Weighted L_1 Penalized Logistic Regression with Principal Components

Wenjun Lang

University of Minnesota

langx239@umn.edu

December 6, 2017

Principal Components

Principal Components

- An orthogonal transformation of input matrix X.
- The first few principal components explain major amount of variation of X.

Classification with Principal Components

- Achieve dimension reduction.
- Get better prediction accuracy.
- E.g.: eigenface.

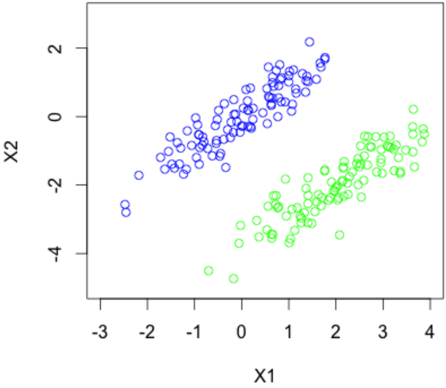
Classification with Principal Components

Eigenface

First row: the leftmost is the average face and the others are top two eigenfaces.
Second row: eigenfaces with least three eigenvalues.

Select First k Principal Components?

- May not work well if y is strongly correlated with last few principal components.



k =1, poor classification
k= 2, no dimension reduction

Better Selection of PC?

We want to achieve...

- PCs with higher variances should be more important than PCs with lower variances.
- Should also consider the association between X and Y.

Weighted L_1 Penalized Logistic Regression with Principal Components

- $X = UDV^T$,
- U: normalized principal components of X.
- D: diagonal matrix of singular values of X.

Weighted L_1 Penalized Logistic Regression in PC space

$$\hat{\gamma} = \arg\min_{\gamma} \; -\ell(\gamma) + \lambda ||D^{-1}\gamma||_1$$

$$\ell(\gamma) = \arg\min_{\gamma} \; -\frac{1}{n}\sum_{i=1}^{n}(y_i U_i^T \gamma - log(1 + exp(U_i^T \gamma)))$$

- $D = diag(d_1, d_2, \cdots, d_p)$. Larger d_i, larger variance of the PC.

$$\hat{\gamma} = \arg\min_{\gamma} \; -\ell(\gamma) + \lambda \sum_{i=1}^{n} \left|\frac{\gamma_i}{d_i}\right|$$

- Smaller penalty on PCs with higher variances.

Coefficients for X:

$$\hat{\beta} = (DV^T)^{-1}\hat{\gamma}$$

Optimization Problem

$$\underset{\gamma}{Minimize} \; -\frac{1}{n}\sum_{i=1}^{n}(y_i U_i^T \gamma - log(1 + exp(U_i^T \gamma))) + \lambda ||D^{-1}\gamma||_1$$

- U: Normalized Principal Components of X.
- D: Diagonal matrix of singular values of X.

MM Algorithm

Majorization Step:

$$Q(\gamma|\hat{\gamma}^{old}) = -\frac{1}{n}[\ell(\hat{\gamma}^{old}) + \nabla\ell(\hat{\gamma}^{old})^T(\gamma - \hat{\gamma}^{old}) - \frac{1}{2}(\frac{1}{4} + 10^{-6})(\gamma - \hat{\gamma}^{old})^T {U^*}^T U^*(\gamma - \hat{\gamma}^{old})] + \lambda \sum_{j=2}^{p+1} \frac{1}{d_j}|\gamma_j|$$

where

$$U^* = (\mathbb{1}_n, U)$$

MM Algorithm

Note that

$$U^{*T} U^{*} = \begin{bmatrix} n & & & \cdots & \\ & 1 & & \cdots & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ & & & \cdots & 1 \end{bmatrix}$$

MM Algorithm

$$Q(\gamma|\hat{\gamma}^{old}) = -\frac{1}{n}\ell(\hat{\gamma}^{old}) - \frac{1}{n}\sum_{j=1}^{p+1}[\nabla_j \ell(\hat{\gamma}^{old}))(\gamma_j - \hat{\gamma}_j^{old})$$
$$- \frac{1}{2}(\frac{1}{4} + 10^{-6})||u_j||^2(\gamma_j - \hat{\gamma}_j^{old})^2] + \lambda \sum_{j=2}^{p+1} \frac{1}{d_j}|\gamma_j|$$

Minimization Step:

$$\hat{\gamma}_j^{new} = S(\hat{\gamma}_j^{old} + \frac{\nabla_j \ell(\hat{\gamma}^{old})}{(\frac{1}{4} + 10^{-6})||u_j||^2}, \frac{n\lambda}{(\frac{1}{4} + 10^{-6})||u_j||^2 d_j})$$

Numerical Experiments

Data

- MNIST Data
- Left to right: first, second, third principal component

- Left to right: twenty-fifth, hundredth, five hundredth principal component

Numerical Results

Methods

- Logistic regression with conventional PCA.
- Penalized logistic regression in the space of principal components
- LDA

Approach	LR with PCA	PLR with PC	LDA
Prediction Error(%)	7.04	6.58	8.13