performance!!
- The goal is not always to as closely fit given data as possible



Bias-Variance Tradeoff

In a perfect world we’d prefer a model that can generate predictions with zero errors for any given dataset as input.

In reality we must choose between how well the model can fit a given set of data i.e. **bias** and how much a given model changes for new datasets **variance**


Optimum Model Complexity
Total Error
Variance
Bias^2
Error
Model Complexity