CSIS5011 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

As we all know, supervised and unsupervised learning are two different branches in machine learning , both of them and something else consist in recently the most exciting area in AI. But, sometimes, there's something underlying which give rise to connection from one to another and vice versa. Over this small project, we explore some prediction methods induced by unsupervised and check out the usefulness corresponding to specific supervised method.

2. Problem

In nature, as a reasonable guess, predators sleep longer than prey, bigger animals sleep longer than smaller animals 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. Over this project, by extracting data of animals sleeping feature, we try to explain and predict the dangerous level for different animals and, meanwhile, make use of and compare different supervised learning method, some of which, of course, are by-product of exploring unsupervised learning .

5. Supervised PCA

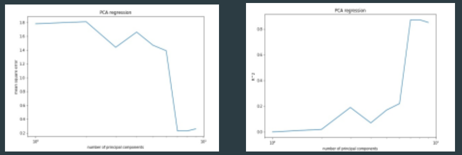
$$Y = \beta X + e$$

$$X_y = \mu + \Gamma \upsilon_y + \delta\varepsilon$$

$$\Leftrightarrow Y = \beta\, \Gamma^T (X_y - \mu) + \tau$$

Where $\upsilon_y \propto Y$ and $\tau \propto \delta\varepsilon$

With such same method as cross-validation, We use PCA for dimension reduction on dataset and linearly regress Y on principal components kept back. we can see that the performance of PCA regression is horrible for little PC kept back but very nice after we keep 7 PC onwards.



Supervised PCA

10. reference

Sliced Inverse Regression for Dimension Reduction
Ker-Chau Li
Journal of the American Statistical Association
Vol. 86, No. 414 (Jun., 1991), pp. 316-327
Fisher Lecture: Dimension Reduction in Regression1, 2 R. Dennis Cook
Regression shrinkage and regression on LASSO R TIBSHIRANI - 1996

3. Ridge Regression

$$minimize\,\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^{\wedge}2\}$$

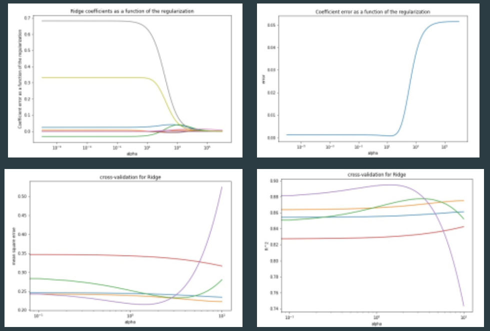
$$subject\ to\ \sum_{j=1}^{p}\beta_j^2 \le s$$

4. LASSO

$$minimize\,\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^{\wedge}2\}$$

$$subject\ to\ \sum_{j=1}^{p}|\beta_j| \le s$$

We invoke cross-validation method such that divide dataset into five subsets, randomly make any four subsets as train sets and leave residual one as test set. We repeat the process five times following the same strategy and do cross-validation to verify each other. We quantify the result by using mean square error and R^2 as benchmark of which the results are as graphical below, we can see MSE remains tiny and steady for both ridge and LASSO at very beginning and the MSE for LASSO skyrocket after α hits 1. MSE for ridge is very steady over entire experiment except one set. R^2 is another side of MSE, we can see R^2 for ridge remains above 0.82 almost but that for LASSO is higher than 0.75 at very beginning and plummet multistep. Overall, the performance for ridge is very nice since the closer the R^2 is to 1, the better the performance of the model. Meantime, performance for LASSO is also nice but a bit worse than that of ridge.



ridge

6. Supervised PCA Plus

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

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

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

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

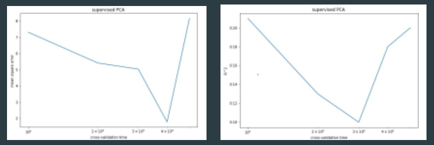
Let $U_\theta = (u_{\theta,1}, u_{\theta,2}, \dots, u_{\theta,m})$ *and*

$$W_\theta = (w_{\theta,1}, w_{\theta,2}, \dots, w_{\theta,m})$$

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

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

In this advanced supervised PCA version, we use train set to calibrate β_θ and use β_θ to predict y, after which we have y and y-test to calculate MSE and R^2. the result is as graphical which is not so ideal.

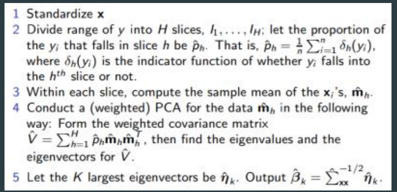


Supervised PCA+

9. Contribution

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

7. Sliced Inverse Regression



1 Standardize $\mathbf{x}$

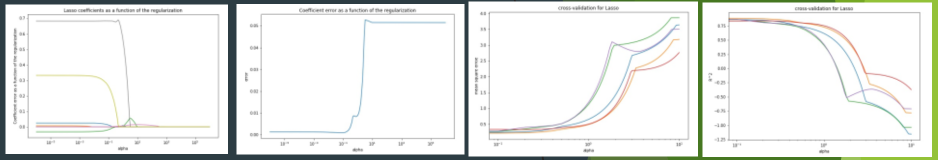
2 Divide range of y into H slices, $I_1, \ldots, I_H$; let the proportion of the y_i that falls in slice h be $\hat{p}_h$. That is, $\hat{p}_h = \frac{1}{n}\sum_{i=1}^{n}\delta_h(y_i)$, where $\delta_h(y_i)$ is the indicator function of whether y_i falls into the h^{th} slice or not.

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 $\hat{\beta}_k = \hat{\Sigma}_{\mathbf{xx}}^{-1/2}\hat{\eta}_k$.

In SIR, we keep two principal components and from the right graph from which the first PC has high correlation with respondent variable Relative to low correlation between second PC and y. anyway, the performance of SIR is nice since ,we can see R^2 attains 0.85 and meanwhile, MSE around 0.28 just when we keep tree or four PC.



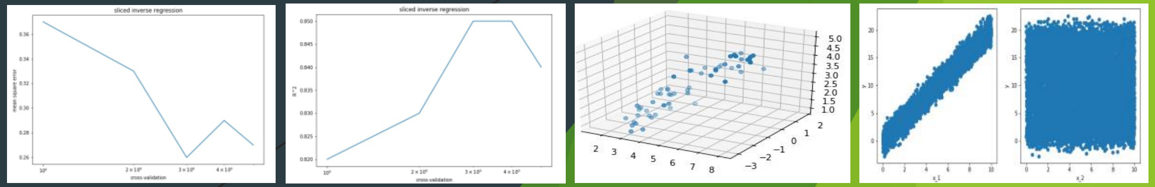
LASSO

8. 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.

Sliced inverse regression


sliced inverse regression
sliced inverse regression
cross-validation
cross-validation