Korea Advanced Institute of Science and Technology

School of Electrical Engineering

EE488 Introduction to Machine Learning Spring 2018

Student's Name: Dinh Vu

Student's ID: 20184187

Homework 3

I. PART 1

1. K-NN

i. Problem 1-i

Because

$P(\theta, \theta_n|x, x_n) = P(\theta|x)P(\theta_n|x_n)$

$\Rightarrow P(\theta = \omega_l, \theta_n = \omega_l|x, x_n) = P(\omega_l|x)P(\omega_l|x_n)$

The error rate can be express as:

$$P(\theta \neq \theta_n|x, x_n) = P(e|x, x_n) = 1 - \sum_{l=1}^{C} P(\theta = \omega_l, \theta_n = \omega_l|x, x_n)$$

$$= 1 - \sum_{l=1}^{C} P(\omega_l|x)P(\omega_l|x_n) \quad (1.1)$$

ii. Problem 1-ii

Bayes error rate is defined as below:

$$P^*(e|x) = 1 - \max_{\theta\in\Omega} P(\omega_m|x) = 1 - P(\omega_m|x) \Rightarrow P(\omega_m|x) = 1 - P^*(e|x)$$

The sum of squares can be divided into 2 parts as follows:

$$\sum_{i=1}^{C} P^2(\omega_i|x) = P^2(\omega_m|x) + \sum_{i\neq m} P^2(\omega_i|x)$$

$$= [1 - P^*(e|x)]^2 + \sum_{i\neq m} P^2(\omega_i|x) \quad (1.2)$$

In order to find lower bound of $\sum_{i=1}^{C} P^2(\omega_i|x)$, the minimum value of $\sum_{i\neq m} P^2(\omega_i|x)$ must be determined while satifying the below condition:

$$\sum_{l=1}^{C} P(\omega_i|x) = 1 \Leftrightarrow P(\omega_m|x) + \sum_{l \neq m} P(\omega_i|x) = 1$$

$$\Leftrightarrow \sum_{l \neq m} P(\omega_i|x) = 1 - P(\omega_m|x) = P^*(e|x)$$

According to Bunyakovsky for $C - 1$ elements:

$$\left(\sum_{i=1}^{C-1} 1^2\right) \times \left[\sum_{i \neq m} P^2(\omega_i|x)\right] \geq \left[\sum_{i \neq m} 1 \times P(\omega_i|x)\right]^2$$

$$\Leftrightarrow (C-1) \sum_{i \neq m} P^2(\omega_i|x) \geq \left[\sum_{i \neq m} P(\omega_i|x)\right]^2 = P^{*2}(e|x)$$

$$\Leftrightarrow \sum_{i \neq m} P^2(\omega_i|x) \geq \frac{P^{*2}(e|x)}{C-1}$$

So, $\min\left[\sum_{i \neq m} P^2(\omega_i|x)\right] = \dfrac{P^{*2}(e|x)}{C-1}$ if only if

$$P(\omega_1|x) = P(\omega_2|x) = \cdots = P(\omega_{m-1}|x) = P(\omega_{m+1}|x) = \cdots = P(\omega_C|x) = \frac{P^*(e|x)}{C-1}$$
$$= \frac{1 - P(\omega_m|x)}{C-1} = \frac{1-S}{C-1}$$

iii. Problem 1-iii

Therefore, from (1.2):

$$\sum_{i=1}^{C} P^2(\omega_i|x) \geq [1 - P^*(e|x)]^2 + \frac{P^{*2}(e|x)}{C-1} \quad (1.3)$$

Hence, the lower bound of $\sum_{i=1}^{C} P^2(\omega_i|x)$ in term of $P^*(e|x)$ is:

$$[1 - P^*(e|x)]^2 + \frac{P^{*2}(e|x)}{C-1}$$

iv. Problem 1-iv

Because when the size of training data is infinite, the nearest neighbor of x will be itself, so:

$$P(e|x, x_n) = P(e|x, x) = P(e|x)$$

$$P(\omega_l|x_n) = P(\omega_l|x)$$

Replace to (1.1):

$$P(e|x) = 1 - \sum_{l=1}^{C} P^2(\omega_l|x) \Rightarrow \sum_{l=1}^{C} P^2(\omega_l|x) = 1 - P(e|x)$$

Replace to (1.3):

$$1 - P(e|x) \geq [1 - P^*(e|x)]^2 + \frac{P^{*2}(e|x)}{C-1} \Leftrightarrow P(e|x)$$
$$\leq 1 - [1 - P^*(e|x)]^2 - \frac{P^{*2}(e|x)}{C-1}$$

$$\Leftrightarrow P(e|x) \leq 1 - [1 - 2P^*(e|x) + P^{*2}(e|x)] - \frac{P^{*2}(e|x)}{C-1}$$

$$\Leftrightarrow P(e|x) \leq 2P^*(e|x) - P^{*2}(e|x) - \frac{P^{*2}(e|x)}{C-1}$$

$$\Leftrightarrow P(e|x) \leq 2P^*(e|x) - \frac{C}{C-1} P^{*2}(e|x)$$

Conclusion, the upper bound of the error rate in term of Bayes error rate is:

$$2P^*(e|x) - \frac{C}{C-1} P^{*2}(e|x)$$

2. Bayesian Statistic

i. Problem 2-i

Prior: $\theta \sim Beta(\alpha_0, \beta_0)$

Observe data $D = \{0,0,1,0,1,1,1\}$, there are $y = 4$ one-labels and 3 one-labels. The data size is $n = 7$.

$P(\theta|D) \sim P(D|\theta)P(\theta)$

$$\sim \binom{7}{4} \theta^4 (1-\theta)^3 \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \theta^{\alpha_0 - 1} (1-\theta)^{\beta_0 - 1}$$

$$\sim \theta^{4+\alpha_0-1}(1-\theta)^{3+\beta_0-1} = \theta^{3+\alpha_0}(1-\theta)^{2+\beta_0}$$

$P(\theta|D) = Beta(\alpha_0 + 4, \beta_0 + 3)$

$$P(x=1|D) = \int_0^1 P(x=1|\theta)P(\theta|D)d\theta = \int_0^1 \theta P(\theta|D)\, d\theta = E[\theta|D] = \frac{4+\alpha_0}{7+\alpha_0+\beta_0}$$

ii. Problem 2-ii

Prior $\mu \sim \mathcal{N}\left(\mu_0, \frac{1}{r_0}\right)$, so:

$$P(\mu) = \frac{1}{\sqrt{2\pi \times \frac{1}{r_0}}} \exp\left[-\frac{(\mu-\mu_0)^2}{2 \times \frac{1}{r_0}}\right] = \sqrt{\frac{r_0}{2\pi}} \exp\left[-\frac{r_0(\mu-\mu_0)^2}{2}\right]$$

According Bayes Rule:

$$P(\mu|D) = \frac{P(D|\mu)P(\mu)}{P(D)}$$

Because given data $D = \{x_1, x_2, \dots, x_n\}$ from Gaussian distribution with mean μ and known variance σ^2:

$$P(x_i|\mu,\sigma^2) = \mathcal{N}(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

Therefore:

$$P(D|\mu) = P(D|\mu,\sigma^2) = \prod_{i=1}^{n} P(x_i|\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

Since $P(D)$ is constant: $P(\mu|D) \propto P(D|\mu)P(\mu)$

$$\Rightarrow P(\mu|D) \propto \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]\sqrt{\frac{r_0}{2\pi}} \exp\left[-\frac{r_0(\mu-\mu_0)^2}{2}\right]$$

$$\Rightarrow P(\mu|D) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{r_0(\mu-\mu_0)^2}{2}\right]$$

Let $r = \frac{1}{\sigma^2}$

$$\Rightarrow P(\mu|D) \propto \exp\left[-\frac{1}{2}r\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{r_0(\mu-\mu_0)^2}{2}\right]$$

$$\Rightarrow P(\mu|D) \propto \exp\left[-\frac{1}{2}\left(r\sum_{i=1}^{n}(x_i-\mu)^2 + r_0(\mu-\mu_0)^2\right)\right]$$

$$= \exp\left\{-\frac{1}{2}\left[r\sum_{i=1}^{n}(x_i^2 - 2x_i\mu + \mu^2) + r_0(\mu^2 - 2\mu\mu_0 + \mu_0^2)\right]\right\}$$

$$P(\mu|D) \propto \exp\left\{-\frac{1}{2}\left[r\sum_{i=1}^{n}(-2x_i\mu + \mu^2) + r_0(\mu^2 - 2\mu\mu_0)\right]\right\}$$

$$= \exp\left[-\frac{1}{2}\left(-2r\mu\sum_{i=1}^{n}x_i + nr\mu^2 + r_0\mu^2 - 2r_0\mu\mu_0\right)\right]$$

Let $S = \sum_{i=1}^{n} x_i$

$$P(\mu|D) \propto \exp\left\{-\frac{1}{2}[(r_0 + nr)\mu^2 - 2\mu(r_0\mu_0 + rS)]\right\}$$

$$P(\mu|D) \propto \exp\left\{-\frac{1}{2}(r_0 + nr)\left[\mu^2 - 2\mu\frac{r_0\mu_0 + rS}{r_0 + nr} + \left(\frac{r_0\mu_0 + rS}{r_0 + nr}\right)^2\right]\right\}$$

$$P(\mu|D) \propto \exp\left\{-\frac{1}{2}(r_0 + nr)\left(\mu - \frac{r_0\mu_0 + rS}{r_0 + nr}\right)^2\right\}$$

Finally, $P(\mu|D) = \mathcal{N}\left(\frac{r_0\mu_0 + rS}{r_0 + nr}, \frac{1}{r_0 + nr}\right)$

Where $r = \frac{1}{\sigma^2}$ $and\ S = \sum_{i=1}^{n} x_i$

iii. Problem 2-iii

Prior: $P(\mu, \gamma) \sim \text{Normal} - \text{gamma distribution}$

According to Bayes Rule:

$$P(\mu, \gamma|D) = \frac{P(\mu, \gamma)P(D|\mu, \gamma)}{P(D)}$$

Since μ and γ are conditionally dependent variables: $P(\mu, \gamma) = P(\gamma)P(\mu|\gamma)$

$P(D)$ is constant, so: $P(\mu, \gamma|D) \propto P(\gamma)P(\mu|\gamma)P(D|\mu, \gamma)$ (2.1)

$$\gamma \sim \text{Gamma}(\alpha, \beta) \Rightarrow P(\gamma) = \frac{1}{\Gamma(\alpha)}\beta^\alpha\gamma^{\alpha-1}\exp(-\beta\gamma)$$

$$\mu|\gamma \sim \mathcal{N}\left(\mu_0, \frac{1}{\gamma_0\gamma}\right) \Rightarrow P(\mu|\gamma) = \frac{1}{\sqrt{2\pi \times \frac{1}{\gamma_0\gamma}}}\exp\left[-\frac{(\mu - \mu_0)^2}{2 \times \frac{1}{\gamma_0\gamma}}\right]$$

$$= \sqrt{\frac{\gamma_0\gamma}{2\pi}}\exp\left[-\frac{\gamma_0\gamma}{2}(\mu - \mu_0)^2\right]$$

$$x_i|\mu, \gamma \sim \mathcal{N}\left(\mu, \frac{1}{\gamma}\right) \Rightarrow P(x_i|\mu, \gamma) = \frac{1}{\sqrt{2\pi \times \frac{1}{\gamma}}}\exp\left[-\frac{(x_i - \mu)^2}{2 \times \frac{1}{\gamma}}\right]$$

$$= \sqrt{\frac{\gamma}{2\pi}}\exp\left[-\frac{\gamma}{2}(x_i - \mu)^2\right]$$

$$\Rightarrow P(D|\mu,\gamma) = \prod_{i=1}^{n} P(x_i|\mu,\gamma) = \left(\frac{\gamma}{2\pi}\right)^{\frac{n}{2}} \exp\left[-\frac{\gamma}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

Therefore, from (2.1):

$$P(\mu,\gamma|D) \propto \gamma^{\alpha-1}\exp(-\beta\gamma)\gamma^{\frac{1}{2}}\exp\left[-\frac{\gamma_0\gamma}{2}(\mu-\mu_0)^2\right]\gamma^{\frac{n}{2}}\exp\left[-\frac{\gamma}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

$$P(\mu,\gamma|D) \propto \gamma^{\alpha+\frac{n}{2}-\frac{1}{2}}e^{-\beta\gamma}\exp\left\{-\frac{\gamma}{2}\left[\gamma_0(\mu-\mu_0)^2+\sum_{i=1}^{n}(x_i-\mu)^2\right]\right\}$$

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i^2-2x_i\mu+\mu^2) = \sum_{i=1}^{n}x_i^2 - 2\mu\sum_{i=1}^{n}x_i + n\mu^2$$

Let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i\,, S = \sum_{i=1}^{n}x_i^2$

$$P(\mu,\gamma|D) \propto \gamma^{\alpha+\frac{n}{2}-\frac{1}{2}}e^{-\beta\gamma}exp\left\{-\frac{\gamma}{2}[\gamma_0(\mu^2-2\mu\mu_0+\mu_0^2)+S-2n\mu\bar{x}+n\mu^2]\right\}$$

$$\Rightarrow P(\mu,\gamma|D) \propto \gamma^{\alpha+\frac{n}{2}-\frac{1}{2}}e^{-\beta\gamma}\exp\left\{-\frac{\gamma}{2}[(n+\gamma_0)\mu^2-2\mu(n\bar{x}+\mu_0\gamma_0)+\gamma_0\mu_0^2+S]\right\}$$

$$\Rightarrow P(\mu,\gamma|D) \propto \gamma^{\alpha+\frac{n}{2}-\frac{1}{2}}e^{-\beta\gamma}\exp\left\{-\frac{\gamma}{2}(n+\gamma_0)\left[\mu^2-2\mu\frac{n\bar{x}+\mu_0\gamma_0}{n+\gamma_0} + \left(\frac{n\bar{x}+\mu_0\gamma_0}{n+\gamma_0}\right)^2\right]\right\}$$

$$\Rightarrow P(\mu,\gamma|D) \propto \gamma^{\alpha+\frac{n}{2}-\frac{1}{2}}e^{-\beta\gamma}exp\left\{-\frac{\gamma}{2}(n+\gamma_0)\left(\mu-\frac{n\bar{x}+\mu_0\gamma_0}{n+\gamma_0}\right)^2\right\}$$

