**Machine Learning**

**Lesson 1:**

* **What is machine learning?**
  + Inductive learning: Process is to provide input data with outputs completed previously and have the machine figure out the best process to complete the desired task
  + Machine learning focuses on most likely future predictions
  + Statistics models are more concerned with relationships between variables (inference)
  + Data mining – discovering previously unknown properties about your data
  + Data science – ML, data mining, optimization
  + Business analytics – predictive, descriptive, prescriptive analytics
* **Supervised Learning**
  + Process of training a predictive model focused on assigning labels to unlabeled data
  + Trained model – inputs and outputs applied with goal of finding pattern for future use
  + Model is tested by providing just inputs and compare how accurate outputs are
* **Unsupervised Learning**
  + Process of building descriptive models which are concerned with summarizing or grouping unlabeled data in new and interesting ways
  + **THERE ARE NO OUTPUTS TO COMPARE TO IN UNSUPERVISED LEARNING**
* **Collecting Data**
  + Ground truth – historical data without output to compare to
  + Exploration – describing, visualization, analyze for better understanding
  + Categorical 🡪 explanatory variable (x-axis), Continuous 🡪 response variable (y-axis)
* **Sampling**
  + Smoothing – minimizing noise
  + Bootstrapping – sampling with replacement
  + Systematic sample – with n = 20, you want a sample of 5: 20/5 = 4 therefore pick every 4th sample subject
  + Stratified is keeping population to sample of proportions
  + Holdout method – testing model w/data it has not seen before
* Misc.: regression for continuous variables, classification for categorical variables

**Lesson 2: Association**

* Unsupervised learning – process of building a descriptive model
* Descriptive model – concerned with summarizing or grouping unlabeled data in a new and interesting ways
* Item set – distinct set of items that are purchased as part of a transaction
  + Can be single or multiple items
* Every item set with 2 or more length has a left-hand sight and a right-hand side
  + EX – {beer, milk} 🡪 {diapers}
  + Left side: condition that needs to be met to trigger rule – precedent
  + Right side: expected result for meeting the condition – antecedent
* **Association Rules –** do not imply causality, they simply imply a strong co-occurrence relationship between the items
* Actionable – rues that provide clear, useful insight that can be acted upon
* Trivial – rules that provide insight to an already well-known rule
* Inexplicable – no clear course of action
* **Approaches** 
  + Brute force – approach to find which association rules are actionable
    - Looking at all possible rules with 2+ in length
    - For data set p distinct items, there exist 3p – 2p + 1 + 1 rules
    - For 20 items 320 – 220 + 1 + 1 = 3,484, 687,250 rules
* Generating frequent item set approach – change focus to rules based on regularly occurring item sets
* Frequency of an item set is measured by **support** or coverage
* **Support –** evaluates strength of item sets in association rules
  + Ex – number of transactions that contain said item set
  + Rules with low support may only occur by change
* **Confidence –** predictive power or accuracy of the rule
  + In other words – out of all transactions where both beer and milk were purchased, 67% of those transactions also had diapers
* **Lift –** increased likelihood of rule occurring relative to its typical rate of occurrence
  + In other words – customers who bought beer and milk are 16% less likely to also buy diapers
* **Apriori Algorithm** 
  + Reduces # of candidates by pruning itemset lattice
* Anti-monotone property of support – if an itemset is infrequent then its supersets are infrequent as well
* Strength: large transactional datasets and useful for discovering unexpected data patterns
* Weaknesses: not good for small datasets
  + Hard to separate insight from common sense
  + Easy to draw misleading patterns from random correlations
* **Unsupervised Clustering** 
  + Clustering – partitioning unlabeled subgroups based on similarity
  + High intra-class similarity – high level of similarity **within** the same group
  + Low inter-class similarity – low similarity **between** 2+ different groups
* Overlapping – clusters can overlap – each item can belong to 1+ group
* Exclusive cluster – no overlapping group
* Complete – every item is assigned to at least 1 cluster
* Partial - # of clusters are not known before hand
  + Usually outliers fall into not being assigned
* **K-Means Clustering**
  + Partitioned, exclusive and complete clustering approach that assigns all items to only one cluster. Differences within minimized differences b/t are maximized
* **K-Means Algorithm**

