**Intro to Machine Learning** – Spring 2018 Course Outline

A lot of this structure has been guided by [An Introduction to Statistical Learning](http://www-bcf.usc.edu/~gareth/ISL/) as it is a very intuitive resource for learning that basics of machine learning. More advanced topics can be found in [The Elements of Statistical Learning](https://web.stanford.edu/~hastie/ElemStatLearn/)

1. Foundations
   1. What’s the point? What does Machine Learning encompass? What will we focus on in this brief course?
   2. Brief history of machine learning and its relationship to traditional statistics
   3. Challenges with healthcare to keep in mind
      1. HIPPA compliance and data security
      2. Interpretability is often important
      3. Data can be extremely complex
      4. Any model errors could be life threatening.
2. Basics
   1. The overarching goal is to find a function which generates desired outputs from the inputs
   2. Prediction vs. Inference
      1. Inference: Given a set of data you want to infer how the output is generated as a function of the data and infer the effect of each variable.
      2. Prediction: Given a new measurement, you want to use an existing data set to build a model that reliably chooses the correct identifier from a set of outcomes.
      3. Prediction does not aim to find most accurate relationship like inference. Instead, prediction only cares about accuracy of the model
      4. “How much extra will a house be worth if it has a view of the river? This is an **inference** problem. Alternatively, one may simply be interested in predicting the value of a home given its characteristics: is this house under- or over-valued? This is a **prediction** problem.”
   3. Parametric vs. Nonparametric
      1. Parametric models tend to be easier to interpret but less flexible, while non-parametric models make fewer assumptions about the underlying relationships
      2. A linear regression, for instance, is parametric as it makes assumptions about the data being normally distributed. It will tend to have larger errors than a non-parametric approach, but it is much easier to understand, explain, and is less prone to over-fitting the data
      3. A smoothing spline is an example of a non-parametric function as it makes no assumptions about the distribution of the input data. However, because it can have many exponential coefficients, it is more difficult to interpret the model coefficients.
   4. Trade-off between accuracy and interpretability
      1. This is really a question of flexibility: a rigid model, such as a linear regression, is extremely good for inference as you can see the direct changes in estimation of Y when you change a single input. Something like neural networks, however, is extremely flexible, and more accurate, but comes with a loss of interpretability.
   5. Supervised vs Unsupervised Learning
      1. In supervised learning, the algorithm is given a set of inputs and desired outputs. In unsupervised learning no labels are given, leaving any pattern or structure discovery up to the algorithm.
      2. Since unsupervised learning has no “target” variable, models such as linear regression do not fit into this category. However, there is still a lot of power in these methods for discovering relationships among the variables, and across observations.
   6. Regression vs Classification vs Clustering
      1. Regression and classification are often supervised learning problems; regression for predicted continuous labels, Classification for predicting discrete labels. Clustering is typically unsupervised classification as the groups (or discrete values) are unknown.
      2. Resource: Machine learning applications in cancer prognosis and prediction <https://www.sciencedirect.com/science/article/pii/S2001037014000464>
3. Model Accuracy
   1. Importance of optimization
      1. There is no machine learning model that fits all problems. There are many factors that will determine which models best explain your data, such as size and complexity of the data, the distributions, available computational power, etc.
      2. Generally, it is helpful to apply different models to your data and assess the accuracy of each separately.
   2. Measuring quality of fit
      1. Mean squared error - a measure of regression model performs
      2. Error rate and AUC – a measure of classification model performs
      3. Importance of training and testing
         1. We start by taking training data to train the models and try to find a minimum error, but this needs to be applied to a test set as well
         2. If the test set errors are much higher, than the model over-fit the training data
         3. It’s crucial to track errors for a model in both a training and test data set to fully understand the potential variability in errors.
         4. As model flexibility increase, training MSE will decrease, but testing MSE may or may not.
         5. In general, the goal should be to minimize testing errors.
         6. Test errors tend to follow a u-shaped curve as model flexibility increase, expressing a trade-off between variance and bias
      4. Bias/Variance trade off in test errors
         1. Variance is the amount our estimation will change if we use a different training data set. Ideally, variance of our estimator is low with different training data sets. This is why cross-validation is so crucial. In general, more flexible models have higher variance.
         2. Bias are the errors that result from trying to fit complex real-life problems into simple, easy to understand models. Linear regression assumes a linear relationship between inputs and targets, but there are often many competing inputs that aren’t captured accurately by the model. In general, more flexibility leads to lower bias in the model.
         3. <https://www.analyticsvidhya.com/wp-content/uploads/2015/07/variance_bias.png>
4. Regression Techniques
   1. Simple linear regression (ordinary least squares)
      1. One input, one target
      2. Inflexible due to big assumptions: linear relationship, normally-distributed data and errors, predictors are independent of one another.
