Closed-Form Supervised Dimensionality Reduction with Logistic Regression

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider supervised dimensionality reduction with logistic regression when labels and features of data are all binary. This model combines learning dimensionality reduction model with learning regression and classification model for the purpose of reducing classification error. Closed-form update rules of the parameters are also derived in detail. This model is useful in statistical genetics where the features are usually binary or multinomial with some binary or multinomial labels. We apply the proposed method to single-nucleotide polymorphism data to classify samples of different races. From the comparison of the results, the supervised dimensionality reduction outperforms unsupervised dimensionality reduction followed by logistic regression.

1 Introduction

Dimensionality reduction is important for many statistical learning tasks, especially when the dimension of the data is much larger than the number of samples. In statistical genetics, the dimension of features (e.g., the allele statuses of millions of locations) is usually huge, but the number of samples (i.e., the number of DNA sequenced) is generally limited since DNA sequencing is expensive. Dimensionality reduction is recommended or even necessary in such cases.

In unsupervised learning, the objective of dimensionality reduction is to find a low-dimensional representation which preserves most important properties of data. Principal component analysis (PCA) is a popular method for unsupervised dimensionality reduction (UDR) for the advantage of computational efficiency. However, interpretation of PCA has always been a difficult task in many problems. Recently, PCA attracts the attention of researchers in genetics again [1] with freely available data from the International HapMap Project (HapMap) and the Human Genome Diversity Panel (HGDP). McVean [2] showed that PCA is not only useful, but also interpretable in genealogy.

In supervised learning, the information of responses can be combined with the information of features to choose the low-dimensional representation in the dimensionality reduction, which is shown useful when the objective is classification [3]. This is of particular interest in genetics. If ethnical labels of participants are given, we can use those informations to assist dimensionality reduction for the features, which are allele statuses in millions of locations. The application of supervised dimensionality reduction (SDR) in genetics is still a new topic.

For parameter estimations, the likelihood function may not be convex for all parameters involved in supervised dimensionality reduction (SDR). Closed form update rules with alternate minimization procedure are proposed for unsupervised dimensionality reduction when features are binary [4], and it is claimed that it can be generalized to supervised dimensionality reduction when both responses and features are exponentially distributed without details [3].

In this project, we consider supervised dimensionality reduction when responses and features are all binary. We derive closed form update rules of parameters with alternate minimization procedures. We apply this SDR method (logistic-SDR) to a classification problem in genetics, and show that misclassification error can be reduced comparing to UDR followed by logistic regression (logistic-UDR).

2 Supervised dimensionality reduction with logistic regression (logistic-SDR)

Let $\mathbf{X}$ be an $N \times D$ data matrix and X_{nd} be the (n, d)-th entry of $\mathbf{X}$, where N is the number of identical and independent distributed (i.i.d.) samples and D is the number of binary features ($n = 1, 2, \cdots, N$ and $d = 1, 2, \cdots, D$). Let $\mathbf{X}_n$ denote the n-th row of $\mathbf{X}$.

Let $\mathbf{Y}$ be an $N \times K$ data matrix and Y_{nk} be the (n, k)-th entry of $\mathbf{Y}$, where D is the number of binary labels ($n = 1, 2, \cdots, N$ and $k = 1, 2, \cdots, K$). Let $\mathbf{Y}_n$ denote the n-th column of $\mathbf{Y}$.

We assume that X_{nd} is a Bernoulli random variable with parameter p_{nd} and density

$$P_{X_{nd}}(x_{nd}|p_{nd}) = p_{nd}^{x_{nd}}(1-p_{nd})^{1-x_{nd}} = \sigma(\theta_{X_{nd}})^{x_{nd}}\sigma(-\theta_{X_{nd}})^{1-x_{nd}},$$

where $\theta_{X_{nd}} = \log\{p_{nd}/(1-p_{nd})\}$ is the natural parameter of X_{nd}, and $\sigma(x) = 1/\{1+\exp(-x)\}$ is the sigmoid function.

