# What is ML?

How does it work? How is it different from regular statistics and econometrics? What types of policy problems can machine learning help inform?

Learning

* Regression problem: combo of linear X’s that predict some continuous or quantitative variable
  + Usually OLS
* Classification problem: predicting what bucket (Y) the combo of X’s will fall into
  + Predicting whether the stock market will increase or decrease tomorrow
  + Usually use logistic regression
* Clustering problem: only observe input variable with no corresponding output, groups similar points together based on observed characteristics

Knowing

* Human sentiment analysis is automatic
  + Magic trick to accuracy is to stop introspecting, stop looking for an insight, just look at data and treat the known like we treat the unknown

ML

* Turn any intelligence task into an empirical learning task
  + Specify what is to be predicted
  + Specify what is used to predict it
* Need a lot of data
* ML in PubPol is good for
  + Prediction: ML predicting re-arrest risk rather than relying on arbitrary judge
  + Causation: ML controlling for X’s and trying to predict effects
* Causal Inference is about trying to estimate B1 (what explains Y?)
  + Hard when X1 and X2 highly correlated
  + Omitted variables bias: people select into most policies
* Prediction is about trying to predict Y
  + We don’t care what in Y is coming from which X, we just want high accuracy

\*\*Do omitted variables pose a problem for prediction??

## What is statistical learning?

* How to estimate f, some fixed but unknown function
  + Y = f(X) + e
* We want to estimate f for prediction and inference
* Usually, we know X’s but do not know Y, so both f and Y are predicted. F is treated as a black box as we don’t care about the exact form or relationship, just the accuracy of predictions
  + Accuracy of Y^ as a prediction for Y depends on reducible and irreducible error
    - We can likely improve the accuracy of f^ by using the best stat learning technique
    - No matter what, our prediction will still have some error in it, no matter how well we predict f, we cannot reduce the error introduced by e

Inference

* When we want to understand how Y is affected by changes in X
  + Can’t treat f has black box because we need its form
* Which predictors are associated with the response?
  + Be able to identify few important predictors among a large set of possible variables
* What is the relationship between the response and each predictor?
* Can the relationship between Y and the predictor be adequately summarized using a linear equation, or is the relationship more complicated?

Inference v Prediction

* Inference: How much extra will a house be worth with a view?
* Prediction: How much will this house cost given some characteristics?
* Linear models allow for simple and interpretable inference, less accurate predictions
* Non-linear models: accurate predictions for Y, less interpretable model

Parametric Methods

1. Make assumption about functional form (shape) of f. (for example, linear)
2. After you select a model, pick a procedure that uses the training data to train the model.
   1. If linear, estimate Beta coefficients (usually though OLS)

* Reduces problem of estimating f down to one of estimating a set of parameters
* More complex models can lead to overfitting (follow errors too closely)

Non-parametric Methods

* No explicit assumptions about f
* Seek an estimate of f that gets as close to the data points as possible
* Can accurately fit a wider range of possible shapes for f
* Need a very large number of observation for accuracy

Supervised vs Unsupervised learning

* Supervised: aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between response and predictors (inference)
  + Linear regression
  + Logistic regression
  + GAM, boosting, support vector machines
* Unsupervised learning: we observe x’s but no y
  + Relationships between the variables or between observations
  + Cluster analysis: to see if observations fall into distinct groups

## Model Accuracy

Measuring Quality of Fit

* Mean Squared Error: extent to which the predicted response value for a given observation is close to the true response
  + We care about MSE on test data not training data
* As model flexibility increases, training MSE will decrease, test MSE may not
  + Overfitting 🡪 When small training MSE and large test MSE
* Can use cross-validation to estimate test MSE using training data
* MSE curves are usually U-shaped

Bias-Variance Trade-Off

* U shape curve because of bias and variance
* Expected test MSE = variance, squared bias, variance of the error
* To minimize test error, we need low variance and low bias
* Variance: amount by which f would change if we estimated it using a different training data set
  + Estimate for f should not vary too much between training sets
  + More flexible statistical methods have higher variance
* Bias: error that is introduced by approximating a complicated problem by a much simpler model
  + Linear regression results in high bias
  + More flexible methods result in less bias

Classification Setting

* For quantifying accuracy of estimate of f using classification, use error rate
* Error rate: the proportion of mistakes that are made if we apply our estimate f to the training observations
* Training error rate vs test error rate
  + Good classifier is one where test error is smallest
* Bayes Classifier
  + Test error rate minimized by classifier that assigns each observation to the most likely class given its predictor values
  + Uses a Bayes Decision Boundary, draws line or classification between two groups, anything on one side is A other is B
  + Produces the lowest possible test error rate (Bayes error rate)
* K-nearest neighbors
  + For real data, we don’t know the conditional distribution of Y given X
  + KNN classifier we try to achieve that golden standard
    - Identifies the K points in the training data that are closest to Xo. Estimates conditional probability for class j as the fractio of points closest to xo whose response values equal j.
    - Applies ayes rule and classifies the test observation xo to the class with the largest probability
    - Choose K. Identify k observations that are closest to the point. Whatever majority classification is in those k, use that.
  + Test error rate is very close to Bayes
  + Choice of K has a drastic effect on KNN classifier
    - When low, method is overly flexible and has low bias, high variance
    - When high, method less flexible and gets closer to linear, low variance but high bias
  + Elbow plot for best value of K (K vs error rate)

# Linear Regression

* Supervised Learning
* Questions to consider
  + Is there a relationship between x and y?
  + How strong is the relationship?
  + Which x is relevant to Y?
  + How accurately can we estimate the effect of each X on Y?
  + How accurately can we predict future Y?
  + Is the relationship linear?
  + Are there interaction effects between the X’s?
    - Simpson’s Paradox: linearizing hides groups, pooled data swamps within-group variation
* Simple linear regression: assumes approximately linear relationship, uses single predictor X
* Least Squares: most common way of estimating Betas
  + Minimizes residual sum of squares (RSS)
* Requirements
  + Assume linearity
  + Assume errors are normally distributed
  + Unbiased estimators
* Diminishing returns for extra sample size

Convex Loss

* Where large x differences, little y differences
* Want regression line that best fits a combination of true underlying f(X) in sample data
* Penalizes “big misses” – want to check data for outliers

Accuracy of Coeff Estimates

* Variance = sd^2 = v
* Standard error = variance/sqrt(n)
* Use standard errors to compute confidence intervals
  + B +- 2SE

Accuracy of Model

* Residual Standard Error (RSE)
  + Measure of the lack of fit of the model
* R^2
  + Proportion of variance explained, between 0 and 1

## Multiple Linear Regression

* Is there a relationship between at least one of the predictors and response?
  + F test > 1
  + P < alpha
  + Always check F too because it adjusts for the number of predictors and observations
* Deciding on important variables
  + Forward selection: Null model,
  + Backward selection: All predictors
  + Mixed selection: combo of forward/backward selection
* Model fit
  + Substantial improvement in R^2 rather than due to adding more variables
  + Plotting data
* Predictions
  + Types of uncertainty: reducible error, model bias, irredudible error
    - Use prediction intervals (wider than confidence interval)

Potential Problems

* Non-linearity of data
  + Use residual plots to identify non linearity
* Correlation of error terms
  + Happens in time series data
* Non-constant variance of error terms: heteroskedasticity
  + Check residual plot
* Outliers
  + Residual plots, check and make smart decision
* High leverage points
  + Unusual value for x, have large impact on regression line
  + Be smart about removing
* Collinearity: two or more predictor variables are closely related to each other
  + Check scatterplots – correlation matrix
  + Causes standard error to grow significantly
  + Causes decline in t stat

Linear Reg vs KNN

* Lin Reg = Parametric
  + Easy to fit, simple interepretations, tests of statistical significance easy
  + Make strong assumptions about form, may be a poor fit to data
* KNN regression = nonparametric
  + Given a value of K and prediction point x, KNN regression identifies the K training observations closest to x, estimates f(x) using average of all training responses near x.
  + Optimal value of K depends on bias variance trade off
    - Small number is most flexible, low bias high variance
* The parametric approach will outperform the nonparametric approach if the parametric form that has been selected is close to the true form of f
* Parametric approach will outperform nonparametric when there is a small number of observations per predictor

# Classification

* When response variable is qualitative
* Classifiers: logistic regression, K-nearest neighbors
* Use training observations to build a classifier
  + Want it to perform on test observations
* For binary response with 0/1 coding, OLS does make sense and base it based on probability
  + But could get results outside of 0/1, results hard to interpret
* Logistic Regression: models the probability that Y belongs to particular category
  + Uses maximum likelihood function
  + Estimates chosen to maximize
  + P^(X) = e^(B0 + B1X)/ (1 + e^(B0 + B1X))
  + One unit change in x changes odds of getting a 1 versus 0 by e^B
* Accuracy
  + Error rate
  + True negative, true positive, false negative, false positive
    - Rates can be misleading
  + Tradeoff between TP rate vs FP rate
  + Can change thresholds

# Week 1 Reading – Prediction Policy Problems

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan and Ziad Obermeyer (2015) “Prediction policy problems,” American Economic Review, Papers & Proceedings, 105(5): 491-5.

Most policy analyses require causation to evaluate their effectiveness. Using a before and after analysis of policy answers questions about what should be done. But, machine learning could add value in terms of prediction problems. These kinds of prediction problems don’t impact the overall action (Y is not caused by X), but, they can be used to improve policy by more efficiently making decisions and predictions. One example is surgery for high risk patients, and redistributing resources to nonfutile surgeries. Another is using ML to predict flight risk in the criminal justice system in order to release or detain people. Prediction potentials have wide applications and should be considered more for better decision making.

Reading

Ch 5, 6 (80 pgs)

Prediction policy problems

Domingos, Pedro (2012) “A few useful things to know about machine learning” https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

Slides

Ch 8 (30 pgs)

Slides

Ch 9 (30 pgs)

Slides

Ch 10 (40 pgs)

Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne, and Vipin Kumar (2018) “Chapter 5, Association analysis: basic concepts and algorithms,” Introduction to Data Mining (Second Edition), pages 357-380 only; available at https://www-users.cs.umn.edu/~kumar001/dmbook/ch5\_association\_analysis.pdf Hanna M. Wallach (2006)

“Topic modeling: Beyond bag-of-words” ICML 2006 Proceedings of the 23rd Int’l Conference on Machine Learning, pp. 977-984. https://dl.acm.org/citation.cfm?id=1143967

Slides