CS 5727: Homework #2

Due on Wednesday, September 27, 2017

Zhan Zhang

zz524@cornell.edu

Jialiang Wang

jw2476@cornell.edu

September 27, 2017

Contents

Figure 1: Sample Face Image

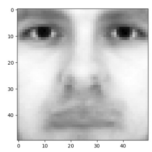


Figure 2: Average Face Image from the Training Set

PROGRAMMING EXERCISES

Problem 1

(b) A sample face image is displayed in figure 1.

(c) The average face is displayed in figure 2.

(d) A mean subtraction from the training set is display in figure 3 (a).
A mean subtraction from the testing set is display in figure 3 (b).

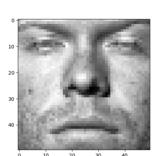
(e) The first 10 eigenfaces are displayed in figure 4.

(f) The plot for the rank-r approximation error is shown in figure 5.

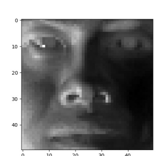
(h) The classification accuracy from the logistic model trained is plotted in figure 6.

Problem 2

(b)



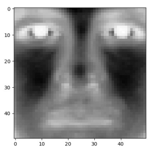
(a) Eigenface 1



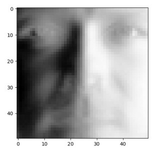
(b) Eigenface 2

Figure 3: Mean Subtraction Face Images

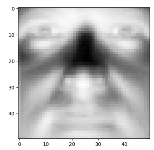
Problem 2 continued on next page. . .