Similarly, we assume that Y_{nk} is a Bernoulli random variable with parameter q_{nk} and density

$$P_{Y_{nk}}(x_{nk}|q_{nk}) = q_{nk}^{y_{nk}}(1-q_{nk})^{1-y_{nk}} = \sigma(\theta_{Y_{nk}})^{y_{nk}}\sigma(-\theta_{Y_{nk}})^{1-y_{nk}},$$

where $\theta_{Y_{nk}} = \log\{q_{nk}/(1-q_{nk})\}$ is the natural parameter of Y_{nk}.

We assume that the parameter matrix $\boldsymbol{\theta}_X = (\theta_{X_{nd}})$ can be represented by a linear model in an L-dimensional $(L < D)$ space:

$$\theta_{X_{nd}} = \sum_{l=1}^{L} U_{nl}V_{ld} + \Delta_{X_d},$$

where the rows of the $L \times D$ matrix $\mathbf{V}$ denote the basis vectors of the low-dimensional space, the columns of the $N \times L$ matrix $\mathbf{U}$ denote the coordinates of $\theta_{X_{nd}}$, and Δ_{X_d} denote the bias. For simplicity, we include $(\Delta_{X_1}, \Delta_{X_2}, \cdots, \Delta_{X_D})$ as the $(L+1)$-th row of $\mathbf{V}$, and include a column of all 1's as the $(L+1)$-th column of $\mathbf{U}$, i.e.,

$$\theta_{X_{nd}} = \sum_{l=1}^{L+1} U_{nl}V_{ld}, \text{ or } \boldsymbol{\theta}_X = \mathbf{UV}.$$

We consider the n-th row of $\mathbf{U}$, $\mathbf{U}_n$, as a low-dimensional representation of the corresponding $\mathbf{X}_n$, and use it to predict the labels $\mathbf{Y}_n$. We assume that the parameter matrix $\boldsymbol{\theta}_Y = (\theta_{Y_{nk}})$ can be represented by a linear model in an L-dimensional space:

$$\theta_{Y_{nk}} = \sum_{l=1}^{L} U_{nl}W_{ld} + \Delta_{Y_d},$$

where the rows of the $L \times D$ matrix $\mathbf{W}$ denote the basis vectors of the low-dimensional space, the columns of the $N \times L$ matrix $\mathbf{U}$ denote the coordinates of $\theta_{Y_{nk}}$, and Δ_{Y_d} denote the bias. Similarly, for simplicity, we include $(\Delta_{Y_1}, \Delta_{Y_2}, \cdots, \Delta_{Y_K})$ as the $(L+1)$-th row of $\mathbf{W}$, and include a column of all 1's as the $(L+1)$-th column of $\mathbf{U}$, i.e.,

$$\theta_{Y_{nk}} = \sum_{l=1}^{L+1} U_{nl}W_{lk}, \text{ or } \boldsymbol{\theta}_Y = \mathbf{UW},$$

Therefore, the log-likelihood for the matrix of features $\mathbf{X}$, is given by

$$\mathcal{L}_{\mathbf{X}}(\boldsymbol{\theta}_X) = \sum_{n,d}[x_{nd}\log\{\sigma(\theta_{X_{nd}})\} + (1-x_{nd})\log\{\sigma(-\theta_{X_{nd}})\}],$$

and the log-likelihood function for the matrix of labels $\mathbf{Y}$, is given by

$$\mathcal{L}_{\mathbf{Y}}(\boldsymbol{\theta}_Y) = \sum_{n,k}[y_{nk}\log\{\sigma(\theta_{Y_{nk}})\} + (1-y_{nd})\log\{\sigma(-\theta_{Y_{nk}})\}].$$

For unsupervised dimensionality reduction, it is shown that $\mathbf{U}$ and $\mathbf{V}$ can be calculated from $\mathcal{L}_{\mathbf{X}}(\mathbf{U},\mathbf{V})$ [4]. For supervised dimensionality reduction, we can write the joint likelihood for both $\mathbf{X}$ and $\mathbf{Y}$ [3]:

$$\mathcal{L}_{\mathbf{X},\mathbf{Y}}(\mathbf{U},\mathbf{V},\mathbf{W}) = \alpha\mathcal{L}_{\mathbf{X}}(\mathbf{U},\mathbf{V}) + \mathcal{L}_{\mathbf{Y}}(\mathbf{U},\mathbf{W}),$$

where α is a parameter to weight the influence of two likelihoods.

