Data Mining and Text Mining

Rei Sanchez-Arias, Ph.D.
rsanchezarias@floridapoly.edu

Introduction

The dimension of a dataset, must be reduced for some of the data mining algorithms to operate efficiently. Approaches include:

- incorporating **domain knowledge** to remove or combine categories
- using **data summarie**s to detect information overlap between variables (and remove or combine redundant variables or categories)
- using **data conversion** techniques such as converting categorical variables into numerical variables
- employing automated reduction techniques, such as **principal components analysis** (PCA) where a new set of variables (which are weighted averages of the original variables) is created.

Curse of Dimensionality (1)

The dimensionality of a model is the number of predictors or input variables used by the model.

The **curse of dimensionality** is the affliction caused by adding variables to multivariate data models. As variables are added, the data space becomes increasingly sparse, and classification and prediction models fail because the available data are insufficient to provide a *useful model across so many variables*.

Curse of Dimensionality (2)

An important consideration is the fact that the difficulties posed by adding a variable increase exponentially with the addition of each variable.

One way to think of this intuitively is to consider the location of an object on a chessboard. It has two dimensions and 64 squares or choices. If you expand the chessboard to a cube, you increase the dimensions by 50%—from 2 dimensions to 3 dimensions. However, the location options increase by 800%, to 512 (8 × 8 × 8).

Curse of Dimensionality (3)

In statistical distance terms, the proliferation of variables means that nothing is close to anything else anymore – too much noise has been added and patterns and structure are no longer discernible. The problem is particularly acute in Big Data applications, including genomics, where, for example, an analysis might have to deal with values for thousands of different genes.

One of the, **key steps in data mining** therefore, is finding ways to **reduce dimensionality with minimal sacrifice of accuracy**.

In other fields, dimensionality reduction is often referred to as factor **selection** or feature **extraction**.

Practical Considerations

The integration of **expert knowledge** through a discussion with the data provider (or user) will lead to good results.

- Which variables are most important for the task at hand, and which are most likely to be useless?
- Which variables are likely to contain much error?
- Which variables will be available for measurement (and what will it cost to measure them) in the future if the analysis is repeated?
- Which variables can be measured before the outcome occurs?