1. Pick K random cluster centers
2. Assign every item to its nearest cluster using a distance metric
3. Move each cluster to the mean of its assigned item
4. Repeat steps 2-3 until convergence is achieved

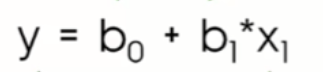
* Convergence – change if cluster assignment is less than threshold
* **Euclidean Distance** – straight line distance between the coordinates of two points in multidimensional space
  + Cluster centroid is average position of the items currently assigned to a cluster
* Random initialization trap – K-means is very sensitive to initial randomly chosen centers leading to possibly different results on same data if process is rerun
  + To overcome this K-means ++ is used which picks initial centroid as far away from other centers to minimize effect of randomness
  + Rule of thumb: k = n/2
* Performance Metrics
  + Elbow method – calculate within cluster sum of squares until it is minimized
  + WCSS us distances b/t items and centroid
  + Silhouette - measures how close item is matched within and how loosely it is with neighbors
    - Higher silhouette value, more likely in right cluster ranges from -1 to 1
  + Gap statistic – compares WCSS to random set

**Lesson 3: K-Nearest Neighbor**

* **Parametric model** – summarizes data with a fixed size of parameters, as/if data changes, number of parameters stays the same ex: linear regression
* Underfitting – when a model does not capture underlying patterns in the data
* **Nonparametric** **models –** does not make assumptions about form of the data and is free to take best fitting form, no matter the shape
* Overfitting – when a model captures the patterns too closely and limiting use ex: memorizing vs understanding
* **K-nearest Neighbor –** process of assigning a class or label to previous unlabeled data based on how similar it is to existing, already labeled data
* Lazy learners –do not generate model but instead perform just-in-time lookups in order to make predictions ex vlookups AKA instance-based learners, rote learners
  + Neighbors decided w/Euclidean distance
  + If multiple neighbors seen, majority wins for classification
  + K = how many neighbors to view
* The larger value for K, the less it is impacted by noise – too large and K might miss some important patterns
* The smaller value for K, allows to pick up on patterns more easily but too small leaves it susceptible to noise
* Rule of thumb starting point – root number of instances in training
  + Larger the data set, the less important K becomes
* Strengths: simple, effective, makes no assumptions, fast training
* Weakness: does not produce model, K selection arbitrary, slow classification, does not handle missing data, outliers, or categorical data well
* **Decision Trees –** goal is to model selected features to expected outcome
  + Classification tree – if model/data is discrete
  + Regression tree – if model/data is continuous
* Greedy learners – uses all data available on a first come, first serve basis
* Recursive partitioning – splitting the data into smaller and smaller subsets to maximize homogeneity of outcomes
* Continue until: all leaf notes are same class OR all features exhausted OR condition has been met
* **Measure of Impurity:** mathematical way to decide where to make best splits based on how similar outcomes are
* Entropy – quantification at the level of randomness or disorder, measure of purity
  + Lower entropy = lower impurity = higher homogeneity
* Gini – quantifies impurity/randomness by measuring how often a particular point will be labeled incorrectly if it were randomly labeled based on the distribution of labels in the partition
* Info gain – change in entropy after a split
  + Biased towards features with a high number of distinct values
* Gain ratio – normalized info gain to reduce bias by using intrinsic info – more values = higher intrinsic info = info. Gain/intrinsic info
* Pruning – process to minimize overfitting to generalize better
* Pre-pruning – setting a limit size during the partitioning process
* C5.0 handles pruning internally and utilizes post pruning
* CART – uses complexity parameter which is calculation of reduction of complexity by adding a new node
  + Pre-prune: smaller complexity parameter = larger tree
  + Post-prune: used more as error rate
* Dealing with imbalances: SMOTE
  + Over sample minority class and under sample majority class simultaneously
  + With oversampling – a synthetic sample is created using ML techniques
  + Only used by training data