The logistic-SDR problem is to find the solution of

$$\widehat{\mathbf{U}},\widehat{\mathbf{V}},\widehat{\mathbf{W}} = \max_{\mathbf{U},\mathbf{V},\mathbf{W}}\mathcal{L}_{\mathbf{X},\mathbf{Y}}(\mathbf{U},\mathbf{V},\mathbf{W}) = \max_{\mathbf{U},\mathbf{V},\mathbf{W}}\{\alpha\mathcal{L}_{\mathbf{X}}(\mathbf{U},\mathbf{V}) + \mathcal{L}_{\mathbf{Y}}(\mathbf{U},\mathbf{W})\}. \tag{1}$$

3 Closed-form update rules

Rish et al. [3] mentioned that equation (1) is not convex for all parameters, but auxiliary functions can be used to find local solutions analogous to the method proposed by Schein et al. [4]. Rish et al. [3] did not provide detailed update rules of parameters. We provide the detailed update rules here for the problem that both $\mathbf{X}$ and $\mathbf{Y}$ are binary.

According to Rish et al. [3], an auxiliary function for $\mathcal{L}(\theta)$ is a function $Q(\widehat{\theta},\theta)$ such that $\mathcal{L}(\theta) = Q(\theta,\theta)$ and $\mathcal{L}(\widehat{\theta}) \geq Q(\widehat{\theta},\theta)$ for all $\widehat{\theta}$. Therefore, $\mathcal{L}(\theta)$ is non-decreasing under the update

$$\widehat{\theta} = \operatorname*{argmax}_{\widehat{\theta}} Q(\widehat{\theta},\theta).$$

Rish et al. [3] also showed that if we have an auxiliary function $Q(\widehat{\boldsymbol{\theta}_X},\boldsymbol{\theta}_X)$ for $\mathcal{L}_{\mathbf{X}}$ and an auxiliary function $Q(\widehat{\boldsymbol{\theta}_Y},\boldsymbol{\theta}_Y)$ for $\mathcal{L}_{\mathbf{Y}}$, then $\alpha Q(\widehat{\boldsymbol{\theta}_X},\boldsymbol{\theta}_X) + Q(\widehat{\boldsymbol{\theta}_Y},\boldsymbol{\theta}_Y)$ is an auxiliary function for $\mathcal{L}_{\mathbf{X},\mathbf{Y}}$.

Schein et al. [4] showed for UDR, we can choose

$$Q(\widehat{\boldsymbol{\theta}}_X,\boldsymbol{\theta}_X) = \sum_{n,d}\left\{\log 2 - \log\cosh(\theta_{X_{nd}}/2) + \frac{T_{X_{nd}}\theta^2_{X_{nd}}}{4} + \frac{(2X_{nd}-1)\widehat{\theta}_{X_{nd}}}{2} - \frac{T_{X_{nd}}\widehat{\theta}^2_{X_{nd}}}{4}\right\}, \tag{2}$$

where $T_{X_{nd}} = \tanh(\theta_{X_{nd}}/2)/\theta_{X_{nd}}$.

Analogous to (2), for labels $\mathbf{Y}$, we can choose

$$Q(\widehat{\boldsymbol{\theta}}_Y,\boldsymbol{\theta}_Y) = \sum_{n,k}\left\{\log 2 - \log\cosh(\theta_{Y_{nk}}/2) + \frac{T_{Y_{nk}}\theta^2_{Y_{nk}}}{4} + \frac{(2Y_{nk}-1)\widehat{\theta}_{Y_{nk}}}{2} - \frac{T_{Y_{nk}}\widehat{\theta}^2_{Y_{nk}}}{4}\right\}, \tag{3}$$

where $T_{Y_{nk}} = \tanh(\theta_{Y_{nk}}/2)/\theta_{Y_{nk}}$.

