Machine Learning Cheat Sheet

*Businesses increasingly make decisions using algorithms. Previously, algorithms had to be manually programmed by a computer scientist; however, machine learning allows algorithms to be **trained**, not programmed. The machine is free to choose its own parameters, with hyperparameters controlled by the data scientist.*

3 types of machine learning:

1 - Supervised Learning: useful when data is labeled, for example 'churned' or 'not churned.' The machine attempts to find the best model and/or coefficients that explain the variance in the labeled data. Input data is split into training and testing sets to gauge model accuracy.

2 - Unsupervised Learning: used when the data is unlabeled for exploratory analysis. Does not requre splitting data into testing and training sets.

3 - Reinforcement Learning: machine is rewarded or penalized with a numerical value. The machine tries to maximize its *cumulative* reward. Reinforcement learning is particularly well-suited to problems that include a long-term versus short-term reward trade-off.

SUPERVISED LEARNING

R **Simple Regression** - minimizes the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. The coefficient estimates for OLS rely on the independence of the features.

Examples: marketing attribution, forecasting sales/capacity/labor/etc.

C **Logistic Regression (Logit)** - models the probability of a certain class or event existing such as pass/fail, win/lose, or healthy/sick (e.g. whether or not a customer will buy, not how *many* they will buy.)

Examples: propensity to buy, propensity to churn.

R C **Decision Trees & Random Forest** - predicts class or value of a target variable based on logic rules developed by the algorithm. DTs are prone to overfitting, so often better to use Random Forest which samples randomly from the dataset to obtain different tree structures. For nonlinear problems. Use Bagging or Boosting to improve performance.

Examples: propensity to buy, propensity to churn.

- **Bagging** - Bootstrapping is the act of creating many decision trees from sub-samples of the dataset and averaging them. The application of Bootstrapping to a high-variance machine learning algorithm (typically decision trees) is known as Bagging (short for bootstrap aggregating).
- **Boosting** - typically used for Decision Trees, boosting combines weak "learners" into a single strong learner in an iterative fashion. Trees are built slowly as each iteration attempts to correct the errors of its predecessor, with future weak learners focusing more on the examples that previous weak learners misclassified. As iterations proceed, examples that are difficult to predict receive ever-increasing influence. Doesn't work well with linear models.

R C **Neural Networks** - good for nonlinear problems not solvable by simple/logistic regression. Uses multiple layers and nonlinear activation functions for classification, clustering, or prediction. Use Simulated Annealing or Dropout to avoid over-fitting.

Examples: propensity to churn, deep learning for self-driving cars (computer vision.)

C **Naive Bayes** - conditional probability family of algorithms used for classification problems. Simple, fast, & scalable; however, can be outperformed by other models (e.g. Random Forests.)

Examples: spam detection, recommendation systems (w/ collaborative filtering).

C **Support Vector Machines** - In SVM, a hyperplane is selected to best separate the points in the input variable space by their class, either class 0 or class 1. The SVM learning algorithm finds the coefficients that results in the best separation of the classes by the hyperplane.

Examples: text goes here.

R C **k-Nearest Neighbor** - primarily used for classification, k-NN calculates distance (or nearness) and then groups the input data with other similar cases. The output can be a class membership (e.g. "churner"), or a value (e.g. propensity to churn = 85%.)

Examples: loan default, propensity to buy, propensity to churn.

R = used for Regression, to discover the relationship between a dependent variable and one or more independent variables.

C = used for Classification.

UNSUPERVISED LEARNING

k-Means Clustering - a method of exploratory analysis, k-Mean Clustering splits the data into 'k' clusters centralized around 'k' means. The value for k is either preset by the data scientist or is the result of experimentation. Usually used as a component in other models, and can also be used for prediction.

Examples: customer segmentation, anomaly detection.

Principal Component Analysis (PCA) - similar to Factor Analysis in qualitative marketing research, PCA attempts to reduce the dimensionality of a data set while losing as little information as possible. PCA can reduce the number of explanatory variables, simplifying the model and lowering the chance of overfitting.

Examples: feature selection.

Neural Networks - can also be used in unsupervised learning. The machine attempts to build a compact representation of the input data. Output tends to be classifications or clustering.

Examples: customer segmentation, anomaly detection.

REINFORCEMENT LEARNING

Reinforcement learning algorithms can perceive and interpret their environment, and through iteration find the best global solution. RL has gained fast adoption recently for consumer-facing scenarios, such as Netflix's recommendation engine.

RL involves an *agent*, a set of *states*, and a set of *actions* for each state. By taking an action in each state, the agent receives a numerical reward, with its goal being to maximize its total reward. RL is well-suited to problems that include a long-term versus short-term reward payoff.

Monte Carlo simulations can be used as reinforcement learning for certain business strategy problems.

Examples: recommender systems, securities trading, robotics, strategy trade-offs