**0 –Introduction to Supervised (aka Statistical) Learning**

Statistical Learning in general refers to a vast set of tools for understanding data. These tools can be classified as either supervised or unsupervised learning. In supervised statistical learning our goal is to develop statistical models to predict a response () using set of predictors ( and in unsupervised statistical learning we do not have a response () to predict but rather our goal is to discover interesting things about a set of measurements (). Unsupervised statistical learning will be the focus of DSCI 415 and in this course obviously are interested is in supervised statistical learning, i.e. building models to predict something.

In ***supervised statistical learning*** we build statistical models for predicting, or estimating, an output (response ) based on one or more inputs (predictors or features ). The response () can either be a numeric or categorical/ordinal variable. If the response is numeric we attempt to build a model that accurately predicts the value of the response (e.g. home price, future sales) using a set predictors. If the response is categorical/ordinal with *k* different levels (e.g. disease type, cancer grade, or whether a credit card transaction is fraudulent or not) then we can develop models using a set predictors that estimate the class an observation belongs to, or in the case of some modeling strategies estimate the probabilities an observation belongs to each of the *k* classes. In this course we will devote about 50% of our time on prediction problems where the response is numeric and about 50% on prediction problems where the response is categorical/ordinal.

Here are some examples of questions where supervised statistical learning could be employed to help answer them.

1. How does the mean selling price of a home relate to the physical features of a home like the # of bedrooms, # of bathrooms, square feet of living space, neighborhood home is located, whether the home has a fireplace or not, etc.? How accurately can we predict what a home will sell for given these characteristics? (Links: [www.zillow.com](http://www.zillow.com) and [www.redfin.com](http://www.redfin.com))
2. Given information about a potential loan applicant (e.g. age, home ownership, annual income, credit score, etc.) can we classify the applicant a good or bad credit risk?
3. Given a large set of genetic expression levels (these are generally numeric) can estimate the disease status of a patient? Can we estimate the probability they will be successfully treated using the current chemotherapy regimen used to treat that type cancer assuming they have it? Can we estimate the probability that someone will develop a certain disease in their lifetime given these genetic expression levels?
4. Given a set of hospital characteristics (e.g. number of orthopedic surgeries per year, # of inpatient beds, # of surgeons, geographic location, etc.) can we estimate the orthopedic sales potential for the hospital?
5. How are the shape and size measurements of breast tumor cells related to the malignancy? Given a set of these measurements can we accurately classify the malignancy of the breast tumor? Can we model the probability that a cell is malignant (or benign) based on these measurements? Can we use these measurements to estimate the probability of tumor recurrence for patients with malignant tumors?
6. Given information about an advertising campaign (e.g. amount spent on radio ads, TV ads, internet ads, length of campaign, etc.) can we estimate the sales impact in dollars?
7. Given characteristics of an e-mail (e.g. word count, capitalization, whether it contains certain keywords, etc.) can we accurately classify it as *spam*?
8. Given information about a credit card transaction (e.g. amount of charge, location of charge, type of purchase, etc.) can we accurately classify it as a legitimate or fraudulent charge?
9. Given a sample of image files can identify certain characteristics about them? For example, can we correctly identify a handwritten digit? How many fish and what species of fish are present in an image from a fixed camera on the deck of a fishing vessel? Can we classify a tumor as being malignant or benign from an image taken using a medical scope or an MRI scan?
10. Given a set of information about a chemical process can we accurately predict the chemical yield or some other numeric characteristic of the chemical such as purity or solubility?
11. If Target.com purchases a banner ad on Google for say HD TV’s can we estimate the probability that if someone clicks on it they will actually make a purchase? What is the estimated ROI (return on investment)?
12. If LL Bean mails a catalog to a potential customer can estimate the probability that they will actually make a purchase given customer information (e.g. age, income, previous purchase history, gender, etc.)? If they do make a purchase can estimate amount of their purchase using these same characteristics?
13. Given information about past medical claims by a group of individuals with Type II diabetes can we forecast/predict the medical claims filed by these same individuals in subsequent years? Can we identify the individuals we anticipate will have the largest claims?
14. Given a set of physical characteristics of a mushroom can we classify it as poisonous or edible?

As you can probably guess the possible applications of supervised statistical learning methods are infinite! Notice that the hypothetical examples above are a mixture of prediction problems where the responses is numeric and where the response is categorical/ordinal. Also notice the broad class is application areas (e.g. business, medicine, chemistry, etc.) where these methods can be used.

In Section 1 we will begin our discussion of supervised statistical learning methods when the response () is numeric. In Section 2 we will examine a common method used to predict/estimate a numeric response you should all be very familiar with - multiple linear regression (STAT 360).