CSIC5011 Mini-project 2: Prediction by Supervised Learning Derived from Unsupervised Learning

Xu siao(sxuao@connect.ust.hk) ID:20377629 Li Juncheng(jlicv@connect.ust.hk) ID:20377124
Department of Mathematics, HKUST

1. Introduction

Supervised and unsupervised learning are two different branches in machine learning. Sometimes, one and other interconnects. Over this small project, we explore and compare prediction methods induced by each of them.

2. Problem

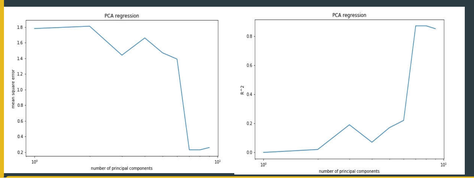
We try to predict the dangerous level for different animals with a number of learning methods by extracting data of animals sleeping features.

In nature, as a reasonable guess, predators sleep longer than prey, bigger animals sleep longer than smaller ones since the formers are unlikely to be attacked by the other animals in natural environment. Moreover, sleeping feature for predators in top food-chain are different from that for preys in lower food-chain. All of which above can be ascribed to dangerous level, each of which is explained by a sequence of dream features specific to a kind of creature. If clustering of features is likely to occur, extraction of the principal components of explanatory variables may improve prediction results. The animal sleeping Data is taken from one of the datasets provided in the project 2, with p =10 and n=62.

3. Supervised PCA

$$Y = \beta X + e$$
$$X_y = \mu + \Gamma \upsilon_y + \delta\varepsilon$$
$$\Leftrightarrow Y = \beta \Gamma^T (X_y - \mu) + \tau$$
$$\text{Where } \upsilon_y \propto Y \text{ and } \tau \propto \delta\varepsilon$$

We invoke cross-validation method with data being divided into five sets, any four of which combined as train set and left one as test set. We use PCA for dimension reduction on dataset and linearly regress Y on principal components kept back. From left-bottom graph, we can see that MSE of PCA regression is horrible for little PC kept back but very nice after we keep 7 PC onwards. From right-bottom graph, correspondingly R^2 remains low for small number of PC kept but skyrocket after we keep 7 PC.



4. Supervised PCA+

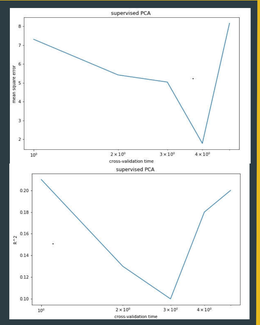
$$s_j = x_j^T y / \|x_j\|$$

Keep back all columns of X such that $s_j > \theta$

$$X_\theta = U_\theta D_\theta V_\theta^T$$
$$U_\theta = X_\theta V_\theta D_\theta^T = X_\theta W_\theta$$

Let $U_\theta = (u_{\theta,1}, u_{\theta,2}, \dots, u_{\theta,m})$ and $W_\theta = (w_{\theta,1}, w_{\theta,2}, \dots, w_{\theta,m})$

$$\hat{y}^{spc} = \bar{y} + \hat{\alpha} u_{\theta,1} = \bar{y} + \hat{\alpha} X_\theta w_{\theta,1} = \bar{y} + X_\theta \beta_\theta \text{ where } \beta_\theta = \hat{\alpha} w_{\theta,1}$$



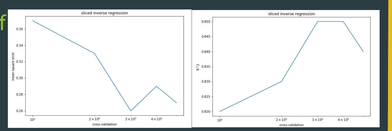
We predict y with β_θ calibrated by data in the training set, and calculate corresponding MSE and R^2. The right-top graph shows MSE for five-fold cross-validation, where the MSE is the lowest at fourth fold. In right-bottom graph, we see that in first and fifth time of cross-validation, R^2 are the highest but the mean of five cross-validation performs not ideal. Perhaps, the data for each dimension locate so close to each other and PCA analysis is not so suitable to this set of data. We appeal to other more suitable method.

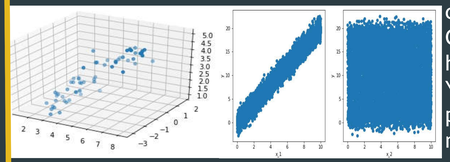
5. Sliced Inverse Regression

In SIR, we begin with keeping top-2 principal components. Estimated R.S.E. and R-square are shown at right side, where five-fold cross-validation is performed. From right-left graph, we can see the mean of five MSE is very low. From right-right graph, we see the lowest R^2 of five cross-validation is 0.82 and highest of that is 0.85. Obviously, the performance of SIR is very nice but the complexity should be in consideration because of heavy computation. Following, we try some simple regression model and see if we can attain our goal of exact prediction at very low cost.

