Classification and KNN

This handout provides an overview of classification and K-Nearest Neighbors**.**

# Classification

Imagine that you are applying for a mortgage. You fill out a long application form with all your personal details, including income, age, qualification, location of the house, valuation of the house, and more. You are anxiously waiting for the bank’s decision on whether your loan application has been approved. The decision is either Yes or No. How does the bank decide? The bank reviews various parameters provided in your application form, and then—based on similar applications received previously and the experience the bank has had with those customers—the bank decides whether the loan should be approved or denied.

Now imagine that you have to make a decision about launching a new product in the market. You have to make a decision of Launch or No Launch. This depends on various parameters. Your decision will be based on similar experiences you had in the past launching products, based on numerous market parameters.

Next, imagine that you want to predict the opinions expressed by your customers on the products and services your business offers. Those opinions can be either positive or negative and can be predicted based on numerous parameters and past experiences of your customers with similar products and services.

Finally, imagine that airport authorities have to decide, based on a set of parameters, whether a particular flight of a particular airline at a particular gate is on time or delayed. This decision is based on previous flight details and many other parameters. Other examples of classification include

* Predicting whether a person will enjoy a movie or music recommendation
* Determining whether an ER patient is having a heart attack, based on their symptoms and vital signs
* Deciding whether an email is spam, given the presence of key words, images, hypertext, header information, and origin.

Each of these cases involves the prediction of a binary categorical outcome (enjoy/ not enjoy, heart attack/no heart attack, spam/not spam) from a set of predictors (also called *features*). The goal is to find an accurate method of classifying new cases into one of the two groups.

Classification is a two-step process. In the first step, a model is constructed by analyzing the database and the set of attributes that define the *class* variable. A classification problem is a supervised machine-learning problem. The training data is a sample from the database, and the class attribute is already known. In a classification problem, the class of Y, a categorical variable, is determined by a set of input variables {x1, x2, x3, …}. In classification, the variable we would like to predict is typically called class variable C, and it may have different values in the set {c1, c2, c3, …}. The observed or measured variables X1, X2, … X*n* are the attributes, or input variables, also called predictor variables or explanatory variables. In classification, we want to determine the relationship between the Class variable and the inputs, or predictor variables. Typically, models represent the classification rules or mathematical formulas. Once these rules are created by the learning model, this model can be used to predict the class of future data for which the class is unknown.

# K-Nearest Neighbors

The *k-nearest neighbors* (*K-NN*) classifier is based on learning numeric attributes in an *n*dimensional space. All of the training samples are stored in an *n*-dimensional space with a distinguished pattern. When a new sample is given, the K-NN classifier searches for the pattern spaces that are closest to the sample and accordingly labels the class in the k-pattern space

(called k-nearest neighbors). The “closeness” is defined in terms of Euclidean distance, where Euclidean distance between two points, X = (x1, x2, x3, … x*n*) and Y = (y1, y2, y3, … y*n*) is defined as follows:

This is a formula for computing Euclidean distance

The unknown sample is assigned the nearest class among the k-nearest neighbors pattern. The idea is to look for the records that are similar to, or “near,” the record to be classified in the training records that have values close to X = (x1, x2, x3, …). These records are grouped into classes based on the “closeness,” and the unknown sample will look for the class (defined by k) and identifies itself to that class that is nearest in the k-space.

A screenshot of a computer

Description automatically generated with medium confidence

## Figure 1 Example of KNN classification (source: Wikipedia)

Figure 1 shows a simple example of how K-NN works. The green dot should be classified either to blue squares or to red triangles. If k = 3 (solid line circle) it is assigned to the red triangles because there are 2 triangles and only 1 square inside the inner circle. If k = 5 (dashed line circle) it is assigned to the blue squares (3 squares vs. 2 triangles inside the outer circle).

K-nearest neighbors does not assume any relationship among the predictors (X) and class (Y). Instead, it draws the conclusion of class based on the similarity measures between predictors and records in the data set. Though there are many potential measures, K-NN uses Euclidean distance between the records to find the similarities to label the class. Please note that the predictor variables should be standardized to a common scale before computing the Euclidean distances and classifying.

After computing the distances between records, we need a rule to put these records into different classes (k). A higher value of k reduces the risk of overfitting due to noise in the training set. Ideally, we balance the value of k such that the misclassification error is minimized.

References:

Umesh R. Hodeghatta & Umesha Nayak(2017). Business Analytics Using R - A Practical Approach. Springer.

Kabacoff, Robert. (2015). R in action: data analysis and graphics with R. Manning Publications Co. <https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm>