DATA6811 - Computational Inference for Machine Learning

Assignment 3

Due: Sunday 29 May 2022, 23:59AEST

1 General Description

This assignment covers evaluates knowledge in the topics of Dimension Reduction, Model Selection and Spatial Temporal Regression. It consists of programming and technical exercises which are closely related to the topics covered in the respective tutorials. The expected completion time is 6 hours.

1.1 Deliverables

Submission will be through Microsoft Teams. Please complete the python notebooks / R markdown files for each section and upload them to the Teams assignment environment in the ”your work” section.

Marking criteria involves a total of 100 points, with 60% assigned to the first section of this assignment (each section sums to 100, and then they are rescaled according to the percentage weight). A penalty of MINUS 5 points per each day after the due date will be applied.

2 Dimension Reduction and Spatial Temporal Modelling (60%)

The country of Mexico contains a vast range of landscapes, with deserts, alpine forests, swamps, and tropical rainforests. Much of this diversity is tied to the spatial variation of climate over this region, both due to the large latitudinal range spanned by Mexico and the presence of mountains through the middle of the country.

You have been provided with a netCDF file named CentralAmericaPrecAnomalyGHCN.nc containing the monthly rainfall anomaly, i.e., the deviation from the expected climate for each month, over land in the Central American region.

1. Plot a map and time series of the 1st Empirical Orthogonal Function (EOF) and Principal Component (PC) of the rainfall anomaly over Mexico. [30 marks]

2. What fraction of the total variance is explained by your first EOF? [30 marks]

3. One common use of EOFs is to guide further analysis of high-dimensional data. Based on your result, which region of Mexico experiences the least variation in rainfall? Can you guess which part of the country experiences the most extreme rainfall events? [40 marks]

   (It’s enough to answer this in broad terms such as ”northeast quadrant” or ”west coast”. Note that the results of your EOF analysis alone are not enough to make a scientific claim about the data - but give us a great starting point from where to decide how to proceed with an analysis or model design).

3 Model Selection and Averaging (40%)

We want to fit a multiple linear regression model to a dataset of $n = 97$ observations and carry out variable selection using Lasso. We find that $\lambda_{\max} = 0.9$ is a value of the shrinkage parameter λ that all the coefficients

λ	0	0.1	0.2	0.3	0.4	0.5	.6	.7	.8	.9
df_λ	8	5	3	3	2	1	1	1	1	0
$\widehat{\sigma}^2_\lambda$	0.9033	0.8116	1.8865	4.1080	4.6396	5.1962	5.7991	6.4485	7.1444	7.4611

are shrunk to zero. To find the optimal shrinkage, we create a range of 10 values for $\lambda : 0, 0.1, 0.2, ..., 0.9$, and compute the Lasso estimates $\widehat{\beta}^{\text{lasso}}_\lambda$ at each of these values. Let $\widehat{\sigma}^2_\lambda = \frac{\|\boldsymbol{y}-\boldsymbol{X}\widehat{\beta}^{\text{lasso}}_\lambda\|^2}{n}$ be the estimate of the variance of the error term ϵ. The table gives the degrees of freedom df_λ and $\widehat{\sigma}^2_\lambda$

(a) Find the best value of λ among these 10 values using the BIC criterion.

(b) Explain why variable selection is often necessary in regression and classification.