Apply Machine Learning to Performance Trend Analysis

Araya Eamrurksiri

Data preprocessing and data transformation had been performed. Some fields in the dataset store multiple values and these values are separated by a tab character. These tab-separated values are split to columns in order to use for later analysis. Moreover, each software product consists of several software packages and numerous test cases are executed in one software package. Test case which has a minimum value of CPU utilization is selected to represent the performance of specific software package.

Markov switching model is implemented for addressing the thesis' problem - we are interested to discover change points and regime shift in time series when the time instant is unknown. This model is one of the most well-known non linear time series models. It is first introduced by Hamilton [1] and is extensively implemented in economics and finance field. It takes the behavior of shifting regime in time series into account and models multiple structures that can explain this characteristic in different states at different time. The shift between state or regime comes from the switching mechanism which is assumed to follow an unobserved Markov chain. Thus, the model is able to capture more complex dynamic patterns, identify the switch in states when change-point is most likely to occur. In speech recognition, such processes are described as hidden Markov model [2].

Given the behavior of the observation sequence in this study, it is noticed that observed value is not completely independent of each other (i.e., performance of current software package depends on the performance from the past version of software package). Therefore, additional dependencies at observation level with the first order autoregression, AR(1), is taken into consideration when fitting the model. It is simply called Markov switching autoregressive model.

The MSwM [1] package in CRAN developed by Josep A. Sanchez-Espigares is used for performing an univariate autoregressive Markov switching model for linear and generalized model. The package implemented expectation-maximization (EM) algorithm to estimate the Markov switching model. Source code and functions used for fitting the model in this package have been studied and reviewed in detail. The purpose of this task is to get a general idea of how the package works and also to understand the algorithm and concept behind. Even though most of the coding has been done for performing the Markov switching model,

[1] https://cran.r-project.org/web/packages/MSwM/index.html

further implementation is still needed in order to properly deal with the problem at hand. Some modifications have been made in the function to handle errors and warnings produced when fitting the model. For instance, when setting variance to have a non-switching effect, warning is generated. It turns out that there is a minor mistake in the code. This issue is now resolved. Furthermore, in some situations when fitting linear regression, Hessian will not be invertible because the matrix is singularity. Consequently, standard error of the estimated coefficients can not be computed. This noninvertible Hessian is solved by using generalized inverse procedure or pseudoinverse [3]. Next, an extension function for handling with categorical predictor variables is implemented in the package. Another error that has been dealt with is when coefficients from fitting the model are NAs. This is mostly happened when predictor variables are categorical. First, the function initializes coefficients in each state by randomly dividing data into different subsets and separately performing regression in each subset. A problem arises when variable in subset does not contain all levels of variable or has only a single value. As a result, the model generates NA coefficient for that particular variable. This issue is now being taken care of.

Results from fitting Markov switching model in the package are described here. The model returned both switching and non-switching estimated parameters in each state. For each observation, states assignment and probability assignment in each state are obtained from the model. The package includes a function to plot periods where the observation is in the specific state and also a function to create several plots for the residual analysis. It shows a plot of residuals against fitted values, a Normal Q-Q plot, and ACF/PACF of residuals.

References

[1] James D Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pages 357–384, 1989.

[2] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[3] Jeff Gill and Gary King. What to do when your hessian is not invertible: Alternatives to model respecification in nonlinear estimation. *Sociological methods & research*, 33(1):54–87, 2004.