**Lesson 4: Regression –** supervised learning, continuous

* Regression – statistical model whose goal is to model size and strength of numeric relationships in order to predict a target variable based on the values of previously observed exploratory variables
* **3 components:**
  + **Y** – a single numeric dependent variable
    - Value we want to predict, also known as the response variable
  + **X** – one or more numeric variables, also known as the predictor variables
  + **Coefficients** – describe relationships between predictor and response variables
    - Variables are unknown from the start
    - We use regression techniques to estimate the coefficients
* **Regression does not establish causation but rather explains the relationships b/t them**
* **Linear Regression**



* + If we have single predictor X and assume relationship is linear b/t X and Y
  + Adjusting coefficient β0 moves the line up or down
  + Adjusting coefficient β1 changes the slope
* Best regression line is the one with the fewest differences b/t actual and estimated points
  + These differences are known as residuals
* Ordinary least squares method – most common approach to finding optimal coefficients
* Linear regression assumes constant change in X leads to constant change in Y ex constant change in temperature leads to constant change in bike rentals
  + Assume if response variable has normal distribution and can vary indefinitely in either direction without fixed 0 value
  + Use exponential model if change in X leads to geometric change in Y
* **Poisson Regression**
  + Use if we assume relationship between predictor X and response Y is exponential or log-linear in relation
  + Useful when predictor is positive and response variable is a count ranging from 0 to infinity
    - Ex estimating # of people who will buy bikes
  + Uses maximum likelihood estimation to estimate coefficient
  + **Use when response variable is continuous** 
    - Can be used for categorical too (logistic)
* **Logistic Regression**
  + Uses non-linear function, logistic function
  + Probabilistic model used to model relationship b/t predictors and the odds of categorical responses
    - Odds they will purchase a bike, based temperature
    - also uses maximum likelihood estimation
  + normal cut-off threshold is 0.5 such as anything over .5 would be classified as yes and anything under will be classified of no but cutoff threshold can vary
  + Types:
    - Binomial – yes/no, true/false
    - Multinomial – 3+ response values ex red, green, north, east, south
    - Ordered – response values have order

**Generalized Linear Models**

* + - Stats technique to unify various regression techniques
    - Does this by using transformation function (link function) to represent relationship b/t predictors and response
* **Generalized Linear Models**
  + Stats technique to unify various regression techniques
  + Does this by using transformation function (link function) to represent relationship b/t predictors and response
* 3 components
  + Random component Y – describes distribution of response
  + Systematic component: linear combo of predictors
  + Link/transformation function – specifies relationship b/t random and systemic components
    - Important b/c it transforms random component that can then be explained as a linear combo of predictors
* Interpreted as a unit change in X changes the odds of P by a multiple of Beta 1
* Strengths:
  + easy to train and implement
  + do not require predictors to be scaled
  + has probability interpretation
  + can be regularized to avoid overfitting
  + able to handle normal data
* Weaknesses
  + Under perform w/multiple non-linear decision boundaries
  + Makes strong assumptions
  + Can miss complex relationships
  + Does not deal well with missing data or outliers
  + Best used for binary problems/predictors

