PREDICTIVE ANALYTICS - REGRESSION AND CLASSIFICATION

PROF. SOURISH DAS
Department of Mathematics
CMI

PRE-REQUISITES : At least one full course in Probability and Statistics

INTENDED AUDIENCE : MTech CS or MSc Statistics students and BTech final year students who have already taken at least one full course on Probability and Statistics

INDUSTRY SUPPORT : Bank and Financial Services, Manufacturing, Auto, ITES

COURSE OUTLINE :

A predictive model is an essential tool used daily in corporate practices. The course will provide an overview of fundamental ideas in statistical predictive models. The objective is to understand how statistical models handle prediction problems. The stress will be on understanding the construction of the models and implementation. It is a core course if students aspire to be Data Scientists.

ABOUT INSTRUCTOR :

Prof. Sourish Das is an Professor at Chennai Mathematical Institute (CMI). He looks after the MSc Data Science program at CMI. Prior to CMI, he worked in the industry for three years. He did his PhD in Statistics from the University of Connecticut, USA and did his postdoctoral work at Duke University, UK. In 2018, he won the Rutherford fellowship of UK and visited the University of Southampton.

COURSE PLAN :

Week 1:

● Introduction: Introduce the course and the objective of the course. I will provide a broad overview of the course and landscape of the predictive models. In our experience, we see that predictive analytics is crucial in the industry and business and part of the daily life of a data scientist in the corporate sector.

● Least Squares method: We introduce the concept of simple and multiple linear regression as regression hyper-plane. Then we discuss the concept of least square methods.

Week 2:

● Normal Equations: We discuss the results that the normal equations will always have at least one solution. If the system is of full rank, then the least square method will give us analytically solvable unique solutions. Also, we discuss why ""mean absolute deviation"" does not have an analytical solution.

● Gauss Markov theorem: Here, we introduce the underlying assumptions of linear regression models. We discuss how the Gauss-Markov theorem is developed under the assumptions of the homogeneity and independence of the residuals. We also discuss the concept of mean squared error (MSE). How the MSE and prediction accuracy are related?

Week 3:

● The geometry of Regression Model and Feature Engineering Here, we discuss some examples. Also, we discuss the ""basis expansion"" in mathematical statistics, known as ""feature engineering"" in ML. With feature engineering, we put the original data into a higher dimension. We hope we find a good fit for a linear hyperplane in a higher dimension, explaining the non-linear relationship between the feature space and the target variable.

● Statistical Inference of Regression Coefficient: Here, we discuss the sampling distribution of regression coefficients and statistical inference of regression coefficient. We discuss how the t-test can be conducted to test whether a predictor/feature has a statistically significant effect on the dependent or target variables.

Week 4:

● Checking Model Assumptions We discuss how to check the model assumptions like (1) Independence, (2) Homogeneity or (3) Normality using statistical tests. We discuss how Bartlet's rank test can be used to check randomness in residuals. How can we apply the Breusch-Pagan Test to check homogeneity? Then we discuss how the Kolmogorov-Smirnov test can be used to check the normality!

● Model Comparison with R-squared, RMSE, AIC or BIC We discuss, how we compare two or several models using selection criteria, such as RMSE, R-squared, adjusted-R-Squared, AIC or BIC and select the best model among the set of possible models.

Week 5:

● Model Complexity and Bias-Variance tradeoff In this part, we tried to understand the issues of model complexities. We discuss the concept of what we mean by a complex model. Sometimes complex models help us to achieve high prediction accuracy. However, we often achieve it as a cost of the interpretability of the model. We try to capture the model complexity concerning the concept of bias-variance tradeoff. Finally, we discuss how we can achieve a parsimonious model by minimising MSE.

● Feature selection and Dimension Reduction Here, we discuss the stepwise feature selection or variable selection algorithms. We discuss how the feature selection technique can reduce the problem's dimension and achieve an interpretable and parsimonious regression model.

Week 6:

● Multicollinearity and Variance Inflation Factor Here, we discuss the problem of multicollinearity in the regression problem. We discuss how the correlation between strong or more features induces a strong correlation between the OLS estimators of coefficients. This makes the sampling distribution a very strong elliptical shape due to tight correlation. As a result, the standard error increases significantly, and the confidence interval becomes very large. We discuss how the Variance Inflation Factor (VIF) can be used to measure the contribution of each feature towards the multicollinearity problem.

- Regularization with LASSO, Ridge and Elastic Net In the second part of Lecture 6, we discuss what is ""Ill-posed problems"". We discuss how Tikhonov Regularization reconstructs the unconstrained minimization of the OLS method into constraint minimization. We discuss how the L2 penalty corresponds to the Ridge solution and the L1 penalty corresponds to LASSO solutions. We discuss why LASSO is a continuous subset selection and one should use LASSO feature selection over stepwise feature selection.
- Ridge Regression with Python

Week 7:

- Regression Analysis with Python
- Regression Analysis with R
- Regression Analysis with Julia

Week 8: Major Applications of Regression Models

- Capital Asset Pricing Model Here, we present a very celebrated application of the statistical regression model in quantitative finance, known as the Capital Asset Pricing Model (CAPM). The CAPM is often used to evaluate whether an asset is overpriced, underpriced, or fairly priced.
- Bootstrap Regression Here, we discuss the concept of Bootstrap statistics and nonparametric bootstrap regression. We discuss the two algorithms of Bootstrap regression: (1) Residual Bootstrap Regression and (2) Paired Bootstrap Regression
- Time Series Forecasting with Regression Model Here, we focus on how the regression model technique can be implemented for long-term and short-term forecasting. We develop a long-term forecasting model by modelling the trend and seasonality as a function of time. On the other hand, we develop the short-term forecasting model using the Auto-regressive model. We used the AirPassengers dataset.
- Granger Causal model. Here, I introduce an excellent application of the regression model, the Granger Causal model. Correlation does not imply Causation. In practice establishing Causation from the data is very difficult. However, the Granger causal regression model tries to answer the question of causality with limited capability.

Week 9:

- Logistic Regression In this video, we introduce the concept of logistic regression for binary class classification problems.
- MLE of coefficient of Logistic Regression Here, we discuss how to estimate the regression coefficient of logistic regression.

Week 10:

- Fit Logistic Regression with optim function in R
- Fit Logistic Regression with glm function in R
- Fit Logistic Regression with sklearn in Python
- Fit Logistic Regression in Julia

Week 11:

- Logistic Regression and Inference In this lecture, we discuss how we can make statistical inferences for the logistic regression model.
- Discriminant Analysis In this lecture, we discuss the concept of discriminant analysis, LDA, and QDA.

Week 12:

- Multinomial Logit Regression
- Generalised Linear Regression
- Poisson Regression
- Negative Binomial Regression