**Section 5: Fitting and Overfitting Models**

**Model Fitting:**

A process of finding the best model for a given data set and business objectives:  
 -Accuracy  
 - Speed  
 -Repeatability  
 -Explanatory Power  
 -Predictive power

**Generalization:**

How well a model can be fit to a universe.  
Example: Customer Churn Set  
- Historical data on customers who have stayed with the company and customer who have departed within six months of contract expiration  
- Task: Build a model to distinguish customers who are likely to churn based on some features  
- We test the model based on historical data and the model is 100% accurate.  
-We can *always* build a perfect model  
- A table model memorizes the training data and performs no generalization  
-Uselss in practice, previously unseen customers would all end up 0% likelihood

**Generalization** is the property of a model where the model is able to apply to data that were not used to build the model.

**Data Mining** needs to create models that generalize beyond training data.

**Overfitting:** tendency of a data mining modelto tailor models to the training data at the expense of generalization to previously unseen data points

All Data Mining procedures tend to be overfitting  
Tradeoff between model complexity and the possibility of overfitting.

**Holdout Data**

Evaluation on training data provides no assessment of how well the model generalizes to unseen cases.  
Idea: “Hold out” some data for which we know the value of the target variable, but which will not be used to build the model

*Predict* the values of the holdout data (test set) with the model and compare them with the hidden true values.

**Fitting Graph**

A fitting graph shows the accuracy of a model as a function of complexity

Generally, *the greater the complexity, the more accurate it fits to data provided for training,* causing overfitting.

**For fitting graphs:**

Once Error Rate reaches 0, the training set has been completely memorized

**Overfitting in tree induction:**

Tree induction – find important predictive individual attributes recursively to smaller and smaller data subsets  
 -Eventually these subsets will be pure, we have found the leaves of our decision tree  
 - The accuracy of this tree will be perfect.  
 -This is the same as the table model, ie an extreme example of overfitting.  
This tree should be slightly better than a lookup table  
Generally: A procedure that grows trees until the leaves are pure tends to be overfit.  
If allowed to grow without bound, decision trees can fit any data to arbirtrary precision  
The complexity of a tree lies in the number of nodes

**Addressing overfitting in trees:**

*Pruning:* simplifies a decision tree to prevent over-fitting to noise in the data  
*Post-pruning*: Takes a fully grown decision tree and discards unreliable parts  
*Pre-pruning*: Stopes growing a branch when information becomes unreliable.  
Post-pruning preferred in practice.

**Decision Boundaries**

Understand balance of cases that are classified or mis-classified when changing boundaries.

**Linear Classifier**

One straight line that effectively separates a data set.  
 - Robust  
 - One function that effectively separates customers

**Regression Analysis**

Predict a value of a given continuous variable based on the values of other variables, assuming a linear or non-linear model of dependency  
- Virtually endless applications:  
- Election outcomes  
-Future product revenues or commodity prices

**Linear Regression**

Usage:  
Two major categories where regression analysis can be leveraged  
1. To predict, estimate or forecast the cvalues: linear regression can be used to fit a predictive model to na observed data set of y and x values  
The model is fit to a set of known values using training data set and validated using a holdout “Test” sample  
Once the model is built it can be leveraged to predict the y values for the records where only X values are available.

2. To quantify the strength of the relationship between y and the Xj. To asses which Xj has a strong relationship or whther a particular Xj has a statistically significant relationship with a target variable  
The models fit to a set of known values using training data set and validated using a holdout test sample, then the relationship between the target variable y and predictors Xj is evaluated.

**Regression Analysis:**  R2 Value and Variable Selection

Goodness of fit in linear regression is generally measured using R2

R2 measure how well the regression line approximates the real data points, it also portrays percent of variance in the data explained by regression model.  
 - If the value is close to 1, the model fits perfectly o very well and explains all variance  
 - If the value is close to 0, then the model does not fit the data and doesn’t explain any variance

Variable Preparation:  
 - Interval variables can be binned or bucketed in order to capture non-linear relationship  
 - Categorical variables must be converted into binary vectors. Data sample must be large enough to accommodate all degrees of freedom.

**Regression Analysis: R2 and Variable Selection cont**

Variable selection: typically, the default significance levels (p values) are set at 0.05

Backward: Training begins with no candidate effects in the model and adds effects until the entry significance level or the stop criterion is met.

Forward: Training begins with no candidate effects in the model and adds effects until the entry significance level or the stop criterion is met

Stepwise: Training begins as in the forward model but may remove effects already in the model. This continues until the stay significance level or the stop criterion is met.

**Others Types of Regression Analysis**

**Quantile Regression**

* Ordinary least squares regression approximates the conditional mean of the response variable, while quantile regression is estimating either the conditional median or other quantiles of the response variable
* This is very helpful in case of skewed data (ie income distribution in the US)

**Logistic Regression**Logistic regression is used to predict categorical target variable  
Most often a variable with a binary outcome  
 Logit and Probit regressions can also be used to predict binary outcome.  
It is frequently used to estimate the probability of an event  
 Bank customer defaulting on a loan  
 Customer responding to a marketing promotion

**Logistic Regression – Sigmoid Curve**

**Non-Linear Functions**

Linear functions can actually represent non-linear models, if we include more complex features such as binning and bucketing and data preparation

**Linear Models versus Tree Induction**

What is more comprehensible to the stakeholders?

-Rules or a numeric function?  
 - How smooth is the underlying phenomenon being modeled?  
 -How “non-linear” is the underlying phenomenon being modeled?  
 -If very much, data engineering needed to apply linear models  
 - How much data do you have?  
 - Key tradeoff between the complexity that can be modeled and the amount of training data available

**Avoiding OverFitting**

Tree Induction:  
 -Post Pruning  
 -Pre pruning

Linear Models:  
 -Feature Selection  
 -Regularization: Optimize some combination of fit and simplicity.