STAT 4750: PROJECT

April 15, 2024

Deadline: May 1, 2024

Instructions:

1. There are three questions in this project. The first question focuses on regression, the second is about classification, and the last one is about causal inference.
2. Show and explain all your work (even for conclusions). Partial credit cannot be given otherwise.
3. Points will be deducted for incorrect work even if the final answer is correct.
4. Attach your code at the end of the project report as an appendix. Note that code cannot be a replacement for explanations or comments.

Question	Points	Your Score
1	100	
2	80	
2	20	
Total	200	

Question 1

We aim to develop a statistical model to predict the progression of Parkinson's disease, quantified by the Unified Parkinson's Disease Rating Scale (UPDRS). Trained medical professionals typically assess the UPDRS in a clinical setting. Specifically, the motor UPDRS score, which measures motor impairment, is an important aspect of this scale. To simplify this process, which is time-consuming and expensive, efforts are being made to enable patients to compute their motor UPDRS score at home, increasing convenience and reducing costs.

Tsanas et al. (2010) advanced this field by creating an approach that uses noninvasive speech tests for at-home assessment, achieving clinically useful accuracy. The data from this study is available in the Parkinson's Telemonitoring dataset at the UCI Machine Learning repository (`https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring`). We have adapted this dataset and stored it in `park.csv`, with a detailed description of the variables provided in Table 1. We treat *motor UPDRS score* as the response and use a variety of regression models to predict the motor UPDRS score of patients using six biomedical voice measures.

Variable	Description
`motor_updrs`	modified clinician's motor UPDRS score
`Abs, PPQ5`	Measures of variation in fundamental frequency
`dB, APQ11`	Measures of variation in amplitude
`NHR`	A measure of ratio of noise to tonal components in the voice
`RPDE`	A nonlinear dynamical complexity measure

Table 1: Description of the response and predictors in `park.csv` file. The response is in the first row and the remaining rows describe the biomedical voice measures.

Answer the following questions based on different types of regression models introduced in the class. Evaluate their predictive performance using mean square prediction error (MSPE) estimates based on 10-fold cross-validation.

1. (10 pts) How does a linear regression model perform when predicting motor UPDRS scores using the given biomedical voice measures? Evaluate the model's predictive performance.

2. (20 pts) Extend the previous model to a polynomial regression of degree 2. Extend it further using the lasso and ridge penalties and principal components regression. Compare their predictive performance with the previous model.

3. (10 pts) How does using the polynomial kernels of degrees 1, 2, 10, and 20 in kernel regression affect the prediction accuracy of motor UPDRS scores? Does the polynomial kernel beat the performance of the previous models for some degree? Justify the similarity between polynomial regressions used in the previous questions and kernel regressions used in this question.

4. (10 pts) How does the performance change if we use the radial basis function and Laplace (or exponential) kernels in the previous quesiton?

5. (10 pts) Investigate the utility of random Fourier feature expansion in improving the prediction of motor UPDRS scores for the polynomial feature maps of order 2 and 5, respectively. Vary the random feature dimension as 6, 12, and 24. How do the random features compare to previous methods?

6. (10 pts) Fit a two-layered (shallow) neural network for predicting motor UPDRS scores. Is its performance better than that of the previous methods? Justify the choice of tuning parameters such as number of hidden units, pre-activation function, learning rate, and epochs.

7. (10 pts) Compare the predictive performance when a deep neural network replaces the shallow neural networks.

8. (10 pts) Assess and compare the MSPEs of all the regression models from (i) to (vii). Comment on this flexibility and their dependence on the tuning parameters. For example, does the ranks of MSPEs match the rankings of the flexibility?

9. (10 pts) Perform a sensitivity analysis on the tuning parameters, such as degree of the polynomial, regularization parameter, random feature dimension, learning rate, number of layers, and number of epochs. How do changes in these parameters affect the model's performance? Which model balances sensitivity to the choice of tuning parameter and predictive accuracy?

Question 2

This question involves creating a statistical model that can differentiate between authentic and counterfeit banknotes. We utilize the Banknote Authentication dataset from the UCI Machine Learning Repository, provided by Volker Lohweg at the University of Applied Sciences, Ostwestfalen-Lippe. Accessible at (`https://archive.ics.uci.edu/ml/datasets/banknote+authentication`), the dataset comprises 1372 samples derived from images of real and fake banknotes. These samples are contained in the `banknote.csv` file. Each sample is described by five attributes: one binary response variable and four predictors. The response variable is 1 for genuine notes and 0 for forgeries. The predictors, extracted using wavelet transform, include measurements of image variance, skewness, kurtosis, and entropy. Table 2 provides a detailed description of the variables in the `banknote` data.

Variable	Description
`class`	Response taking two values: 0 for a forged banknote and 1 for a genuine banknote
`variance`	Variance of wavelet transformed banknote image
`skewness`	Skewness of wavelet transformed banknote image
`curtosis`	Curtosis of wavelet transformed banknote image
`entropy`	Entropy of the banknote image

Table 2: Description of the response and predictors in `banknote.csv` file. The response is in the first row and the remaining rows describe the four predictors, which are continuous and can take any real values.

Answer the following questions based on classification models. Evaluate their predictive performance using accuracy, TPR, and FPR estimates based on 10-fold cross-validation.

1. (10 pts) How does logistic regression and the perceptron perform in classifying banknotes as genuine or counterfeit based on the four given image-derived features? Summarize predictive performance.

2. (10 pts) Assess the predictive performance improvement of logistic regression and perceptron with polynomial features of degree 2.

3. (10 pts) Compare the predictive performance of the previous models with SVMs with polynomial kernels of degrees 1, 2, 10, and 20. Does SVM perform better than the previous approaches?

4. (10 pts) Implement a kernel approximation using random Fourier features for banknote classification using logistic regression with degree 3 polynomial features. Vary the random feature dimension as 6, 12, and 24. How does this approach compare with the SVM in terms of predictive performance?

5. (10 pts) Repeat 6. in Question 1 for classifying banknotes as genuine or counterfeit.

6. (10 pts) Repeat 7. in Question 1 for classifying banknotes as genuine or counterfeit.

7. (20 pts) Repeat 9. in Question 1 for classifying banknotes as genuine or counterfeit.

Question 3

The question is about causal inference with a continuous response y, 0-1 valued treatment t, and confounders $\mathbf{x} = (x_0, x_1, x_2, x_3)$. The sample size $n = 2000$ and training data $(y_i, t_i, \mathbf{x}_i)$ $(i = 1, \ldots, n)$ are in files `resp.csv`, `trt.csv`, and `conf.csv`.

1. (10 pts) Estimate the average treatment effect (ATE) using propensity scores. Employ logistic regression to estimate the propensity scores necessary for computing the ATE.

2. (10 pts) Instead of logistic regression, use a deep neural network to estimate the propensity scores, and then compute the ATE. Compare the ATE estimate with the previous ATE estimate.

References

1. A Tsanas, MA Little, PE McSharry, LO Ramig (2010), 'Accurate telemonitoring of Parkinson.s disease progression by non-invasive speech tests', IEEE Transactions on Biomedical Engineering.