ECE 792-41 Statistical Foundations for Signal Processing and Machine Learning

Project 1: Data Analytics and Machine Learning

The project should be completed individually. You must use the IEEE Transactions template (Word or LaTex) for writing your report. Your submission must be a concise write-up of the results and findings. It should include figures and tables, and the descriptions and discussions of them. There is no need to include detailed background reviews that mainly repeat the techniques taught in lectures. You are allowed to use off-the-shelf packages.

You should sign the **Honor Pledge** at the beginning of the report: "*I pledge in my honor that I have not given or received any unauthorized assistance on this report*".

Road Safety Analysis and Prediction

In this project, you are given a road accident dataset that contains such variables as the number of accidents (safety), route type, segment type, segment length, the number of lanes, annual average daily traffic (AADT), congestion level. You have two related goals, namely, i) to conduct a statistical analysis on the dataset to reveal the major predictors for the number of accidents, and ii) to design a prediction system that can predict the number of accidents from a vector of input variables.

The dataset contains entries for five segment types that have already been separated into subdatasets. Each subdataset may contain a distinct subset of variables. You will initially work on the analysis and the prediction system design for each subdataset. If time permits, you can choose to work on the complete dataset for bonus points. Your main focus should be on segment types 0, 1, and 2.

Variables in the excel file are coded in different colors by a domain expert. A green column means the variable is a known transportation indicator. A red column means that the traffic engineer sees no relationship to safety. White columns are to be explored. You may want to put more respect to such domain knowledge in the analysis task, and be more open when designing the prediction system as far as the generalization principle of machine learning (i.e., never design a learning system using testing data) is strictly followed.

Note that a few variables in the dataset are categorical/qualitative, including route type, segment type. To correctly use such an L-level variable as a predictor/feature, you must code it into $(L-1)$ binary variables. Please refer to pages 85-86 of ISLR book for the detailed explanation and how the coding can be done. There are also some other variables that look like categorical but are ordinal/numerical, including #Lanes:OFR, Weave.Segment.LCFR.

The preferred programming language is R for Task 2 (statistical analysis) and Python for Task 3 (machine learning), but you are free to use other programming languages. Learn how you can start effective coding in a new programming language within two hours.

1. Data Cleaning and Variable Coding

There are some missing data for variable safety that was filled by 49.99, 49.999, or alike. Keep track of how many such records (manually or programmatically) you removed, and report their percentage.

Examine the records in which the safety value is coded to 0. Be prepared to remove them from the dataset if they are missing data per your observation, or be prepared to keep them in the dataset if they are truly 0 (which means that no accident was recorded). There are also some entries with safety values equal to 0.4 or 0.6. Investigate whether they are filler values or true values. Keep them in or remove them from the dataset per your judgment, and justify your decision. You may reassess your decision based on new understandings obtained from Tasks 2 and 3.

2. Statistical Analysis

In this task, you should conduct a regression-based task to reveal how variable safety is related to other variables. Using off-the-shelf statistical analysis packages such as those in R is recommended.

a) Examine the distribution of the variable safety. If necessary, transform the variable using a link function to create a new response variable. Justify your choice.

b) Examine every scatter plot of any pair of variables in the dataset. Record those highly correlated (collinear) pairs.

c) Run a regression analysis against various variables in the dataset. You need to decide which ones to include and not to include. You may want to gradually add more predictors to regression. If doing a pure linear regression, avoid putting in collinear variables into the regression problem simultaneously. You may also want to try penalizing the weights in a certain way. Note that in this case, centering and normalizing each independent variable may be needed. Based on your results, which predictors are more useful to explain the variable safety?

3. Prediction System Design

In this task, you will build a prediction system for variable safety using a vector of variables of your choice. The generalization principle of machine learning must be stringently followed: Training and testing must be clearly separated and performance must be reported for both training and testing. You can use any model for designing the prediction system, including (generalized) linear model, regression tree, support vector machine, and neural network. Make sure that the categorical variables are coded correctly when they are used as input.

Glossary:

AADT	Annual Average Daily Traffic
MM	Mile Marker