Applied Analytics in Finance

Matilda Bosire

Strathmore University

September 7, 2023

Outline

Why not Linear Regression?

- The linear regression model assumes that the response variable Y is quantitative. However, in many situations, the response variable is qualitative.
- In the credit exercise, the probability of default is qualitative, as it categorizes debtors into default or non-default events (often, qualitative variables are referred to as categorical).
- Note that in a binary case it is not hard to show that even if we flip the coding (if $\hat{Y} > 0.5$, we have a default event and non-default otherwise), linear regression will produce the same final predictions.

Classification

- Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category or class.
- Widely-used classifiers include logistic regression, linear discriminant analysis, quadratic discriminant analysis, naive Bayes, and K-nearest neighbors.

Logistic Regression

- Models the probability that the response variable belongs to a particular category.
- For the Credit data, logistic regression models the probability of default, i.e.:

$$Pr(Default = Yes|age, savings.balance, checking.balance, past.debt, etc.),$$

- The probabilities range from 0 to 1, and one might predict a default event for any individual for whom $p(age, savings.balance, checking.balance, past.debt, etc.) > 0.5$.

Logistic Regression

- To avoid negative or very large probabilities for predictor values that are close to zero or very large respectively, we must model $p(X)$ using a function that gives outputs between 0 and 1 for all values of X. In logistic regression, we use the logistic function:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}},$$

  which translates to:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}.$$

- The quantity $\frac{p(X)}{1-p(X)}$ is called the *odds ratio*, and takes on values between 0 and infinity. Values close to 0 indicate very low probabilities, and vice versa.

Logistic Regression

- By taking the logarithm on both ends:

$$log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X,$$

the LHS is called the *log odds* or *logit,* that is linear in X, and can be generalized to p predictors as:

$$log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p.$$

- The Logistic regression is also known as a Generalized Linear model (`https://cran.r-project.org/web/packages/glm2/glm2.pdf`).
- MLE is preferred in parameter estimation, with p-values (< 0.05) being utilized to test for significance of predictor variables (reject the null hypothesis, $H_0 : \beta = 0$).
- A unit increase in a predictor is associated with an increase in the log odds of default by it's β coefficient.

K Means Clustering

- Non-parametric, un-supervised learning; measures the distance between different data points.
- Given a value for K, and a prediction point x_0, KNN first identifies the training observations closest to x_0, represented by N_0.
- It then estimates $f(x_0)$ (response), using the average of all training responses in N_0, i.e., the prediction in a region is the average of several points:

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

K Means Clustering

- The idea is to first specify the number of clusters, K, then the algorithm will will assign each observation to exactly one of the K-clusters.
- Clusters are non-overlapping: no observation belongs to more than one cluster.
- A good clustering technique is one where the **within-cluster variation** is as small as is possible.
- Generally, an optimal K-Means algo is such that the total within-cluster variation of all the independent clusters is as small as is possible.

K Means Clustering

- Denote the within-cluster variation for cluster, C_k as a measure, $W(C_k)$. Then, the optimization problem to be solved in K-Means is:

$$\underset{C_1,\ldots,C_k}{minimize} \sum_{k=1}^{K} W(C_k)$$

- To solve this in an actionable way, we use the **squared Euclidean distance**. That is:

$$W(C_k) = \frac{1}{|C_k} \sum_{i,i' \in C_k} \sum_{j=1}^{P} (x^{i,j} - xi',j)^2$$

where xi',j is the centroid, and $|C_k|$ represents the observations in the K^{th} cluster.

- The within-cluster variation is the sum of all pairwise squared **Euclidean Distances** between observations in a cluster, divided by the total number of observations in that cluster.

K Means Clustering

- The optimal value of K will depend on the **bias-variance trade-off**.
- A small value of K provides a flexible fit which will have low bias but high variance. The high variance is due to the fact that the prediction in a given region is entirely dependent on one observation.
- Larger values of K provide smoother and less variable fits, i.e., changing one observation has a smaller effect. However, the smoothing may cause some bias by masking some of the structure of $f(X)$, i.e, it runs the risk of ignoring small, but important patterns.

K Means Clustering

- The algorithm converges to a local rather than a global optimum, and results will depend on the initial (random) cluster assignment for each observation. It is important to run the algo multiple times from random different configurations, then select the solution as one with the lowest within-cluster variation.

K-Means vs KNN

- KNN is a supervised machine learning algorithm, while on the other hand, K-Means is an unsupervised machine learning algorithm.
- Nearest neighbor classifiers are defined by their characteristic of classifying unlabeled examples by assigning them the class of the most similar labeled examples.
- KNN analyzes the ′K′ nearest (labeled) data points and then classifies the new data based on the same. It selects the label of the new point as the one to which the majority of the ‘K’ nearest neighbors belong to.