Advanced Regression Assignment 1

2024-05-02

Instructions

1. Please hand in both an R markdown (or Quarto) file and a compiled pdf (only) document version of your R markdown file on Vula. Your compiled document should not contain any code. If you need to describe an algorithm this should be done in text or in the form of pseudo-code.
2. Your report should contain enough information, description of methods, interpretation of results and conclusions to explain clearly what you have done, without having to refer to any code. This means that you need to specify all settings used, even if these are default settings. A good guideline is to give enough detail to allow another person to replicate what you have done exactly, even if they want to replicate this using a different software package.
3. Please attach a plagiarism statement to your hand-in. Your code and write-up need to be entirely your own work, even though you are allowed to discuss the work with others. Please also submit your document to Turnitin. Reference all sources of information and code.
4. The assignment is due: 22 July 2024

Question 1: Splines

There are many ways to construct a basis for cubic regression splines. We have only looked at the truncated power basis and B-splines.

Wood (2006 and 2017) describes two additional bases. The first one is described below. The 2nd is described in detail in his book (Wood 2017 Section 5.1.3) and is the basis used for cubic regression splines in the mgcv package (`s(x, bs = "cr")`).

The former is defined in Wood 2006 p. 124-127. He just calls this 'a cubic spline basis'. The basis functions are:

$b_1(x) = 1, b_2(x) = x, b_{i+2}(x) = R(x, x_i*), \qquad \text{for } i = 1, \ldots, q-2$

where x_i^* are knot locations: {x_i^*: i = 1, …, q-2}.

There is more detail on this type of basis in the books by Gu (2002) and Wahba (1990). I haven't been able to find these books.

One significant advantage of this basis is that the penalty matrix can be found as:

$$S_{i+2,j+2} = R(x_i^*, x_j^*), \qquad i, j = 1, \ldots, q-2$$

$$\begin{aligned} R(x,z) = &\left((z-0.5)^2 - 1/12\right)\left((x-0.5)^2 - 1/12\right)/4 \\ &- \left((|x-z|-0.5)^4 - 0.5(|x-z|-0.5)^2 + 7/240\right)/24 \end{aligned}$$

For this assignment, use the above basis to fit a cubic regression spline to the following scenario. Choose evenly spaced knot locations.

$Y_i \sim N(\mu_i, 0.5^2), \qquad \mu_i = 5 + \sin(3\pi(X_i - 0.6))$

Simulate data $n = 100$ from the above model. Simulate $X_i \sim N(0, 1)$.

a. Construct a basis with 6 basis functions, on the range $x \in [0, 1]$, and illustrate these.

b. Fit a penalized regression spline to the simulated data points, using the above basis with 30 knots.

c. Obtain confidence bands for the fitted spline.

d. Show GCV as a function of the smoothing parameter.

Outline your methodology, illustrate your results, and discuss briefly.

Question 2: Two-Dimensional Splines

The data are from brain imaging by functional magnetic resonance scanning (Landau et al. 2003). The data are available in the data frame `brain`, in package `gamair`.

Each row of the data frame corresponds to one voxel. One slice of the image is provided, described as a near-axial slice through the dorsal cerebral cortex.

The variables are:

X: voxel position on horizontal axis.

Y: voxel position on vertical axis.

medFPQ: median of three replicate 'Fundamental Power Quotient' values at the voxel: this is the main measurement of brain activity.

region: code indicating which of several regions of the brain the voxel belongs to. The regions are defined by the experimenters. 0 is the base region; 1 is the region of interest; 2 is the region activated by the experimental stimulus; NA denotes a voxel with no allocation.

meanTheta: mean phase shift at the Voxel, over three measurements.

For this assignment consider models for medFPQ as a function of X and Y only.

1. Use a tensor product spline. Use the type of basis functions from Question 1, starting with 10 basis functions for each dimension (X and Y). Then penalize.
2. Construct a thin-plate spline basis with 100 evenly spaced knots. Fit to the brain data, again penalizing for overfitting.

Wood 2006 Ch. 5 pg. 230 did a similar analysis using functions from the `mgcv` package. You should fit the models from first principles (e.g. penalized likelihood), not using `mgcv`.

