Cheat Sheet

- Estimators: MLE, MAP, Bayesian Estimation
- Model-Free Methods: Empirical Risk Minimization, Performance Measures
- Model Selection, Generalization, Overfitting, Regularization, Bias-Variance Tradeoff, Model Complexity
- Regression: Linear Regression, Regularized Linear Regression(Ridge, Lasso), Polynomial Regression Model
- Classification: Naive Bayes, Logistic Regression, Support Vector Machine, Boosting, Decision Tree, Conditional Independence in Naive Bayes Model, Generative vs Discriminative Classifiers, Linearly/Nonlinearly separable SVM, Role of Slack Variables in SVMs, Primal and Dual forms of constrained optimization problems, weak duality and strong duality
- Non-parametric Models: KNN, Kernal Regression, Kernel Trick
- Deep Learning and Neural Networks: Backpropagation, Overfitting, Regularization, Hyperparameter role and tuning for learning rate, momentum and regularizer, Stochastic and batch gradient descent
- Parametric vs Non-parametric Models
- K-means Clustering

Review Lecture: Model-based Approach: assume a model and find the parameters of this assumed model
MLE: maximize the probability of seeing the sample given some parameter theta $P(D|\theta)$ MLE is unbiased, asymptotic, consistent
MAP: choose θ whose probability given data is highest, ie maximize posterior distribution
Practice calculating MLEs and MLPs
Generalization and Risk Minimization True Risk + Empirical Risk
As the number of sample increases, the Empirical risk does down but true risk goes down and then up
Regularization term
Decision Tree:
Importance Measure: information gain based on entropy, how much uncertainty we remove when splitting on this certain attribute
Outlier removal with pruning cuz outlier is just a single branch sticking out and will be removed during pruning when testing the model
Neural Networks: learn a function f s.t. f(x) is approximately equal to y the true labels. $w_0 + w_1^T x$
Activation functions: linear, tanh, sigmoid, ReLU
Essentially creating complex decision boundary with multiply layers and activation functions
Training is essentially minimizing the loss function.
Compute gradient of loss function with respect to parameters. Gradient descent. Back propagation.
Batch gradient descent