**Neural networks**

Components of NN:

x – Input Vector

Wi – Weight Matrix

b - Bias

σ(sigma) – Elementwise Nonlinear Transformation

Structure of NN:

f(x) = σ(Wx + b) – x->FC->f(x)

a = σ1(W1x + b1)

f(x) = σ2(W2a + b2) – x->FC->a->FC->f(x)

How does an individual artificial neuron work: takes input, compute with weight and bias, generate output

Examples of activation functions:

Sigmoid (0-1) positive y

TanH (-1 – 1) negative and positive y

ReLU (0-0-infinite) y=0 until x=0, then positive y

Leaky ReLU negative y until x=0, then positive y

What is backpropagation: a method used in artificial neural networks to calculate a gradient that is needed in the calculation of the weights to be used in the network

How does it work: a = σ(Wx + b)

What is a convolutional layer:

Why use it:

What is a recurrent neural network: a class of artificial neural network where connections between nodes form a directed graph along a temporal sequence.

a = σ(Wx + b), s’ = a

Why use it:

**Classification and regression trees**

Components of a classification tree: root node, leaf nodes, internal nodes, branches

Shape of possible decision boundaries for CTs: square shape

Basic method for learning the structure of a CT:

Basic method for estimating the parameters of a CT:

Typical hyper-parameters of a CT construction algorithm:

Tradeoffs of CTs vs. other classifiers:

Interpretability: the learned model is easy to examine and understand as a collection of rules

Test Complexity: shallow trees can be extremely fast classifiers at test time

Train Complexity

Representation

Selectivity

Cost-complexity pruning and how it works:

Definition of regression tree: each internal node contains a rule that tests the value of a single feature d against a threshold value t and assigns the data case to its left or right sub-tree according to the result

How it differs from a CT: regression tree takes numerical (continuous) variable, classification tree takes categorical (discrete) variable.

**Feature selection**

Best-subset selection: begin with the null model and select the best model among (p selects k) models for each value of k (1, 2, … p). Then select the single best model among these p models using cross-validated prediction error.

Why don’t we use it: it’s computationally expensive, and since there are many possibilities for k, the model we select has high variance.

Forward selection begins with the null model and fit p linear regression models and pick the best one among them (best means least RSS or highest R-squared), then search through the remaining p-1 and add the variable that can improve RSS of current model. Continue doing this until stopping rule is satisfied.

Backward selection begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor on at a time.

**Overfitting, generalization and regularization**

Examples of estimators, estimates, estimands within machine learning:

Estimator: the model, the rule

Estimand: a quantity of interest

Estimate: the result

Evaluating a model:

Estimator – held-out test set and error rate

Estimand – misclassification error

Estimate – 0.153

Observed data – different independent held-out test sets

Learning a model:

Estimator – the logistic regression learning procedure (and the model it produces)

Estimand – y for a specific data instance X

Estimate – 1

Observed data – different independent training sets

How can estimates of the generalization error of models produced by ML algorithms be decomposed:

How do different components of generalization error change as model capacity changes:

Model capacity is the ability of a model to represent complex mappings from inputs to outputs.

Why don’t we always use the highest capacity model:

1. Accuracy tends to vary across datasets.
2. Accuracy estimates that use training data systematically overestimate the accuracy that will be obtained on test data.
3. High capacity models can increase training set accuracy without increasing test set accuracy.

Hyperparameter is a parameter of a model family that controls the capacity of the learned model.

Example of hyperparameter is the value k of k-nearest neighbor model or bandwidth parameter of a kernel density estimator.