Describe and illustrate your methods and results, and briefly discuss. Possibly compare to Wood's results or to results using `mgcv` functions.

Question 3: GAMs and MARS

Pulsars are a rare type of neutron star that produce radio emissions detectable on Earth. They are of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter. The data can be found in `pulsar.csv`.

Data and Information Sources:

https://www.kaggle.com/datasets/colearninglounge/predicting-pulsar-starintermediate

https://www.kaggle.com/pavanraj159/predicting-a-pulsar-star https://www.kaggle.com/datasets/brsdincer/pulsar-classification-for-class-prediction?resource=download

https://archive.ics.uci.edu/dataset/372/htru2

Each candidate star is described by eight continuous variables, and a single class variable. The first four are summary statistics obtained from the integrated pulse profile (folded profile). This is an array of continuous variables that describe a longitude-resolved version of the signal that has been averaged in both time and frequency. The remaining four variables are similarly obtained from the DM-SNR curve: Radio waves emitted from pulsars reach earth after traveling long distances in space which is filled with free electrons. Pulsars emit a wide range of frequencies, and the amount by which the electrons slow down the wave depends on the frequency. Waves with higher frequency are slowed down less as compared to waves with higher frequency.

Use GAMs and MARS for this classification task. As benchmark, also fit a logistic regression model (`glm`). Compare these three approaches.

Outline your methods, illustrate and interpret your results, briefly discuss.

Question 4: Wavelets

Hastie et al. (2008) use functional logistic regression models to distinguish between two sounds: "aa" and "ao". For this assignment use wavelets for this task.

I am not sure how well this is going to work, but just see what you can do. I.e., the task is to find areas of the spectrum that are important for predicting which sound was spoken.

The phonemes data can be found here:

- package `fds`
- Elements of Statistical Learning Data: https://hastie.su.domains/ElemStatLearn/

For each observation / case, transform the signal into 'features' using a discrete wavelet transform. Limit to just the first 256 wavelengths to simplify the problem. I.e., transform the x_{ij} into wavelet coefficients using $\mathbf{W}\theta$ where $\mathbf{W}$ is a wavelet basis matrix.

Then use this in a logistic regression model.

It may be necessary to threshold the coefficients at the first stage (when creating the wavelet features), or sufficient to penalize during the logistic regression model stage. You should explore these alternatives.

The `glmnet` function in package `glmnet` is able to fit penalized logistic regression. You are allowed to use this function or any other function that implements penalized logistic regression.

Describe your methodology, illustrate and interpret your results and give conclusions.

Question 5: Functional Data Analysis

The octane rating of fuel determines its 'knocking' resistance. The higher the octane rating the higher the compression ratio that an engine can run at. Traditionally, octane measurement involves comparing the knocking resistance of fuel samples to standard mixtures in special variable compression ratio engines. This is an expensive process relative to obtaining the near-infra-red spectrum of a sample. It would be good to be able to predict octane rating from the spectrum.

Use the octance rating data in the R package `gamair` (`gas`). The data set contains the NIR spectra and octane numbers of 60 petrol samples. The NIR spectra were measured using diffuse reflectance as log(1/R) from 900 nm to 1700 nm in 2 nm intervals, giving 401 wavelengths.

Fit a scalar-on-function regression model to model the octane rating as a function of the NIR spectrum from the petrol sample.

For this question you should not use any specialised R functions for functional data analysis (except for R's `lm` function, which you can use).

Describe your methodology, illustrate and interpret your results and give conclusions.

References

1. Hastie, Tibshirani, Friedman, J. H. 2008. The Elements of Statistical Learning. Data Mining, Inference and Prediction. 2nd Edition. Springer.
2. Kalivas, J. H. (1997). Two data sets of near infrared spectra. Chemometrics and Intelligent Laboratory Systems, 37(2), 255–259.
3. Landau S. et al (2003) 'Tests for a difference in timing of physiological response between two brain regions measured by using functional magnetic resonance imaging'. Journal of the Royal Statistical Society, Series C, Applied Statistics, 53(1):63-82
4. Wood, S.N. (2006). Generalized Additive Models. An Introduction with R. Chapman & Hall / CRC.
5. Wood, S.N. (2017). Generalized Additive Models: An Introduction with R, Second Edition (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279