ADS 503: Applied Predictive Modeling

Module 1 Presentation 1.1: Course Terminology & Data Pre-Processing Concepts

Hello class and welcome to the Module 1 Presentation of ADS-503, Applied Predictive Modeling.

In this module we are going to define predictive modeling and understand the difference between prediction & interpretation. We will introduce course terminology, discuss the need for preprocessing of data and the common methods used to preprocess data. Finally, we will introduce dimensionality reduction methods and focus on principal component analysis (PCA) method.

Predictive modeling has become a key feature of lot of applications that we use in our day to day life. Examples include movie recommendations by Netflix, product recommendations on Amazon, travel time estimation by google maps, and auto correct/sentence completion features in emails. Building a predictive model usually requires careful understanding of the data & modeling assumptions, more importantly it should not substitute but complement subject matter expert knowledge and intuition. Chapter 1 of your textbook discusses more about this topic.

The most important thing to understand before starting on the modeling process is the requirements of the business problem being addressed. In general, models that produce higher accuracy are harder to interpret, if the goal of the problem is interpretation then care should be taken to consider models that are interpretable. Section 1.1 of your textbook discusses more about this topic.

This slide covers some of the key terminology that will be used for the rest of the course. Section 1.3 of your textbook discusses more about this topic.

This slide covers some more terminology that will be used for the rest of the course. Section 1.3 of your textbook discusses more about this topic.

Predictive modeling process involves the following steps: splitting data into training, validation and or test data, choosing predictors also known as feature selection or feature extraction, estimating/evaluating model performance and finally model selection.

The “No Free Lunch” Theorem argues that, without having substantive information about the modeling problem, there is no single model that will always do better than any other model. Because of this, it’s always recommended to try a wide variety of techniques, then determine which model to focus on. Chapter 2 of your textbook discusses more about this topic.

Data pre-processing is a critical step in the predictive modeling process, it can significantly impact model performance, hence care should be taken to preprocess data according to the modeling needs. We will delve into these data pre-processing steps using a case study of cell segmentation in a high content screening data set. Section 3.1 of your textbook has more details about this case study.

Medical researchers often seek to understand the effects of medicines or diseases on the size, shape, development status, and number of cells in a living organism or plant. To do this, experts can examine the target serum or tissue under a microscope and manually assess the desired cell characteristics. This work is tedious and requires expert knowledge of the cell type and characteristics. Another way to measure the cell characteristics from these kinds of samples is by using high-content screening.

Briefly, a sample is first dyed with a substance that will bind to the desired characteristic of the cells. For example, if a researcher wants to quantify the size or shape of cell nuclei, then a stain can be applied to the sample that attaches to the cells’ DNA. The cells can be fixed in a substance that preserves the nature state of the cell. The sample is then interrogated by an instrument (such as a confocal microscope) where the dye deflects light and the detectors quantify the degree of scattering for that specific wavelength. If multiple characteristics of the cells are desired, then multiple dyes and multiple light frequencies can be used simultaneously. The light scattering measurements are then processed through imaging software to quantify the desired cell characteristics.

Using an automated, high-throughput approach to assess samples cell characteristics can sometimes produce misleading results. Hill et al. (2007) describe a research project that used high-content screening to measure several aspects of cells. They observed that the imaging software used to determine the location and shape of the cell had difficulty segmenting cells (i.e., defining cells boundaries). Consider Figure 3.1, which depicts several example cells from this study. In these images, the bright green boundaries identify the cell nucleus, while the blue boundaries define the cell perimeter. Clearly some cells are well segmented, while others are not. Cells that are poorly segmented appear to be damaged, when in reality they are not. If cell size, shape, and/or quantity are the endpoints of interest in a study, then it is important that the instrument and imaging software can correctly segment cells. Section 3.1 of your textbook has more details about this case study.

Centering and scaling is a transformation applied to individual predictors to make predictors have a common scale. Section 3.2 of your textbook discusses more about this topic.

A common method to make a skewed distribution symmetric is using logarithmic transformation if the predictor takes on only positive values. The impact of applying log transformation to one of the predictors, standard deviation of the intensity of pixels in actin filaments can be seen in the figure 3.2, after transformation the distribution is close to being symmetric. Section 3.2 of your textbook discusses more about this topic.

Box Cox transformations are a family of transformations indexed by the parameter lambda that make a skewed distribution symmetric if the predictor takes on only positive values. This family includes logarithmic, square, square root & inverse transformations when lambda takes on different values. Value of lambda is estimated using maximum likelihood estimation method based on the training data. Section 3.2 of your textbook discusses more about this topic.

Outliers in a data set should be handled with thought, they may convey important information about data in some cases and hence should not be deleted without thorough understanding of the data. Some predictive models are sensitive to outliers, spatial sign data transformation handles outliers by projecting the predictor values onto a multidimensional sphere. This has the effect of making all the samples the same distance from the center of the sphere.

Since the denominator is intended to measure the squared distance to the center of the predictor’s distribution, it is important to center and scale the predictor data prior to using this transformation.

Note that, unlike centering or scaling, this manipulation of the predictors transforms them as a group. Section 3.3 of your textbook discusses more about this topic.

Data reduction techniques are another class of predictor transformations. These methods reduce the data by generating a smaller set of predictors that seek to capture a majority of the information in the original variables. In this way, fewer variables can be used that provide reasonable fidelity to the original data. For most data reduction techniques, the new predictors are functions of the original predictors; therefore, all the original predictors are still needed to create the surrogate variables. This class of methods is often called signal extraction or feature extraction techniques.

PCA is a commonly used data reduction technique which seeks to find linear combinations of the predictors, known as principal components (PCs), which capture the most possible variance. Section 3.3 of your textbook discusses more about this topic.

Each principal component is a linear combination of predictors and is uncorrelated with other principal components. The component weights help us understand what predictors are most important in each principal component. The primary advantage of PCA, and the reason that it has retained its popularity as a data reduction method, is that it creates components that are uncorrelated. Section 3.3 of your textbook discusses more about this topic.

Looking at the correlation between average pixel intensity & entropy intensity from our case study data set, we could infer that average pixel intensity and entropy of intensity values measure redundant information about the cells and that either predictor or a linear combination of these predictors could be used in place of the original predictors. Section 3.3 of your textbook discusses more about this topic.

After extracting principal components the practitioner has to make a decision on how many principal components to include in the final model. There are couple of ways to make this decision, the first method is using scree plot & the second method is using cross validation which will be covered in more detail in later modules. Section 3.3 of your textbook discusses more about this topic.

This slide covers some of the useful preprocessing functions that are available to achieve the data transformations discussed in this module.