      3. Easy to interpret the coefficients (intercept and slope)
      4. Coefficients determined by minimizing the residual sum-of-squares, which are the collective squared errors of the model across all observations
      5. The main method for testing the accuracy of the coefficients is to construct confidence intervals around them. If “0” is included in the range of confidence, then the coefficient and the variable itself provide very little explanation of the target and it is a weak relationship.
      6. T-tests are used to assign a p-value to each coefficient, expressing it’s “usefulness” in the final model.
      7. Model accuracy is assessed with RSE, R^2, and F-statistic
   2. Multiple linear regression
      1. Two or more predictors for a single target. This brings up a lot of new issues that must be dealt with
      2. The first step should be to compute the F statistic for all coefficients in the model. If at least one coefficient points to a significant predictor, then we have traction.
      3. Highly correlated predictors will result in odd coefficients. So, how can we choose which predictors to keep? There are many variable selection methods. We can use AIC, BIC, or adjusted R^2 on many different subsets of predictors, aiming for the most accurate group.
         1. Forward selection – Start with one predictor, and build on it until lowest RSS is achieved
         2. Backwards selection – Start will all predictors, and trim those variables with the largest p-values
         3. Mixed selection – Start by adding predictors, and once a new addition has a p-value above some threshold, it is removed, while another is added.
         4. If predictors > observations, backward selection cannot be used as there is not enough data to optimize the predictors. But, you can always start with forward selection.
      4. Qualitative variables are handled using dummy variables.
      5. Polynomial terms can be used in a linear regression for more complicated relationships.
      6. Potential problems
         1. Non-linear relationship between predictors and the target – Residual plots can be used to investigate this. IF the appropriate fit for residual errors is not a straight line, this is a good indication that the relationship is not linear. Using polynomial terms can help with this.
         2. Correlation of error terms – very common in time series data. In this case, errors are essentially “Carried forward” through proceeding observations.
         3. Heteroscedasticity – Residual plots work great for identifying this problem as well.
         4. Outliers – Can arise for many reasons, including erroneous data. An outlier can significantly skew coefficients and the final model, so often times we may want to remove them.
         5. Collinearity – When two or more predictors are highly correlated, it can be hard to separate the effect of each on the target. Can either drop one of the variables, or combine them into a single predictor.
      7. Other regression techniques
         1. Ridge Regression – Very similar to linear regression, except it contains a shrinkage penalty and a tuning parameter, lambda. If lambda set to zero, we are left with the OLS estimator. AS it approaches infinity, all coefficients (aside from the intercept) are penalized to zero. Turns out this method can find a set of coefficients which has much less variance and the cost of slightly higher bias. It also is computational easier to implement than subset selection. Works best when the OLS estimates have high variance.
         2. The Lasso – Unlike ridge, which keeps (but penalizes) all predictors, Lasso offers a slight variation which actually drops predictors along the way. As the tuning parameter becomes large enough, some of coefficients for the less informative predictors will be forced to zero, essentially removing them from the model.
         3. Principal Component Regression – Applies PCA to the data set to transform it into a set of principal components which are used as the predictors. The key idea is that you often only need a few PCs to be able to explain the variability of the data set. However, we don’t know if the linear combinations that make up the PCs and best explain the predictors also best explain the response. Unsupervised.
5. Classification Techniques
   1. From a healthcare perspective we might want to classify patients based on various outcomes: Alive or dead, emergency department visit or not, tumor stage, etc.
   2. Regression will not work for multi-level classifiers even if we create a dummy variable for the target. Even for binary classification, there are still errors. A linear regression model will often predict values less than 0 or greater than 1!
   3. Logistic regression
      1. Models the **probability** the target belongs to a category, rather than the **response** of the target as in a linear regression setting.
      2. We must generate a function which takes any values for the predictors and puts them on a [0,1] range, which is a logistic function.
      3. Fitting the model coefficients uses a method called maximum likelihood, which aims to accurately classify as many observations as possible.
      4. Like linear regression’s t-test, we can apply a z-test to the coefficients of our logistic model to make decisions on which predictors to use.
      5. Multiple logistic regression is a simple extension of the model to more than one predictor.
      6. However, we are limited by the binary nature of logistic regression. We can only assign to two categories.
   4. Linear discriminant analysis
      1. LDA solves some of the short comings of logistic regression by harnessing Bayes’ theorem
      2. For one predictor, LDA basically solves for the greatest separation of the distributions of the predictor for each category. If the distributions overlap, there will be some uncertainty. It does this by approximating the Bayes classifier (which in many real life situations we cannot calculate).
      3. “To reiterate, the LDA classifier results from assuming that the observations within each class come from a normal distribution with a class-specific mean vector and a common variance, and plugging estimates for these parameters into a Bayes classifier.”
      4. This framework can be extended to more than one predictor by assuming that the set of predictors are drawn from a multivariate normal distribution, with class-specific means and a shared covariance matrix.
      5. Quadratic discriminant analysis is an extension which releases the assumptions of common covariance across predictors.
   5. K-Nearest Neighbor
      1. A very simple, non-parametric method for assigning or discovering classes. It estimates the conditional probability that the observation in question is each of the classes based on k-neighbors. The highest probability wins.
      2. If k=1, the class of the nearest neighbor to each point becomes its assignment while predicting. This is the most flexible approach, with very low training errors but high testing errors.
   6. Support Vector Machines
      1. Perform well in many settings, often a great benchmark. IT classifies data by finding a “hyperplane” which maximizes separation of the classes.
      2. Works well even when p > n
      3. Can also be used for regression
   7. Confusion matrix
      1. Resource: [https://en.wikipedia.org/wiki/Sensitivity\_and\_specificity#Confusion\_matrix](https://en.wikipedia.org/wiki/Sensitivity_and_specificity%23Confusion_matrix)
      2. A matrix showing predicted versus true classifications. Overall error rate can be low, but that might not be enough.
      3. Class-specific performance
         1. Sensitivity measure the proportion of positives that are correctly identified
         2. Specificity measures the proportion of negatives that are correctly identified
      4. Overall, the model might perform very accurately, but if the classes are poorly weighted, you might find that sensitivity or specificity perform very poorly. LDA tries to approximate Bayes’ classifier, which has the lowest total error rate across, regardless of class.
   8. ROC Curves measure the false positive rate against the true positive rate. The “area-under-the-curve” (AUC) tells us how accurate the model overall is. An area of 1 expresses a perfect model, while an area of 0.5 expresses that the model is as good as random guessing.
6. Cross-Validation
   1. As noted before, it is crucial to have a testing and training data set to get a better idea of out-of-sample error variance.
   2. We can take it even further, and train the same model on many different subsets of the data, and see by how much the model coefficients are changing. This also allows us to pick an appropriate level of flexibility for the model (one that minimizes the errors of the testing data).
   3. Used to estimate test error associated with the machine learning technique and evaluate the model, or to test different levels of flexibility
   4. In its simplest form, it’s just the training/testing split known as the “validation set” approach
   5. K-fold cross validation is another method where the practitioner divides the set of observations into k-groups. Each group is then, in turn, saved as the validation set and we are left with k-MSEs, one for each validation set. We can then see the variance of the MSE. High variance means that the model we selected is sensitive to the underlying data.
   6. Cross validation does not need to be applied to many data splits of the same model, we can also use it across different models to find which ones provides a minimum MSE and minimum MSE variance.
   7. For classification problems, the logic still works the same way, but we measure error rates instead of MSE
7. Tree-based methods
   1. Can be used for both regression and classification, these methods involve applying a hierarchy to the set of predictors based on significance. These are simple and useful for interpretation, but are often not the most accurate predictive models.
   2. Better over linear models for non-linear and complex relationships.
   3. Can handle qualitative variables without use of “dummy” variables
   4. Bagging, boosting, and random forests
8. Ensemble methods
   1. Multiple learners are trained to solve the same problem.
   2. Bagging, boosting, AdaBoost
   3. Requires voting to weight model classifications
   4. Resource: Classification of lung cancer using ensemble-based feature selection and machine learning methods [http://pubs.rsc.org/-/content/articlelanding/2015/mb/c4mb00659c/unauth#!divAbstract](http://pubs.rsc.org/-/content/articlelanding/2015/mb/c4mb00659c/unauth%23!divAbstract)
9. Neural Networks
   1. A collection of artificial neurons, connected in various ways
   2. Each neuron takes a set of inputs and calculates an output with a non-linear “activation” function.
   3. All connections between neurons are weighted, which adjust as the learning process takes place.
   4. Neurons are often organized into layers, often with different purposes.
   5. The input layer contains one input for each predictor in your data set. The output layer will be one neuron if regression, or k-neurons if classifying into k-groups. Everything in between are “hidden layers”.
   6. Impossible to interpret “why” an output value was given for a set on inputs as it’s a highly nonlinear and intractable problem
   7. Many variants in this space for solving different
   8. <http://pages.cs.wisc.edu/~bolo/shipyard/neural/local.html>
10. Deep Learning , Evolutionary Computation, and other recent developments (if there is time, wrap-up stuff)