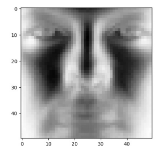
(a) Eigenface 1



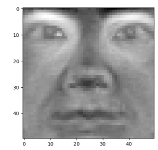
(b) Eigenface 2



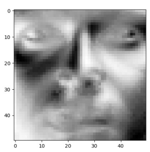
(c) Eigenface 3



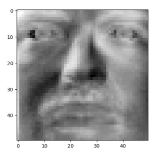
(d) Eigenface 4



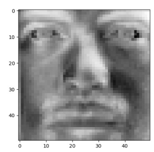
(e) Eigenface 5



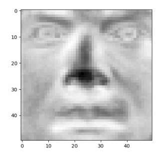
(f) Eigenface 6



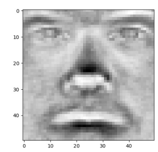
(g) Eigenface 7



(h) Eigenface 8



(i) Eigenface 9



(j) Eigenface 10

Figure 4: Eigenfaces

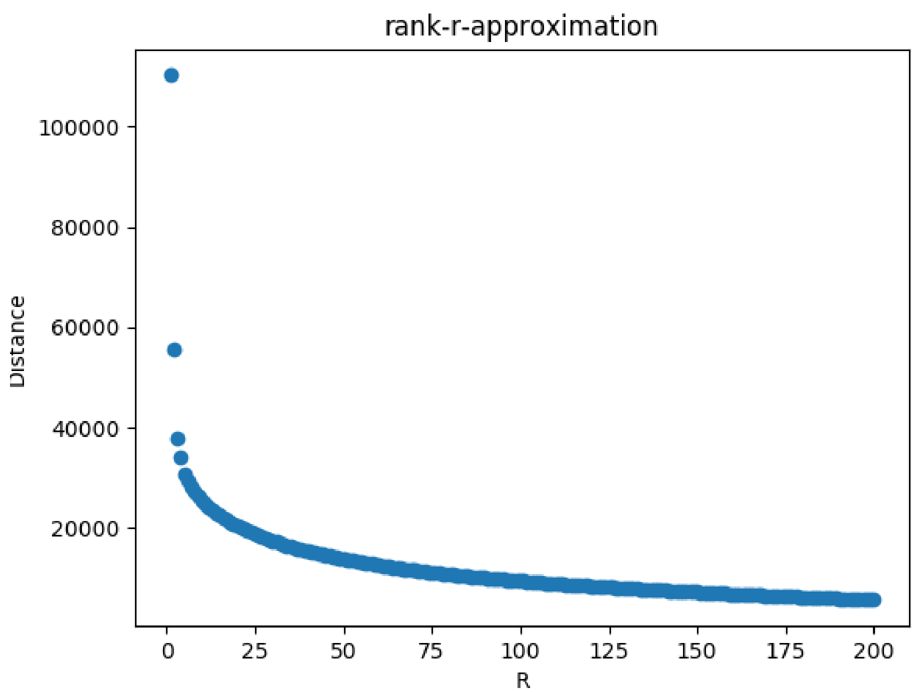


Figure 5: Rank-r Approximation Error

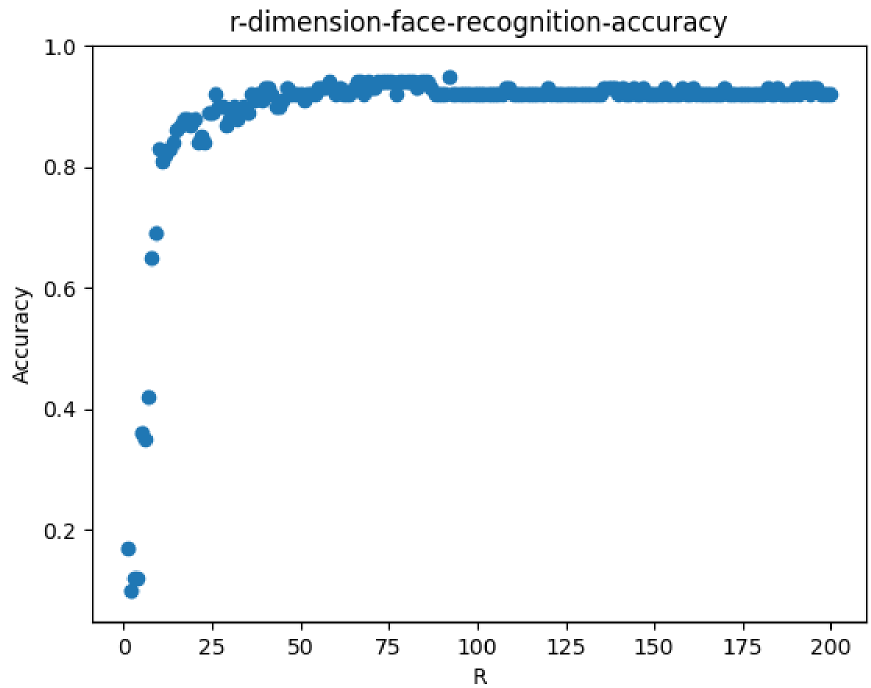


Figure 6: Classification Accuracy of the Logistic Model

- There are 39774 samples in the training set;
- There are 20 categories in the training set;
- There are 6714 unique ingredients appearing in the training set;
- There are 7137 unique ingredients appearing in both the training and testing set.

(d) The average accuracy for 3-fold cross-validation for both Gaussian prior and Bernoulli prior are 0.38215893891 and 0.678408428328.
System with Bernoulli prior have a performance much better than the one with Gaussian piror.
This could be explained from that the features (ingredients) are mark as 0 or 1 (exists or non-exists) which is a better fit for the Bernoulli Distribution.
(f) The average accuracy for 3-fold cross-validation for Logistic Regression is 0.775758670409.
(g) The outcome has been submitted to kaggle, shown in figure 7.

WRITTEN EXERCISES

Problem 3

It subjects to Lagrange multiplier condition :

$$d\mathcal{L}/da = a^T B^T + a^T B - \lambda(a^T W^T + a^T W) = a^T[(B^T - \lambda W^T) + (B - \lambda W)] = 0$$

So this problem becomes $Ba = \lambda W a$

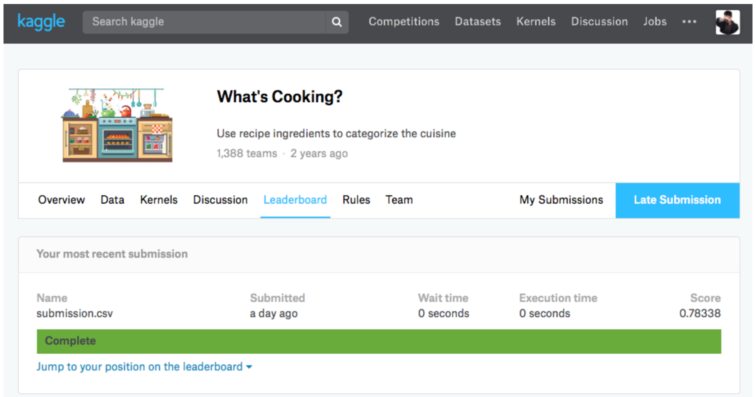


Figure 7: Kaggle What's Cooking Submission Record

By Cholesky decomposition, $W = LDL^T = DD^T, B = DCD^T$ so

$$DCD^T a = \lambda DD^T a$$

So that $C(D^T a) = \lambda(D^T a)$
let $(D^T a) = y$, then

$$Cy = \lambda y$$

In this way, we convert the problem of find x to maximize $a^T Ba$ to compute a matrix C and find a y st.

$$Cy = \lambda y$$

and $y = D^T a$

Problem 4

Please see the handwritten appendix for the solution.

Problem 5

(a)

$$M^T M = \begin{bmatrix} 39 & 57 & 60 \\ 57 & 118 & 53 \\ 60 & 53 & 127 \end{bmatrix} \tag{1}$$

and

$$MM^T = \begin{bmatrix} 10 & 9 & 26 & 3 & 26 \\ 9 & 62 & 8 & -5 & 85 \\ 26 & 8 & 72 & 10 & 50 \\ 3 & -5 & 10 & 2 & -1 \\ 26 & 85 & 50 & -1 & 138 \end{bmatrix} \tag{2}$$

(b) The eigen values for both two matrices are 214.6705 and 69.3295.

(c) The eigen vectors for matrix $M_T M$ are the row vectors in the matrix:

$$\begin{bmatrix} 0.904534033733291 & 0.0146040411173086 & 0.426151268684285 \\ -0.301511344577764 & 0.728597992755815 & 0.615008840621910 \end{bmatrix} \tag{3}$$

The eigen vectors for matrix MM_T are the row vectors in the matrix:

$$\begin{bmatrix} 0.143201385117688 & -0.944594949352975 & 0.00448837927376996 & -0.244973232790239 & 0.164929423163427 \\ 0.525570082350403 & 0.0470413984460477 & 0.541872014065717 & 0.453306436544331 & 0.471647315618646 \end{bmatrix} \tag{4}$$

(d) The SVD for the original matrix M is:

$$U = \begin{bmatrix} -0.164929423163427 & -0.244973232790239 \\ -0.471647315618646 & 0.453306436544331 \\ -0.336470547433036 & -0.829439646984777 \\ -0.00330585055309077 & -0.169746590702150 \\ -0.798200311392409 & 0.133106561666003 \end{bmatrix} \tag{5}$$

$$\Sigma = \begin{bmatrix} 14.6516377649769 & 0 \\ 0 & 8.32643445923304 \end{bmatrix} \tag{6}$$

$$X = \begin{bmatrix} -0.426151268684285 & 0.0146040411173084 & -0.904534033733291 \\ -0.615008840621911 & 0.728597992755815 & 0.301511344577764 \end{bmatrix} \tag{7}$$

(e) The one-dimensional approximation to the matrix M is:

$$M(recovered) = \begin{bmatrix} 1.02978864496302 & -0.0352904633127328 & 2.18579397826745 \\ 2.94487812262642 & -0.100919847830279 & 6.25069707133886 \\ 2.10085952200109 & -0.0719956529420139 & 4.45921220323874 \\ 0.0206411160375204 & -0.000707363158274219 & 0.0438121233519250 \\ 4.98381429651495 & -0.170793411297470 & 10.5784729045218 \end{bmatrix} \tag{8}$$

4.2 (1)

$$\delta_i(x) = x^T \cdot \hat{\Sigma}^{-1} \hat{u}_i - \frac{1}{2} \hat{u}_i^T \hat{\Sigma} \hat{u}_i + \log \hat{\pi}_i$$

where π_i is ith prior probability

$$f = \arg\max \delta_i(x)$$

in order to classify as class 2

$$\Rightarrow \delta_2(x) > \delta_1(x) \Rightarrow \delta_2(x) - \delta_1(x) > 0$$

$$\Rightarrow x^T \hat{\Sigma}^{-1}(\hat{u}_2 - \hat{u}_1) - \frac{1}{2} \hat{u}_2^T \hat{\Sigma} \hat{u}_2 + \frac{1}{2} \hat{u}_1^T \hat{\Sigma} \hat{u}_1 + \log\left(\frac{N_2}{N}\right) - \log\left(\frac{N_1}{N}\right) > 0$$

$$\Rightarrow x^T \hat{\Sigma}^{-1}(\hat{u}_2 - \hat{u}_1) > \frac{1}{2}\left[\hat{u}_2^T \hat{\Sigma} \hat{u}_2 - \hat{u}_1^T \hat{\Sigma} \hat{u}_1\right] - \log\left(\frac{N_2}{N_1}\right)$$

$$\Rightarrow x^T \hat{\Sigma}^{-1}(\hat{u}_2 - \hat{u}_1) > \frac{1}{2}(\hat{u}_2 + \hat{u}_1)^T \hat{\Sigma}^{-1}(\hat{u}_2 - \hat{u}_1) - \log\left(\frac{N_2}{N_1}\right)$$

(2) $\cancel{\sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2}$

$$= \cancel{\sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)(y_i - \beta_0 - x_i^T \beta)}$$

(2) ~~let~~ $\hat{u}_j = \frac{1}{n_j} \sum_{y_i=j} x_i$

estimated parameters

$\hat{\pi}_j = n_j / n$, is prior probability

$\hat{\Sigma} = \frac{1}{n-k} \sum_{j=1}^{k} \sum_{y_i=j} (x_i - \hat{u}_j)(x_i - \hat{u}_j)^T$

① let U_i be n element vector with j-th element is 1 if j-th observation ~~in~~ is class i and 0 otherwise.

② let t_i be our target labels so that $t_i \in \{1, 2\}$.

③ let Y be our target vector. so that $Y = t_1 U_1 + t_2 U_2$ and $\vec{1} = U_1 + U_2$

④ ~~$N_j \hat{u}_j = \sum_{y_i=j} x_i$~~

$\therefore N_i \hat{U}_i = \sum_{y_j=i} x_j = X^T U_i$

where $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

$\therefore X^T Y = t_1 N_1 \hat{u}_1 + t_2 N_2 \hat{u}_2$

That's $\begin{cases} 2X^T X\beta - 2X^T Y + 2\beta_0 X^T \vec{1} = 0 & ① \\ 2N\beta_0 - 2\vec{1}^T (Y - X\beta) = 0 & ② \end{cases}$

Use ② $\Rightarrow \hat{\beta}_0 = \frac{1}{N} \cdot \vec{1}^T (Y - X\beta)$ ③

③ $\rightarrow$ ①, we get:

$$X^T X \hat{\beta} - \cancel{\#} X^T Y + \frac{1}{N} \cdot \vec{1}^T (Y - X\hat{\beta}) X^T \vec{1} = 0$$

$$\cancel{\Rightarrow X^T X \beta - X^T Y + \frac{1}{N} \cdot \vec{1}^T Y}$$

$$X^T \cdot X \cdot \hat{\beta} - X^T Y + \frac{1}{N} X^T \cdot \vec{1} \cdot \vec{1}^T \cdot Y - \frac{1}{N} X^T \cdot 1 \cdot 1^T X \cdot \hat{\beta} = 0$$

$$\therefore (X^T \cdot X \cdot \cancel{\beta} - \frac{1}{N} X^T \cdot \vec{1} \cdot \vec{1}^T \cdot X) \hat{\beta} = X^T Y - \frac{1}{N} X^T 1 1^T Y$$

$$\Rightarrow X^T Y - \frac{1}{N} (X^T 1)(1^T Y)$$

$$= t_1 N_1 \hat{u}_1 + t_2 N_2 \hat{u}_2 - \frac{1}{N} [X^T \cdot (u_1 + u_2)] \cdot [(u_1 + u_2) \cdot Y]$$

$$= t_1 N_1 \hat{u}_1 + t_2 N_2 \hat{u}_2 - \frac{1}{N} [N_1 \hat{u}_1 + N_2 \hat{u}_2] \cdot [(u_1 + u_2)^T \cdot [t_1 u_1 + t_2 u_2]]$$

$$= t_1 N_1 \hat{u}_1 + t_2 N_2 \hat{u}_2 - \frac{1}{N} [N_1 \hat{u}_1 + N_2 \hat{u}_2] [t_1 N_1 + t_2 N_2]$$

$$= \frac{\cancel{N\cdot(t_1\hat{u}_1 + t_2 u_2) - [N_1^2 t_1 \hat{u}_1 + N_1\hat{u}_1 N_2 t_2}}{N}$$

$$\frac{N(\overset{N_1}{t_1}\hat{u}_1 + \overset{N_2}{t_2}\hat{u}_2) - [N_1^2 \hat{u}_1 t_1 + N_1 N_2 \hat{u}_1 t_2 + N_1 N_2 \hat{u}_2 t_1 + N_2^2 \hat{u}_2 t_2]}{N}$$

$$\Big[(N t_1 N_1 \hat{u}_1 - N_1 t_1 N_1 \hat{u}_1) + (N t_2 N_2 \hat{u}_2 - N_2 t_2 N_2 \hat{u}_2)$$

$$- N_1 N_2 \hat{u}_1 t_2 - N_1 N_2 \hat{u}_2 t_1\Big] / N$$

$$= \big[N_2 t_1 N_1 \hat{u}_1 + N_1 t_2 N_2 \hat{u}_2 - N_1 N_2 \hat{u}_1 t_2 - N_1 N_2 \hat{u}_2 t_1\big] / N$$

$$= \frac{N_1 N_2}{N}\left[t_1 \hat{u}_1 + t_2 \hat{u}_2 - \hat{u}_1 t_2 - \hat{u}_2 t_1\right]$$

$$= \frac{N_1 N_2}{N}\left[(t_1 - t_2)\hat{u}_1 - (t_1 - t_2)\hat{u}_2\right]$$

$$= \frac{N_1 N_2}{N}\cdot(t_1 - t_2)(\hat{u}_1 - \hat{u}_2)$$

$$\therefore X^T Y - \frac{1}{N} X^T \mathbf{1}\mathbf{1}^T Y = \frac{N_1 N_2}{N}(t_1 - t_2)(\hat{u}_1 - \hat{u}_2)$$

Then $X^T X - \frac{1}{N} X^T 1 1^T Y$

We know $\frac{1}{N} X^T 1 \cdot 1^T Y$

$$= \frac{[N_1^2 \hat{u}_1 t_1 + N_1 N_2 \hat{u}_1 t_2 + N_1 N_2 \hat{u}_2 t_1 + N_2^2 \hat{u}_2 t_2]}{N}$$

$$X^T X = \sum (X - u_i)(X - u_i)^T + N_1 \hat{u}_1 \hat{u}_1^T + N_2 \hat{u}_2 \hat{u}_2^T$$

We have $\hat{\Sigma} = \frac{1}{N-2} \sum_{j=1}^{k} \sum_{y_i = j} (x_i - \hat{u}_j)(x_i - \hat{u}_j)^T$

$\therefore X^T X = (N-2) \cdot \hat{\Sigma} + N_1 \hat{u}_1 \hat{u}_1^T + N_2 \hat{u}_2 \hat{u}_2^T$

$\therefore X^T X - \frac{1}{N} X^T \vec{1} \vec{1}^T \cdot X = (N-2)\hat{\Sigma} + \frac{N_1 N_2}{N} \cdot \hat{\Sigma}_B$

where $\hat{\Sigma}_B$ is definition in problem

$$\therefore \left[(N-2)\hat{\Sigma} + \frac{N_1 N_2}{N} \hat{\Sigma}_B \right) \hat{\beta} = \frac{N_1 N_2}{N} (t_1 - t_2)(\hat{u}_1 - \hat{u}_2)$$