What is generalization error:

What is the relationship between the data set used to estimate generalization error and whether that estimate will be biased:

**Multiple comparisons**

Three necessary and sufficient conditions for a multiple comparison procedure:

Multiplicity: generate multiple items

Variation: estimate scores with variance

Selection: select extreme-scoring item

Is the score of the item selected by an MCP an unbiased estimate of its true score: no

Why: the expected value of the maximum score will vary based on the number of models from which the argmax is selected

What can be done to obtain an unbiased estimate of the score of that item: increase the number of models

**Bias and variance of learned models**

Three components of error for a learned statistical model: bias, variance and noise

Definitions of those components, both mathematically and conceptually:

Bias is an error from erroneous assumption in the learning algorithm.

Bias(f’(x)) = E[f’(x) – f(x)]

Variance is an error from sensitivity to small fluctuations in the training set.

Var(f’(x)) = E[f’(x)^2] – E[f’(x)]^2

Noise

Noise(f’(x)) = sigma^2

The connection between alternative samples from the population and the bias-variance decomposition:

Under what conditions is the bias of a learned model virtually guaranteed to be large: the model learned from different samples produce predictions that show consistent error for specific test-set instances.

Under what conditions is the variance of a learned model guaranteed to be large: the model learned from different training sets produce predictions that vary about a mean for specific test-set instances.

Why bias and variance can be “traded off” against each other:

Typical behavior of bias, variance and noise as model capacity changes: as model capacity grows, the bias goes down, variance goes up, noise stays constant.

**Cross-validation**

How to create cross-validation samples: partition a dataset D into k equally sized partitions, and then use k-1 partitions for training and estimate generalization performance based on the held out partition.

Common values of k are 5, 10, and N, the latter is referred to as leave-one-out cross-validation.

What to do with those samples to estimate model accuracy: do the same process k times and then select the best-performing model based on D and use the k partitions as the estimate of generalization performance.

Why is cross-validation useful: cross-validation combines average measures of fitness in prediction to derive a more accurate estimate of model prediction performance.

What are the alternatives to CV: separate training and test sets.

Why use nested CV (with both an inner and outer loop): because the best hyperparameter value and an unbiased estimate of accuracy can be obtained simultaneously with nested CV.

**ROC Curves**

What are the dimensions of ROC space: (0-1) false positive rate on x-axis, (0-1) true positive rate on y-axis

What do points in ROC space correspond to: ranking classifiers and corresponding thresholds

What does a ROC curve correspond to: a classifier

How to construct a ROC curve for a classifier:

The diagonal line from (0,0) to (1,1) in ROC space correspond to: a classifier with random accuracy

What does the curve of a perfect classifier look like in ROC curve: a vertical line and a horizontal line, (0,1) is the point represents the perfect classification.

What is a common summary measure used to compare ranking classifier based on their ROC curve: Area under the curve (AUC)

**Ensembles**

What is regression toward the mean: the phenomenon that if a variable is extreme on its first measurement, it will tend to be closer to the average on its second measurements, and if it is extreme on its second measurement, it will tend to have been closer to the average on its first.

What is an ensemble: an ensemble is simply a collection of models that are all trained to perform the same task.

Why do ensembles improve generalization error: ensembles combine several machine learning techniques into one predictive model in order to decrease variance(bagging), bias(boosting), or improve prediction(stacking). Thus as the capacity of model grows, the bias goes down and ensemble decreases variance, thus the generalization error will goes down.

Ways that you can construct an ensemble:

Bagging

Bootstrap sampling

Boosting

Stacking

What is bagging: it takes a single training set and randomly sub-samples from it K times to form K training sets, and use these sub-samples to train different models.

What is a bootstrap sample:

What are random forests and how are they constructed:

What is boosting: boosting is an ensemble method that iteratively re-weights instances in the dataset instead of randomly resampling them.

How boosting differs from bagging: bagging can be thought of as learning a set of **parallel** classifiers, boosting can be thought of as learning a set of **serial** classifiers.

What is stacking: stacking is an ensemble method to form a train-validation-test split and train many classifiers fk(x) on the training data.

What are the key differences between stacking and other ensemble techniques (bagging and boosting): stacking is typically used to combine models from several different model families.