Final Exam Review

Sravani Vadlamani

Final Exam

- In Class on Tuesday, April 26, 2022
- Please bring your computer as the exam is on canvas
- Do not forget your calculator

Supervised vs Unsupervised Learning

- Supervised Learning
  - Predict or estimate an output based on one or more inputs
  - Linear regression, logistic regression, boosting, support vector machines
- Unsupervised Learning
  - Learn relationships and structure from data with inputs but no outputs
  - Cluster Analysis

Input vs Output Variables

- Input Variables
  - Independent variables, predictors features
- Output Variables
  - Response or dependent variable
- We believe there is a relationship between Y and at least one of the X's. We model the relation as

$$Y_i = f(X_i) + \varepsilon$$

- Where ***f*** is an unknown function and ε is a random error with mean zero
- Parametric vs Non-Parametric Approaches

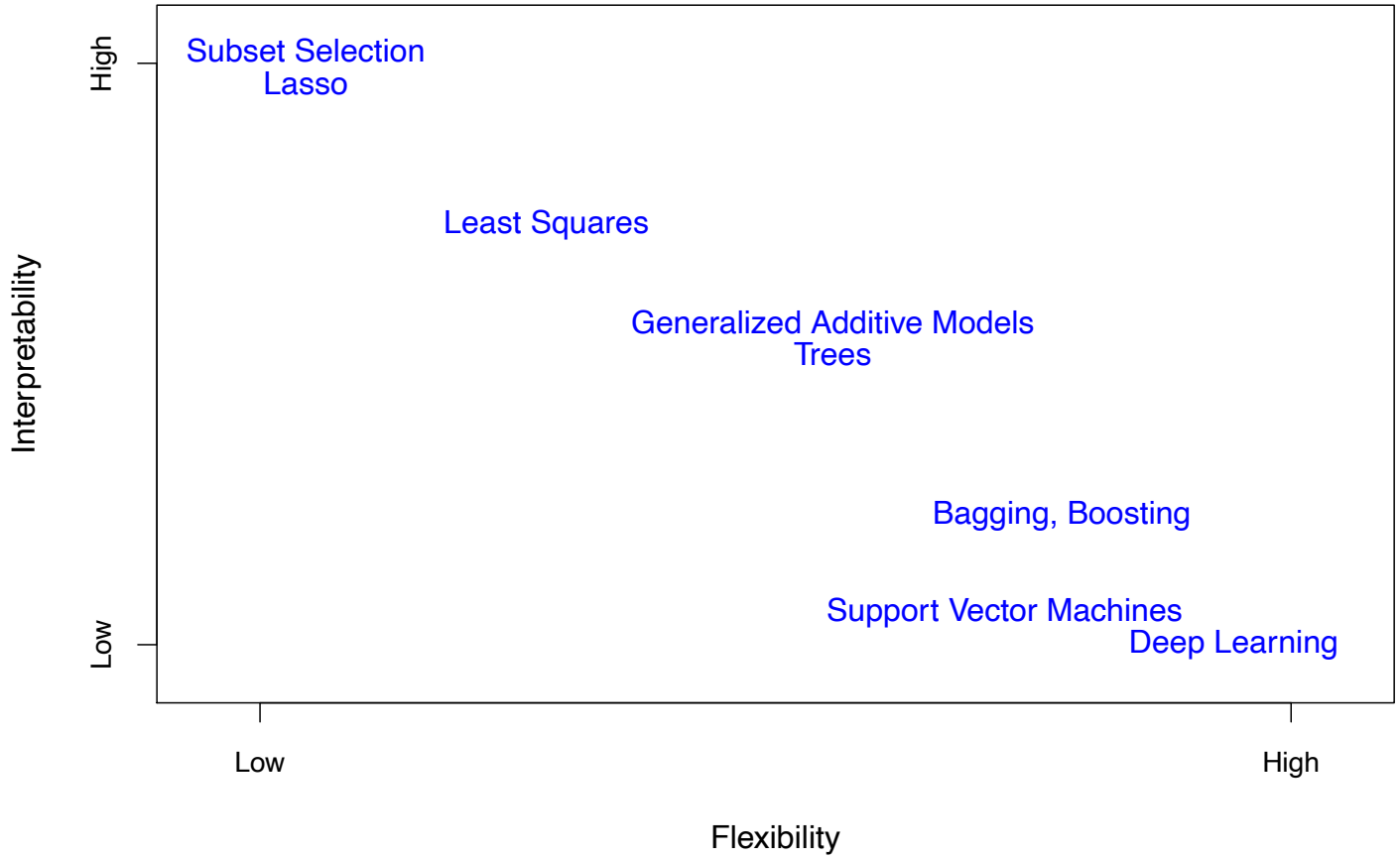
Flexible Models Vs Overfitting

- The more flexible the model, the more realistic it is. However, more flexible models have the disadvantage of requiring a greater number of parameters to be estimated and they are also more susceptible to overfitting.
- Overfitting is a phenomenon where a model closely matches the training data such that it captures too much of the noise or error in the data. This results in a model that fits the training data very well but does not make good predictions under test or in general.