~~with~~ $\cancel{\begin{cases} t_1 = \frac{-N}{N_1} \\ t_2 = \frac{N}{N_2} \end{cases}}$ with $\begin{cases} t_1 = \frac{-N}{N_1} \\ t_2 = \frac{N}{N_2} \end{cases}$

$$\Rightarrow \left[(N-2)\hat{\Sigma} + \frac{N_1 N_2}{N} \hat{\Sigma}_B \right) \hat{\beta} = \frac{N_1 N_2}{N} \left[\frac{-N}{N_1} - \frac{N}{N_2} \right] [\hat{u}_1 - \hat{u}_2]$$

$$= \frac{N_1 N_2}{N} \cdot \frac{N^2}{N_1 N_2} (\hat{u}_2 - \hat{u}_1)$$

$$= N (\hat{u}_2 - \hat{u}_1)$$

(3) show $\hat{\Sigma}_B \beta$ in direction of $(\hat{u}_2 - \hat{u}_1)$

$$\hat{\Sigma}_B = \frac{N_1 N_2}{N^2} (\hat{u}_2 - \hat{u}_1)(\hat{u}_2 - \hat{u}_1)^T$$

$\therefore$ $\cancel{\hat{\Sigma}_B \hat{\beta}}$

$\hat{\beta} = C \cdot (\hat{u}_2 - u_1)$. from (2)

where $C \in R$

Thus. $\hat{\Sigma}_B \hat{\beta} = \left(C \cdot \frac{N_1 N_2}{N^2}\right) \cdot (\hat{u}_2 - \hat{u}_1)(\hat{u}_2 - u_1)^T \cdot C(\hat{u}_2 - u_1)$

$= \lambda \cdot (\hat{u}_2 - \tilde{u}_1)$

where $\lambda \in R$.

$\therefore \hat{\beta} \propto \hat{\Sigma}^{-1} (\hat{u}_2 - \hat{u}_1)$

(4) t_1, t_2 are arbitrary and distinct, so it satisfies for any (distinct) coding of two classes.

(5) From (2), we have

$$\hat{\beta}_0 = \frac{1}{N} \cdot \vec{1}^T (Y - X\hat{\beta})$$

$$= \frac{1}{N}(t_1 N_1 + t_2 N_2) - \frac{1}{N} \cdot 1^T \cdot X \hat{\beta}$$

$$\cancel{= \frac{1}{N}}$$

$$\cancel{= \frac{1}{N}(t_1 N_1 + t_2 N_2) - \frac{1}{N}(N_1 \mu_1^T + N_2 \mu_2^T) \hat{\beta}}$$

$$= \frac{1}{N}(t_1 N_1 + t_2 N_2) - \frac{1}{N}[N_1 \mu_1^T + N_2 \mu_2^T] \cdot \hat{\beta}$$

$\because$ $t_1 N_1 + t_2 N_2 = \frac{-N}{N_1} \cdot N_1 + \frac{N}{N_2} N_2 = 0$

$\therefore$ $\hat{\beta}_0 = -\frac{1}{N}[N_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2^T] \cdot \hat{\beta}$

$\therefore$ $f(x) = \frac{1}{N}(N x^T - N_1 \hat{\mu}_1^T - N_2 \hat{\mu}_2^T) \hat{\beta}$

$$= \frac{1}{N}(N \cdot x^T - N_1 \hat{\mu}_1^T - N_2 \hat{\mu}_2^T) \cdot \lambda \cdot \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1)$$

~~classify to x~~

~~if f~~

if $f(x) > 0$

$$\Rightarrow N x^T \cdot \lambda \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) > (N_1 \mu_1^T + N_2 \mu_2^T) \lambda \cdot \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1)$$

$$\Rightarrow x^T \cdot \cancel{\lambda} \cdot \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{N}(N_1 \mu_1^T + N_2 \mu_2^T) \cdot \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1)$$

which is different from LDA rule unless ~~N~~ $N_1 = N_2$, so that $\log\left(\frac{N_1}{N_2}\right) = 0$