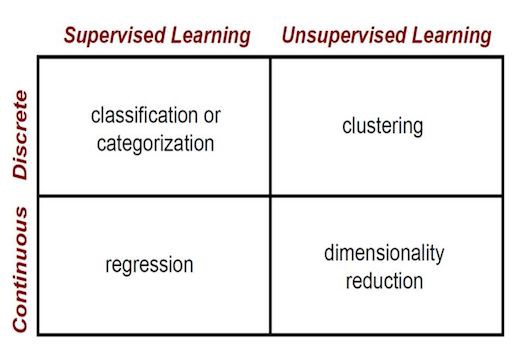
Lesson 13 – Introduction to Machine Learning

**Questions for Mentor:**

**Intro to Machine Learning for Beginners:**

* Used for
  + Prediction
  + Image recognition
  + Speech recognition
  + Medical diagnoses
  + Financial industry and trading
* Machine learning (ML) is a category of an algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed
* Can be classified into 3 types of algorithms
  + Supervised learning
    - With labeled data
    - Has ‘ground truth’ associated with it
    - Goal is to approximate the mapping function so well that when you have new input data you can predict the output variables
    - Examples: classification or regression
  + Unsupervised learning
    - Unlabeled and uncategorized data
    - Does not have ‘ground truth’ associated
    - Goal is to infer natural structure present within a set of data points
    - Examples: clustering, association (people who buy x also buy y)
  + Reinforcement learning
    - Learning using rewards and punishments
* Supervised vs Unsupervised learning
  + supervised learning common algorithms:
    - logistic regression
    - naive bayes
    - support vector machines
    - artificial neural networks
    - random forests
  + don’t overfit – then you’re just fitting to get training data as output, not to learn of the actual trend or structure
  + Unsupervised learning- we wish to learn the inherent structure of our data without using explicitly provided labels
  + Unsupervised learning is very useful in exploratory data analysis because it can automatically identify structure in data
* 
* Batch learning
  + Algorithm performs batch learning if the system can’t learn incrementally and must be trained using all available data
  + System is trained then launched
  + Will need to be updated with new data
* Online learning
  + Trains system by breaking data up into groups and feeding the system those groups over a longer period of time
  + Model learns in close to real time and updates accordingly
  + Must set learning rate appropriately so it can adapt to new data but without being at the expense of learning done with previous data
* K-Nearest Neighbor
  + Predictions made for new data point by searching through entire training set for K most similar instances (neighbors) and summarizing output variable for K instances
* Learning Vector Quantization (LVQ)
  + Allows you to choose training instances to hang onto and learns what instances look like so you don’t have to hang onto whole training set
* Self Organizing Map (SOM)
  + Similar to clustering similar values
* Locally weighted learning (LWL)
  + Local model created for each point of interest based on neighboring data of the query point
* Machine Learning 101
  + Learn about relationship of underlying data
  + Most times, we learn an approximation of underlying relationship
  + The workflow to build a machine learning model is centralized around the data
  + Bias is a learner’s tendency to consistently learn the same wrong thing. Variance is the tendency to learn random things unrelated to the real signal
  + Bias vs Variance coordinate
* A close up of a logo

  Description automatically generated
* A close up of text on a white background

  Description automatically generated
* Data Scientists and Machine Learning
  + Feature engineering
    - Preparing proper input dataset, compatible with ML algo requirements
    - Improving performance of ML models
  + Techniques for feature engineering
    - Web page saved in springboard folder
* Overfitting vs Underfitting
  + Validation
    - Validation sets used in addition to training and testing set
    - K-fold cross validation
* Rules of machine learning
  + Don’t be afraid to launch a product without ML
  + Make metric design and implementation a priority
  + Choose machine learning over a complicated heuristic
  + Keep first model simple and get infrastructure right
  + Detect problems before exporting models
  + Don’t overthink which objective you choose to directly optimize first
  + Choose a simple, observable and attributable metric for first model
  + Log features at serving time
  + Don’t waste time on new features if online objectives have become issue
  + Launch decisions are a proxy for long term goals