Prediction Accuracy vs Model Interpretability

- Non-linear regression methods are more flexible and can potentially provide more accurate estimates.
- Why not just use a more flexible method if it is more realistic?
- A simple method such as linear regression produces a model which is much easier to interpret (the Inference part is better). For example, in a linear model, βj is the average increase in Y for a one unit increase in Xj holding all other variables constant.
- Even if you are only interested in prediction, so the first reason is not relevant, it is often possible to get more accurate predictions with a simple, instead of a complicated, model. This seems counter intuitive but has to do with the fact that it is harder to fit a more flexible model.

Prediction Accuracy vs Model Interpretability



Regression vs Classification

- When are both methods used?

Quality of Fit for Regression

- To evaluate the performance of a model, it is necessary to quantify how close the predicted responses are to the observed/actual data
- One common measure of accuracy in regression method is the mean squared error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Where, $\hat{y}_i$ *is the predicted responses*

Bias Variance Tradeoff

- Choice of learning method is governed by two competing factors – bias and variance
- Bias refers to the error introduced by modeling a usually extremely complicated problem using a simple problem
- For example – a linear regression model assumes a linear relation between Y and X which may be unlikely in real life thus introducing some bias
- More flexible (or complex) models have less bias

Bias Variance Tradeoff

- Variance refers to the amount by which $\boldsymbol{f}$ would change if it were estimated with a different training set
- The more flexible a method is, greater is its variance

- In general, as the flexibility of the statistical method increases, its variance increases and bias decreases
- The relationship between bias, variance, and test set mean squared error is referred to as the bias-variance trade-off. It is called a trade-off because it is a challenge to find a model that has both a low variance and a low squared bias.

Bias Variance Tradeoff

- For any given $x = x_0$, the expected test mean squared error can be decomposed into the sum of the following three quantities:
  - Variance of $f(x_0)$
  - Squared bias of $f(x_0)$
  - Variance of the error term ($\boldsymbol{\varepsilon}$)

$$Expected\,Test\,MSE = E\left(Y - f(x_0)\right)^2 = Bias^2 + Var + \underbrace{\sigma^2}_{\text{Irreducible Error}}$$

- To minimize the expected test error, it's necessary to choose a method that achieves both low variance and low bias. It can be seen that the expected test mean squared error can never be less than , the irreducible error.

Linear Algebra Basics

- Matrices
  - Row and Column Vectors
  - Transpose
  - Symmetric Matrix
  - Diagonal Matrix
  - Matrix Addition/Subtraction
  - Matrix Multiplication
  - Hilbert Matrix
  - Trace of a square matrix

Linear Regression

- Simple and Multiple Linear Regression
- Parameter estimation using least squares
- Forward vs backward variable selection
- Quantitative and Qualitative predictors
- Model Fit
  - Residual Standard Error
  - R-square – proportion of variance explained by the model

Linear Regression

- Issues
  - Non-linearity (if present transform predictor variables)
  - Correlation of error terms
  - Homoscedasticity violation (transform the response variable)
  - Outliers
  - Multicollinearity (VIF)

Classification Techniques

- Logistic regression
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Naïve Bayes
- K-nearest neighbors
- Generalized additive models
- Decision Trees, Random forests, Boosting
- Support Vector Machines