Table 1 Procedures of SIR

1. Standardize **x**
2. Divide range of y into H slices, $I_1, \dots, I_H$; let the proportion of the y_i that falls in slice h be $\hat{p}_h$. That is, $\hat{p}_h = \frac{1}{n}\sum_{i=1}^{n} \delta_h(y_i)$, where $\delta_h(y_i)$ is the indicator function of whether y_i falls into the h^{th} slice or not.
3. Within each slice, compute the sample mean of the $\mathbf{x}_i$'s, $\hat{\mathbf{m}}_h$.
4. Conduct a (weighted) PCA for the data $\hat{\mathbf{m}}_h$ in the following way: Form the weighted covariance matrix $\hat{V} = \sum_{h=1}^{H} \hat{p}_h \hat{\mathbf{m}}_h \hat{\mathbf{m}}_h^T$, then find the eigenvalues and the eigenvectors for $\hat{V}$.
5. Let the K largest eigenvectors be $\hat{\eta}_k$. Output $\hat{\beta}_k = \sum_{\mathbf{xx}}^{-1/2} \hat{\eta}_k$.





Graph 1

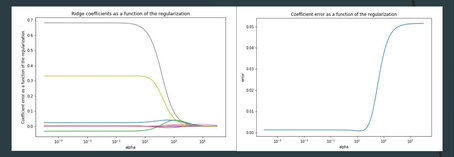
Graph 2

Graph 1 shows two PCs kept corresponding to Y in 3-D coordinate.
Graph 2 shows that the first PC has nice linear relationship with Y but the second PC doesn't perform in the same way, meaning the first PC is most useful and more PC kept is unlikely to improve the performance.

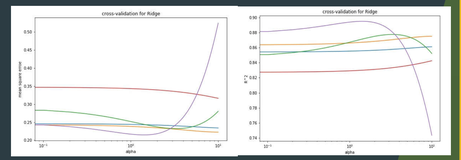
6. Ridge Regression

$$minimize\ \{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})\text{^}2\}$$
$$subject\ to\ \sum_{j=1}^{p}\beta_j^2 \le s$$

We quantify the result by using mean square error and R^2 as benchmark of which the results are as graphical below. The left-bottom graph compare the ridge coefficients and normal linear regression coefficients for controlling different alpha. We see the difference between ridge and normal regression converges to zero as alpha gets larger and larger. The right-bottom graph shows with-in difference in coefficients among different cross-validation for different alpha.



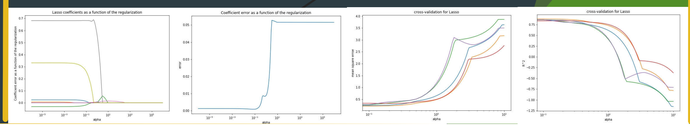
The left-bottom shows MSE for five cross-validation while controlling alpha. We MSE remains steady for different alpha. The right-bottom graph show R^2 for five cross-validation while controlling alpha. We see R^2 are all high for five cross-validation, each of which is higher than 0.82 and the highest attains 0.88.



7. LASSO

$$minimize\ \{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})\text{^}2\}$$
$$subject\ to\ \sum_{j=1}^{p}|\beta_j| \le s$$

LASSO is similar to ridge and what the graph shows below is similar to graph above. We see the performance of ridge and LASSO are both nice and more importantly, the model of ridge and LASSO is so simple and implementation of which is so cheap.



8. Conclusion and Discussion

The result of prediction shows that the dangerous level is correlated with some animals' sleeping feature. But there are some drawbacks as well such that the methods I used right now are all something related to shrinkage, meaning that we make the prediction as exactly as possible at big costs of model unexplainable. Since we shrink down or even delete some predictable variables Which have very big covariance with each other in order to calibrate the prediction as exactly as possible. In one word, we can exactly predict what will happen but we don't know why something will happen. In our project, we can predict the dangerous level of animals but can not find out illuminable relationship between sleeping feature and danger level, leaving the model unexplainable.

9. Contribution

Xu and Li evenly
Contribute to such everything as coding, math modeling, report writing and data preprocessing.

10. Reference

(1) Sliced Inverse Regression for Dimension Reduction
Ker-Chau Li
Journal of the American Statistical Association
Vol. 86, No. 414 (Jun., 1991), pp. 316-327
(2) Fisher Lecture: Dimension Reduction in Regression1, 2 R. Dennis Cook
(3) Regression shrinkage and regression on LASSO R TIBSHIRANI - 1996
(4) Prediction by Supervised Principal Components, Eric Bair, Trevor Hastie, Debashis Paul & Robert Tibshirani