MATH561 Final Project Description

Background and Motivation

Climate model simulation can produce extremely large output data on hundreds of climate variables relevant to scientists. For example, the Community Earth System Model - Large Ensemble (CESM-LE) climate model created by the National Center for Atmospheric Research is estimated to produce 10 petabytes of climate model data to participate in the Coupled Model Intercomparison Project Phase 6 (CMIP6), a centerpiece of the Intergovernmental Panel on Climate Change. Storing that much data is extremely expensive, and traditional lossless compression techniques are not very effective on floating-point climate (or other) data. Consequently, scientists are interested in applying lossy compression, a form of compression where some of the original information is lost in exchange for much higher reductions in file size. Care must be taken to prevent compression from affecting the results of scientific analysis. Optimal compression for a particular dataset means that the data size has been reduced by as much as possible without affecting scientific conclusions. Compression is applied to every time step of every variable in each simulation, which means the optimal compression level must be selected millions of times. Ideally, selection of the optimal compression level would be performed in an automated fashion based on certain features of the data.

In this project, you will be given climate model datasets and the corresponding optimal compression levels. These optimal compression levels were found by performing a brute-force search over many candidate compression levels. Performing this exhaustive search for every time slice is computationally infeasible; scientists need a way to predict the optimal compression level without checking every possibility. Your job is to identify features of the data that may be useful in predicting the optimal compression level, and use these features as predictors in classification models to predict the optimal compression level. The goal is to maximize the classification accuracy of your model.

Data

The data is located in Files -> Final Project. The three subfolders High, Medium, and Low correspond to the optimal compression level for the datasets within. Inside each folder is a set of data to train your model on, and a validation set to test the accuracy of each model. These data files contain ten time steps from each of a curated selection of CESM-LE variables output at 6-hourly, daily and monthly frequencies. Also included in the Final Project folder is a sample R script that shows how to load the data files, gives detailed information on the structure of the data files, uses the data to create basic plots and features, and finally puts features into a data frame object for use in classification models.

Some details about the format of each data file: Each data object is structured as a list containing two variables, mat and var. The data consists of climate model output at ten separate time slices for multiple climate variables. The mat variable is a list where each element is the dataset for a single variable at a single time slice, stored as a matrix. The elements in the var variable are the variable names, frequencies, and a time step identifier corresponding to each dataset. The ith element in the var variable corresponds to the name and output frequency of the ith element in the mat variable.

Finally there is a folder with your test data (270 datasets). These data have no classification labels and as part of your deliverables, you will submit your predicted compression levels for these test data.

Requirements and Deliverables

As a first step for this project your goal is to construct and identify features that are effective in predicting the optimal compression level of the dataset. These may include simple features such as the range of the dataset, or more sophisticated features you develop that describe the sparsity or smoothness characteristics of the dataset.

You are required to develop at least three different classification models that use your features to predict the optimal compression level of each dataset. These can include any of the classification methods discussed in class (logistic regression, regression trees, LDA, QDA, SVMs, KNN, etc.) or methods not discussed in class (e.g. neural networks). At least two of your three models have to use a method we covered in class.

Your deliverables are

1. A Final Project Report (see below for more details on what it needs to contain)
2. Predictions for the test data as an R data frame with column names “var” for the variable name and “pred” for the predicted class. Each row will correspond to the prediction for a single time step of a variable.
3. Your well-documented code to run your models and reproduce all your results.

Your Final Project Report should at a minimum contain:

- Description of the project and its goals
- Description of the data and findings from your exploratory analysis
- Description, rational and diagnostic of the models you select
- Evaluation of model performance on the validation datasets
- Discussion of results for the test data
- Summary and overall conclusions

You can also find an example of a final project report in the Final Project folder on Canvas.

Enjoy practicing your skills by engineering features, and exploring and developing models!