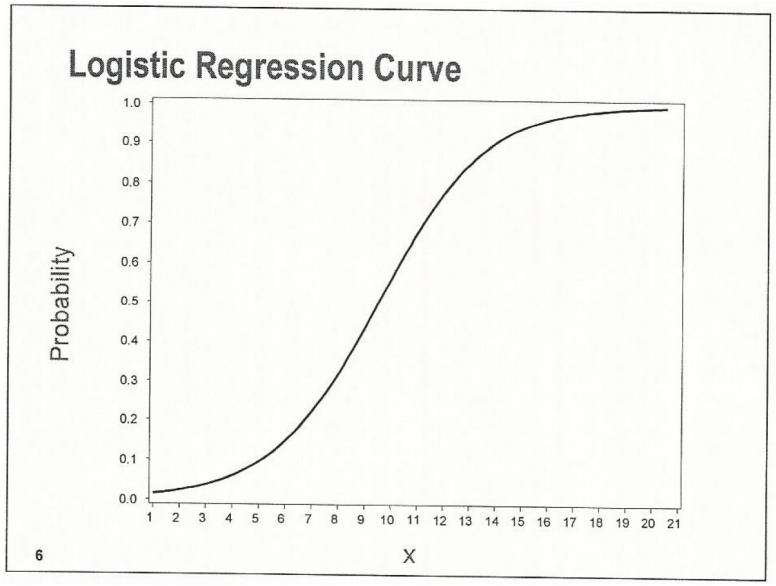
Logistic Regression

- We use a logistic function to model p(x) such that the output is always between 0 and 1 for all values of X

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

- The above equation can be rewritten as

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x}$$



Logistic Regression Parameter Interpretation

- In linear regression, β_1 gives the average change in Y associated with a one-unit increase in X
- In logistic regression, a one-unit change in X yields a β_1 change in the log-odds
  - This is equivalent to multiplying the odds by e^{β_1}
- If β_1 is positive, increasing X will increase $p(x)$
- If β_1 is negative, increasing X will decrease $p(x)$
- The rate of change in $p(x)$ per unit change in X depends on the value of X

Estimating Parameters

- The coefficients $\beta_0 \; and \; \beta_1$ must be estimated based on the training
- Logistic regression uses maximum likelihood to estimate $\beta_0 \; and \; \beta_1$ such that the predicted probability is as close to the observed classes using the following likelihood function

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i\prime:y_{i\prime}=0} (1 - p(x_i))$$

- The estimates $\beta_0 \; and \; \beta_1$ maximize the above function

Logistic Regression vs LDA

- In logistic regression we model the conditional distribution of the response variable, given the predictor(s) X i.e., $Pr(Y=k \mid X = x)$
- In LDA we model the distribution of the predictors separately in each of the response classes Y ($Pr(X|Y=k)$ and then use Bayes theorem to invert the probabilities to estimate the conditional distribution.
- Reasons to prefer LDA over Logistic regression
  - Parameter estimates from logistic regression are surprisingly unstable when there is a substantial separation between classes. Discriminant analysis approaches do not have this issue
  - Logistic regression estimates are less accurate when the predictors are approximately normally distributed, and the sample size (n) is small

LDA with One Predictor - Summary

- In summary, LDA assumes that the observations from each class follow a normal distribution with a class specific mean vector and constant variance across the classes to build a Bayes' theorem-based classifier.

Multiple Predictor LDA

- Assumes $X = (X_1, X_2, ...X_p)$ follows a multivariate normal or multivariate Gaussian distribution with a class-specific mean vector and a common covariance matrix.
- Multivariate normal distribution implies that each predictor follows a one-dimensional normal distribution with some correlation between the predictors. The bell shape of the normal distribution will be distorted if the predictors are highly correlated.

Quadratic Discriminant Analysis

- An alternative approach to LDA
- Same assumptions as LDA regarding the observations from each class following a Gaussian/Normal distribution.
- Difference is in the covariance estimation
  - QDA assumes each class has its own covariance matrix.
- This results in assuming that an observation from the k th class follows a distribution of the form $X \sim N(\mu_k, \Sigma_k)$ where Σ_k is the covariance matrix for k th class

Comparison of Methods

- Generally, LDA is better than QDA if there are relatively few training observations and so reducing variance is relevant.
- If the training set is very large that the variance of classifier is not an issue or if the assumption of common covariance matrix is unrealistic, QDA can be a better choice.
- LDA and logistic regression work well when the decision boundary is linear
- QDA gives better results when the decision boundary is moderately non-linear
- K-nearest neighbors (KNN) is a non-parametric approach and outperforms LDA and logistic regression when the decision-boundary is highly non-linear

Metrics

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I Error)
Test Negative	False Negative (Type II Error)	True Negative

- Sensitivity =Recall = $P(Test+ \mid Condition+) = \frac{TP}{TP+FN}$
- Specificity = $P(Test- \mid Condition-) = \frac{TN}{FP+TN}$
- False Negative Rate (β) = $P(Test- \mid Condition+) = \frac{FN}{TP+FN}$
- False Positive Rate (α) = $P(Test+ \mid Condition-) = \frac{FP}{FP+TN}$
- Precision = $\frac{TP}{TP+FP}$
- Sensitivity = 1 – False Negative Rate = Power
- Specificity = 1 – False Positive Rate
- F1 Score = $\frac{2*Precision*Recall}{Precision+Recall}$
- Accuracy = $\frac{Correct\ Predictions}{Total\ Predictions}$