Therefore, we can choose the auxiliary function

$$Q(\widehat{\boldsymbol{\theta}}_X,\boldsymbol{\theta}_X,\widehat{\boldsymbol{\theta}}_Y,\widehat{\boldsymbol{\theta}}_Y) = \alpha Q(\widehat{\boldsymbol{\theta}}_X,\boldsymbol{\theta}_X) + Q(\widehat{\boldsymbol{\theta}}_Y,\boldsymbol{\theta}_Y). \tag{4}$$

For $\mathbf{V}$ update, we fix $\mathbf{U}$, i.e., let $\boldsymbol{\theta}_X = \mathbf{U}\mathbf{V}$, $\widehat{\boldsymbol{\theta}}_X = \mathbf{U}\widehat{\mathbf{V}}$, and calculate the derivative of (4) with respect to $\widehat{V}_{ld}$

$$\begin{aligned}
\frac{\partial Q}{\partial \widehat{V}_{ld}} &= \alpha \sum_{n=1}^{N} \left\{ \frac{(2X_{nd}-1)}{2} U_{nl} - \frac{T_{X_{nd}}}{4} 2\widehat{\theta}_{X_{nd}} U_{nl} \right\} \\
&= \frac{\alpha}{2} \sum_{n=1}^{N} \left\{ (2X_{nd}-1) U_{nl} - T_{X_{nd}} \sum_{l'=1}^{L+1} U_{nl'} \widehat{V}_{l'd} U_{nl} \right\} \\
&= \frac{\alpha}{2} \left\{ \sum_{n=1}^{N} (2X_{nd}-1) U_{nl} - \sum_{l'=1}^{L+1} \left(\sum_{n=1}^{N} T_{X_{nd}} U_{nl'} U_{nl} \right) \widehat{V}_{l'd} \right\} \\
&\equiv \frac{\alpha}{2} \left\{ b_{X_{dl}} - \sum_{l'=1}^{L+1} A_{X_{dll'}} \widehat{V}_{l'd} \right\}.
\end{aligned} \tag{5}$$

Therefore, the d-th column of $\widehat{\mathbf{V}}$, $\widehat{\mathbf{V}}_d$, can be solved from the linear equations

$$\sum_{l'=1}^{L+1} A_{X_{dll'}} \widehat{V}_{l'd} = b_{X_{dl}},$$

where $A_{X_{dll'}}$ and $b_{X_{dl}}$ are defined in (5).

For $\mathbf{W}$ update, we fix $\mathbf{U}$, i.e., let $\boldsymbol{\theta}_Y = \mathbf{U}\mathbf{W}$, $\widehat{\boldsymbol{\theta}}_Y = \mathbf{U}\widehat{\mathbf{W}}$, and calculate the derivative of (4) with respect to $\widehat{W}_{lk}$:

$$\begin{aligned}
\frac{\partial Q}{\partial \widehat{W}_{lk}} &= \sum_{n=1}^{N} \left\{ \frac{(2Y_{nk}-1)}{2} U_{nl} + \frac{T_{Y_{nk}}}{4} 2\widehat{\theta}_{Y_{nk}} U_{nl} \right\} \\
&= \frac{1}{2} \sum_{n=1}^{N} \left\{ (2Y_{nk}-1) U_{nl} + T_{Y_{nk}} \sum_{l'=1}^{L+1} U_{nl'} \widehat{W}_{l'k} U_{nl} \right\} \\
&= \frac{1}{2} \left\{ \sum_{n=1}^{N} (2Y_{nk}-1) U_{nl} + \sum_{l'=1}^{L+1} \left(\sum_{n=1}^{N} T_{Y_{nk}} U_{nl'} U_{nl} \right) \widehat{W}_{l'k} \right\} \\
&\equiv \frac{1}{2} \left\{ b_{Y_{kl}} + \sum_{l'=1}^{L+1} A_{X_{kll'}} \widehat{W}_{l'k} \right\}.
\end{aligned} \tag{6}$$

Therefore, the k-th column of $\widehat{\mathbf{W}}$, $\widehat{\mathbf{W}}_k$, can be solved from the linear equations

$$\sum_{l'=1}^{L+1} A_{Y_{kll'}} \widehat{W}_{l'k} = b_{Y_{kl}},$$

where $A_{Y_{kll'}}$ and $b_{Y_{kl}}$ are defined in (6).

