Unit 4: Data Modeling – Part 1

Theory and Basics

# Overview of Data Modeling

We now turn to the Data Modeling stage of CRISP-DM. Modeling data may take many various forms, and it can often be confused with other terminology from the IT world since it so often involves data structures managed as a part of critical IT infrastructure. However, in the CRISP-DM framework we are specifically performing data modeling to build predictive models based on the data gathered in prior steps. Or, to put it another way: we *train* our model on existing data, *validate* the model with some hold-out sample data, and then *test* our results with new data.

Figure 1: CRISP-DM

What types of tools are used to model data? The list can be endless! The earliest data models were the simple linear regression. Linear regression was developed by ??? in ???. Linear regression gave rise to the similar logistic regression. You may have heard the phrase “regression to the mean”; this phrase harkens back to the original idea that given a sample of a population, eventually all populations would tend to revert back towards the mean of the entire population. Tall people would have children who would tend to be at least a bit shorter, and short people would tend to have children who were a bit taller, all things being equal, because the tendency of the population was to revert to the mean.

Another type of early data model was that of recognizing patterns and trends. A moving average is simply a measurement designed to identify an overall trend in a series of data over time.

Data models today can take many various forms, and we will explore more advanced techniques in Unit 5. For now, let’s go in-depth with the basics.

# Linear Regression

The first type of model we'll look at are linear models also known as linear regressions. In basic terms, we use a linear model if the data looks like a line could be drawn through it. More specifically, for two-dimensional data, we try to model the relationship between two variables, the independent or input variable, ***X***, and the dependent or response variable, ***y***, by an equation of the form:

where and are *coefficients*. We can generalize this equation to more than two dimensions. We can generalize this to more than two dimensions. If are independent variables, and is the dependent variable, we try to model the relationship by an equation in the form

where are the coefficients.

To find the coefficients of these equations, we'll rely on the [ordinary least squares](https://en.wikipedia.org/wiki/Ordinary_least_squares) method that attempts to minimize the sum of squares of the differences between the observed values and the predicted values - we'll explore this further in a bit.

An alternative formulation of this problem involves vectors and matrices. For each observation, that is for each pair of values of , , , we have the following system of equations:

We can write this in terms of vectors.

We can write this in terms of matrix multiplication. We also swap the left- and right-hand sides.

This is typically an overdetermined linear system of equations of the form

where is the matrix of ones and observed values of the independent variable, is the vector of unknown coefficients, and is the vector of obeserved values of the dependent variable. A variety of methods exist to find meaningful solutions to this linear equation. While this formulation is equivalent to ordinary least squares, it is commonly referred to as [linear least squares](https://en.wikipedia.org/wiki/Linear_least_squares_(mathematics)).

# Linear-Like Regression

Although linear regression works for many instances, we sometimes want to represent curves in our data via transposition to a linear relationship for simplifying our calculations. We can then transpose the regression line back to the original relationship to represent the regression line against the actual data. Relationships can be any type of non-linear relationship such as logarithmic relationships, hyperbolic curves, parabolic curves, or more complicated types. Sometimes it helps to model linear-like relationships against multiple curve types, and measure the accuracy of each before selecting a final type. In this course we will show an example during this unit’s exercises, but will not explore the topic in great depth.

# Logistic Regression

A [logistic regression](https://en.wikipedia.org/wiki/Logistic_regression) is used to model data where the dependent variable is categorical. In the simplest case, the dependent variable is binary and has only two possible values. A logistic model, provides an estimate of the probability that one of the two categories applies given the values of the independent variables; it fits a [logistic probability distribution](https://en.wikipedia.org/wiki/Logistic_distribution) to the data. We'll only look at simple case where the dependent variable is binary and there is only one independent variable.