Fundamental concepts

Fundamental concepts regarding ML

NULL SPACE

0.1 Regression basics

- The central idea behind regression is that we want to fit a model $\hat{y} = b_1 + b_2 x$ by selecting values for parameters b_1 and b_2 such that the resulting line is as close as possible to all the data values. This measure of closeness is linked to the idea of minimizing the **least squares criterion**.

- The residual is defined as: $e_i = y_i - \hat{y}$. The residual sum of squares (RSS) is defined as:

$$RSS = e_1^2 + \cdots + e_n^2 \tag{1}$$

- The central idea is to choose b_1 and b_2 to minimize RSS.

- When it comes to **multivariate regression** then for a predictor X_j we define β_j as the measure of association between that predictor and the response variable. β_j is interpreted as the average change in response Y with a unit increase in X_j keeping all other predictors fixed.

- **Residual plots** can help us identify whether the true relation might be nonlinear or not. We can plot the residuals vs. the predicted values. If the plot shows a discernible pattern then our data might be nonlinear.

- To ascertain if our error terms might be correlated or not we can plot the residuals as a function of time. If there is a pattern then they are.

- To check if our **homoscedasticity assumption** is violated we again look at residual plots against predicted y (down) and actual y (up) and see if a funnel shape emerges or not.

- **Multicollinearity problems** are assessed through a technique called the **variance inflation factor**. VIF is the ratio of variance of β_j when fitting the full model divided by the variance of β_j when fit on its own.

- The **dummy variable trap** is when dummy encoded independent variables are multicollinear.

0.2 Logit basics

- Rather than modeling the response Y directly, **logistic regression** models the probability that Y belongs to certain classes.
- To model a function that gives outputs between 0 and 1 we use the logistic function which is fit using the method of **maximum likelihood**. It is an S shaped curve.
- Interpretation: increasing X by one unit changes log odds by β_1.
- Shrinkage techniques like Ridge and Lasso are used to constrain the coefficient values in regressions. This tends to reduce the variance of our model.

0.3 DT

- Tree based methods revolve around stratifying the predictor space into regions. Classification or prediction of classes is based on the mean or mode of the classes of training observations to which the region where this observation belongs.
- Decision trees use a top down greedy approach to split the tree into regions. Split is made such that each point the error is minimum.
- Classification error rate is the fraction of training observations in a region that do not belong to the most common class.
- **Gini index** is a measure of total variance across the k classes in a particular region. Small values indicate higher node purity.
- Bagging and boosting are aggregation techniques that help us reduce the variance of a DT model.
- In **boosting** the trees are grown sequentially. Given a current tree, we fit another tree on top of that by using the current residuals rather than the outcome Y.

0.4 Regression accessories

- When testing for the presence of **heteroscedasticity** that is non constant error term variance we typically run regressions between the estimated error variance from a standard OLS and with the predictor variables. We then check if the coefficients of this new regression are statistically significant.
- While testing for **autocorrelation** presence we use the **Durbin Watson d-stat** which is nothing but the ratio of the sum of squared differences in successive estimated residuals to the RSS. Basically we check the lower and upper bounds of the computed ratio to see if autocorrelation is present or not.

0.5 Laws

- We state that after conducting repeated experiments, each time sampling independently and identically distributed random variables with finite mean and varince, we can obtain a distribution of the sample averages of the random variables of each repeated experiment. The WLLN states that the tails of this distribution tends to 0 as the sample size increases. The probability mass of the entire distribution falls on the value of the true parameter and we say that the sample mean converges in probability to the population mean under WLLN.

- CLT states that if we conduct an experiment repeatedly, each time taking n iid samples, then the standardized statistic of the sample mean follows a standard normal distribution as the sample size increases. Therefore the sample mean converges in distribution to the standard normal.