Metrics

- Accuracy
  - % of correct predictions
  - One value for the entire model
- Prediction
  - Exactness of the model
  - Each class/label has a value
- Recall
  - Completeness of model
  - Correctly detected over total observations
  - Each class/label has a value
- F1 Score
  - Combines precision and recall (Harmonic mean of precision & recall)
  - Each class/label has a value

ROC Curves

- A curve for simultaneously displaying the positive and negative error types for various thresholds
- ROC stands for “Receiver Operating Characteristics” and derives its name from communications theory (historical)
- The x-axis is False Positive Rate (=1-sensitivity)
- The y-axis is True Positive Rate (= specificity)
- Overall performance is given by the Area Under the Curve, denoted as AUC.
- Larger the AUC, better the classifier i.e., the curve is closer to the top left corner

Linear Model Selection

- Best subset selection
- Stepwise selection
- Model selection criteria
  - Cp, AIC, BIC (lower the better)
  - Adjusted R-square (higher the better)
- Shrinkage methods
  - Ridge Regression (l-2 norm)
  - Lasso Regression (l-1 norm)

Resampling Techniques

- Validation Set Approach
- LOOCV
- K-fold
- Bootstrapping

Extensions to Linear Models

- Polynomial regression
  - Add extra predictors obtained by raising each of the original predictors to a power
- Step Functions
  - Split a continuous variable into k distinct regions to produce a qualitative variable. This has the effect of fitting a piecewise function
- Regression Splines
  - Extension of polynomial regression and step function and are more flexible
  - Divide the range of X into k distinct regions and fit a polynomial function within each region
  - The polynomials are constrained so that they join smoothly at the region boundaries called knots
  - When there are sufficient regions, the splines can result in an extremely flexible fit

Extensions to Linear Models

- Smoothing splines
  - Similar to regression splines but they minimize a residual sum of squares criterion subject to a smoothness penalty
- Local regression
  - Similar to smoothing splines but the regions are allowed to overlap in a smooth way
- Generalized Additive Models
  - Extend above methods to deal with multiple predictors

Regression Trees

- Regression trees identify variables that are important for prediction in a different way
  - Stratifying the prediction space into several simple regions
  - Identifies variable and cut point on the variable to partition the data and does this repeatedly
  - Allows for non-linear associations and interaction effects
  - Simple methods useful for interpretation but not competitive in terms of accuracy
  - Can be applied to both regression and classification problems

Trees vs Linear Models

- Which one is better?
- It depends
  - If the relation between the response and predictors is well approximated by a linear model, linear regression will outperform trees
  - If there is a highly non-linear and complex relation between the response and predictor variables, decision trees will outperform classical methods

Advantages and Disadvantages of Trees

- Advantages
  - Easy to explain (than linear regression!)
  - More closely mirrors human decision-making than the previously seen regression and classification methods
  - Easy to display and interpret for a non-technical audience
  - Handle qualitative predictors without creating dummy variables
- Disadvantages
  - Do not have the same level of predictive accuracy as other regression and classification approaches
  - Not very robust i.e., small changes in data and lead to drastic changes in the final estimated tree

Bagging

- Bagging = Bootstrap Aggregation
- Bootstrap
  - Resampling procedure where we sample a certain number of cases from our dataset with replacement
- Aggregation
  - The act of collecting together

Random Forests

- Extension of bagging with a goal to decorrelate the trees
- Bagged trees are correlated i.e., they are independent but not identically distributed
- To decorrelate the trees, we choose a random sample of observations AND a random sample of predictors

Boosting

- A general approach that can be applied to many statistical learning methods
- Shares similarities with Bagging and Random Forests where a collection of trees are grown
- Differences with Bagging and Random Forests
  - Boosting does not involve bootstrapped sampling
  - Trees are grown sequentially instead of independently
    - Each tree is grown based on the information from previously grown trees
    - Each tree fits to the residuals from the previous tree

Support Vector Machines

- A classification technique developed in the 1900s. Typically used for Two class classification
- A generalization of simple and intuitive classifier called the maximal margin classifier that can be applied to only classes separated by a linear boundary
- Support vector classifiers extend the application of maximal margin classifier to non-separable cases
- Support vector machines extend the support vector classifiers to accommodate non-linear class boundaries
- SVMs are originally intended for binary classification but can be extended to handle more than two classes

Unsupervised Methods

- Two Methods
  - Principal Component Analysis
    - A useful tool for data visualization and data pre-processing before applying supervised techniques
  - Clustering
    - A range of methods to discover unknown subgroups in the data