**Lesson 5: Evaluating Performance**

* Estimating Future Performance
  + Using all data as training data or using same data to build and evaluate = resubstitution
  + **Resubstitution error –** measure of model’s performance against training data
    - Too optimistic b/c it is being tested on data that has already been encountered
    - Ex: given answers to a quiz then taking quiz with the same questions asked
* **Holdout method** – partitioning data into testing and training data, method we have been using
  + Partitions are independent = if a data observation is used in training set it cannot be used in testing
  + Test data must be representative of sample original data
  + Do not use test dataset to influence final data set model
    - Do not go back after evaluation and adjust training to perform well on specific test set. This causes overfitting and makes a model less versatile to perform wel on any other data set
  + Validation data set – 3rd partition to apply to tested data set again, typical split is 50-25-25
  + Weaknesses – does not work well with small data sets b/c might not be representative of sample when partitioned
    - Partitions might have too many simple/difficult patterns
* **Repeated Holdout Method –** uses resampling to maintain testing evaluation capabilities when partition shrinks training dataset too much
  + Folds represent the data that will be used to validate the model during the K iterations
  + Propensity – probability an instance belongs to a specified class
* **Cross-validation –** partitions data into K sets of training data. Within each K, the data is partitioned again into groups and each group is used as training and validation data and then accuracy is averaged
  + **Leave-one-out** – similar to K-fold but K is set to n # of instances
  + Benefits ensures greatest amount of data is used each time model is trained
  + Weakness is that it is high in computation cost, validation not stratified
* **Monte Carlo** – validation sets are randomly sampled during each iteration
  + draw back is the same, instances may never be used for validation/training
  + size of validation and training are independent = good
* **Bootstrapping –** resampling technique
  + Uses random sample with replacement for training set
  + Instances not randomly selected are used for validation
  + Probability of an item being selected in validation is 36.8% and 63.2% for training
  + Results in pessimistic outcome
  + Best used in small data set, or there is limited memory or speed
* Limits of predicate accuracy
  + **Accuracy paradox** – digging deeper into piece accuracy instead of overall could yield different results and class imbalance may lead to high accuracy by chance
  + **Kappa statistic** – adjusts accuracy by accounting for possibility of a correct prediction by change alone, range is 0-1
* Beyond Predictive Accuracy
  + Performance should be judged by utility, its use and purpose rather than raw accuracy
  + Precision/positive predicted value = proportion of positive predictions that are truly positive = trustworthy
  + Recall – correcting identifies positive measures
  + Having high recall usually means low precision
  + F-measure/F-score combines precision and recall into a single # using harmonic mean
  + Sensitivity/true positive rate – examples classified correctly
  + Specificity – true negative rate, negative classified correctly
  + ROC curve = commonly used visualization to examine trade-off between detection of true positive while avoiding false positive
  + AUC – treats entire ROC diagram as 2-D square and measure total area under ROC curve, ranges from 0.5 - 1
* Hyperparameter Tuning
  + Hyperparameters = parameters set before learning process begins
  + Process of selecting optimal hyperparameter values is known as parameter/hyperparameter tuning
  + Most common metrics are accuracy and Kappa for classification models and RMSE for regression
  + Cost sensitive measures are sensitivity, specificity, and ROC/AUC
  + **Grid search – automated parameter tuning to identify best combo of hyperparameters**
* Ensemble Methods – assumes we may not always find optimal set of hyperparameters for a single model
  + Instead of optimizing a single model, we use several complementary weak models to build effective system of models = ensemble
  + Homogeneous ensemble models – collection of single type, weak learners like KNN and decision trees
  + Heterogeneous – varied collection of weak learners
  + Allocation function – dictates how much of training data each model receives – can assign all of subset of data
  + Combination function governs how disagreements among models are reconciled dependent of type of problem solved
* Bagging – AKA bootstrap aggregating is a technique that builds several models by bootstrap sampling
  + Usually made of homogeneous learners built independently in parallel
  + Bagging attempts to reduce variability of single base estimator
  + Random forest – takes random subset of features from bootstrapped sample to run through variety of methods
  + Does well with noisy, missing or data with large number of features
    - Can do categorical or continuous data
    - Con: not easy to interpret, complex to compute
* Boosting – boosts performance of weak learners to construct a stronger classifier
  + Uses linear combination of homogeneous learners
  + Models are trained in sequence
  + Attempts to reduce bias error of estimator
  + Adaptive boosting – adds weights to incorrect classified data to best correct it
  + Easy to implement and tune
  + Tendency to overfit and slow to train, sensitive to noise and outliers
* Stacking – uses heterogeneous collection of base learners
  + Combination is non-deterministic meaning it does not follow pre-determined set of rules
  + Meta learning – learns from outputs of other models