For $\mathbf{U}$ update, we fix $\mathbf{V}$ and $\mathbf{W}$, i.e., let $\boldsymbol{\theta}_X = \mathbf{U}\mathbf{V}$, $\widehat{\boldsymbol{\theta}}_X = \widehat{\mathbf{U}}\mathbf{V}$, $\boldsymbol{\theta}_Y = \mathbf{U}\mathbf{W}$, $\widehat{\boldsymbol{\theta}}_Y = \widehat{\mathbf{U}}\mathbf{W}$, and calculate the derivative of (4) with respect to $\widehat{U}_{nl}$:

$$\begin{aligned}
\frac{\partial Q}{\partial \widehat{U}_{nl}} &= \alpha \sum_{d=1}^{D} \left\{ \frac{(2X_{nd}-1)}{2} V_{ld} + \frac{T_{X_{nd}}}{4} 2\widehat{\theta}_{X_{nd}} V_{ld} \right\} + \sum_{k=1}^{K} \left\{ \frac{(2Y_{nk}-1)}{2} W_{lk} + \frac{T_{Y_{nk}}}{4} 2\widehat{\theta}_{Y_{nk}} W_{lk} \right\} \\
&= \frac{\alpha}{2} \sum_{d=1}^{D} (2X_{nd}-1) V_{ld} + \frac{1}{2} \sum_{k=1}^{K} (2Y_{nk}-1) W_{lk} \\
&\quad - \sum_{l'=1}^{L+1} \left(\frac{\alpha}{2} \sum_{d=1}^{D} T_{X_{nd}} V_{l'd} V_{ld} + \frac{1}{2} \sum_{k=1}^{K} T_{Y_{nk}} W_{l'k} W_{lk} \right) \widehat{U}_{nl'} \\
&\equiv b_{XY_{nl}} - \sum_{l'=1}^{L+1} A_{XY_{nll'}} \widehat{U}_{nl'}.
\end{aligned} \tag{7}$$

Therefore, the n-th row of $\widehat{\mathbf{U}}$, $\widehat{\mathbf{U}}_n$, can be updated from the linear equations

$$\sum_{l'=1}^{L} A_{XY_{nll'}} \widehat{U}_{nl'} = b_{XY_{nl}},$$

where $b_{XY_{nl}}$ and $A_{XY_{nll'}}$ are defined in (7). Notice that the last column of $\mathbf{U}$ is fixed as 1, so the update rule is only used for the first L column of $\mathbf{U}$.

4 Application to the SNP data

A single-nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide in the genome differs between members of a biological specie or paired chromosomes in an individual. It has been estimated that the entire human population harbors 10 million so-called "common" SNPs with a minor allele frequency (i.e., the percentage of all living humans that have the rarer nucleotide (allele) for this SNP, as opposed to the other more frequent nucleotide) of greater than 5% in the human population. According to Li et al. [1], over 6 million of these SNPs have been identified across the human genome.

Human Genome Diversity Panel (HGDP) contains 49,553 SNPs of 938 participants from 53 different races. For each SNP, the minor allele status is coded as 1 and the major allele status is coded as 0. For this project, we only use the data of participants who are Han-Chinese (68), Japanese (56) or Yoruba (42).

We apply the logistic-SDR discussed in this project with comparison to logistic-UDR. For logistic-SDR, we apply SDR on the data, and learn the logistic regression model at the same time as the dimensionality reduction model. For logistic-UDR, we first learn the UDR from data, and then learn the logistic regression model on the coordinates $\mathbf{U}$.

Figure 1 shows the projections of samples on the first three bases (PCs) if UDR is applied to data with features only. From Figure 1, the samples are not separated on the low-dimensional space ($L = 2$), especially for Han-Chinese and Japanese. We label Han-Chinese as 1 and other as 0, i.e., we want to classify whether the samples are Han-Chinese or not.

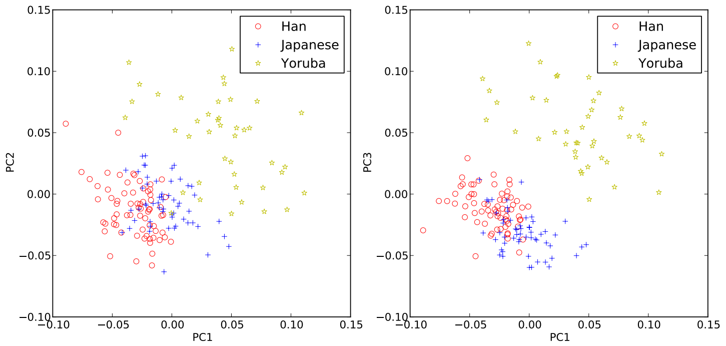

Figure 1: The projections of samples on the first PC, the second PC and the third PC using UDR.

A logistic regression is applied to the data using the the coordinates on the first three PCs, but the misclassification error is quite high (0.1807).

Figure 2 shows the projections of samples on the first three bases (PCs) if logistic-SDR is performed ($\alpha = 0.01$). From Figure 1, the samples are clearly separated on the low-dimensional space ($L = 2$), as a consequence, the misclassification error is 0.

We also try different values of α, and notice that the misclassification error decreases when α decreases.

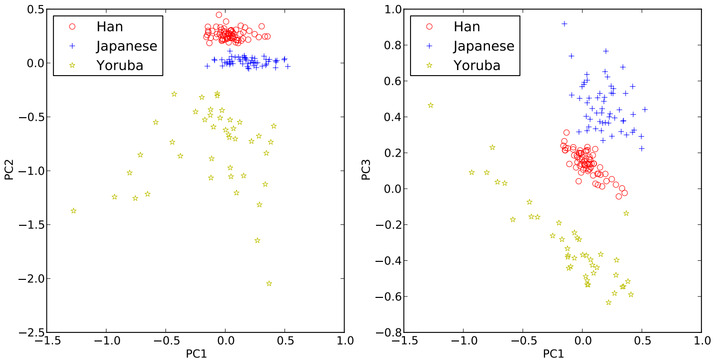

Figure 2: The projections of samples on the first PC, the second PC and the third PC using normal PCA.

5 Conclusion and future work

In this project, we derived closed-form update rules for supervised dimensionality reduction when responses and features are all binary. This work is a generalization of Schein et al [4] on UDR to SDR. Rish et al. [3] claimed a general result about SDR with generalized linear models, but they did not provide detailed closed-form update rules. We went through detailed calculations in this project when responses and features are all binary, and wrote code for logistic-UDR and logistic-SDR in `python`.

We applied logistic SDR method to a classification problem in statistical genetics. We showed that using logistic SDR method, we can find a low-dimensional representation of data and significantly reduce misclassification error rate. This result is of importance in statistical genetics.

Rish et al. [3] proposed a general prediction step for new data without labels, but actually their approach need the labels of those new data to proceed in the update steps. I do not think it makes sense to use the labels of test data in prediction step. Future work may provide an alternative prediction step for new data without labels using logistic-SDR. With such steps, we can divide data into training and testing, and then make fair comparisons with other methods.

The choice of tuning parameter α is still an open question. Rish et al. [3] provided general comments on the choice of α. Possible methods which subjectively choose α based on the data may be helpful in future.

References

[1] Li, J.Z. et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, Vol 139: 1100-1104.

[2] McVean, G. (2009) A genealogical interpretation of principal component analysis. *PLoS Genetics*, 5(10): e1000686.

[3] Rish, I., Grabarnik, G., Cecchi, G., Pereira, F. & Gordon, G.J. (2008) Closed-form supervised dimensionality reduction with generalized linear models. *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pp. 832-839. Helsinki, Finland.

[4] Schein. A., Saul, L., & Ungar, L. (2003) A generalized linear model for principal component analysis of binary data, *Ninth International Workshop on Artificial Intelligence and Statistics*.