Finally:

$$P(\mu,\gamma|D) = \frac{\beta^{\alpha}\sqrt{n+\gamma_0}}{\Gamma(\alpha)\sqrt{2\pi}}\gamma^{\alpha+\frac{n}{2}-\frac{1}{2}}e^{-\beta\gamma}e^{-\frac{(n+\gamma_0)\gamma}{2}\left(\mu-\frac{n\bar{x}+\mu_0\gamma_0}{n+\gamma_0}\right)^2}$$

$$P(\mu,\gamma|D) = \mathrm{Normal-Gamma}\left(\mathrm{mean} = \frac{n\bar{x}+\mu_0\gamma_0}{n+\gamma_0}, \mathrm{variance} = \frac{1}{n+\gamma_0}, \alpha, \beta\right)$$

iv. Problem 2-iv

Exponential family: $P(x|\eta) = h(x)g(\eta)e^{\eta^T u(x)}$

Conjugate prior: $p(\eta|x,\nu) = f(x,\nu)g(\eta)^{\nu}e^{\nu\eta^T x}$

- Poisson distribution

$$P(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!} e^{x \ln \lambda} e^{-\lambda} = \frac{1}{x!} e^{-\lambda} e^{x \ln \lambda}$$

$$\eta = \ln \lambda, \qquad u(x) = x, \qquad h(x) = \frac{1}{x!}, \qquad g(\eta) = e^{-e^{\eta}}$$

Conjugate prior:

$$p(\eta|x,\nu) = f(x,\nu) g(\eta)^{\nu} e^{\nu \eta^T x} = f(x,\nu) e^{-\lambda\nu} e^{\nu x \ln \lambda} = f(x,\nu) \lambda^{\nu x} e^{-\lambda\nu}$$

$$= f(x,\nu) \lambda^{\nu x} e^{-\nu\lambda} \propto \mathrm{Gamma}(\alpha, \beta)$$

Where: $\alpha = 1 + \nu x,\ \beta = \nu$

- ❖ Multinomial distribution

$$P(x_1, x_2, \dots, x_K|\mu, N) = \frac{N!}{x_1!\, x_2! \dots x_K!} \prod_{k=1}^{K} \mu_k^{x_k} = \frac{N!}{x_1!\, x_2! \dots x_K!} \prod_{k=1}^{K} e^{x_k \ln \mu_k}$$

$$= \frac{N!}{x_1!\, x_2! \dots x_K!} \exp\left(\sum_{k=1}^{K} x_k \ln \mu_k\right) = \frac{N!}{x_1!\, x_2! \dots x_K!} \exp\left(\begin{bmatrix} \ln \mu_1 \\ \vdots \\ \ln \mu_K \end{bmatrix}^T \begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix}\right)$$

$$\eta = \begin{bmatrix} \ln \mu_1 \\ \vdots \\ \ln \mu_K \end{bmatrix}, \qquad u(x) = \begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix}, \qquad h(x) = \frac{N!}{x_1!\, x_2! \dots x_K!}, \qquad g(\eta) = 1$$

Conjugate prior:

$$p(\eta|x,\nu) = f(x,\nu) g(\eta)^{\nu} e^{\nu \eta^T x} = f(x,\nu) \exp\left\{\nu \begin{bmatrix} \ln \mu_1 \\ \vdots \\ \ln \mu_K \end{bmatrix}^T x\right\}$$

$$= f(x,\nu) \prod_{k=1}^{K} \mu_k^{\nu x_k} \propto \mathrm{Dir}(\mu, \alpha)$$

Where: $\alpha_k = 1 + \nu x_k$

- ❖ Laplace distribution

$$P(x|\mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} = \frac{1}{2b} e^{-\frac{1}{b}|x-\mu|}$$

$$\eta = -\frac{1}{b}, \qquad u(x) = |x - \mu|, \qquad h(x) = 1, \qquad g(\eta) = -\frac{\eta}{2}$$

Conjugate prior:

$$p(\eta|x,\nu) = f(x,\nu) g(\eta)^{\nu} e^{\nu \eta^T x} = f(x,\nu) \left(-\frac{\eta}{2}\right)^{\nu} e^{\nu\left(-\frac{1}{b}\right)x} = f(x,\nu) \left(\frac{1}{2b}\right)^{\nu} e^{-\frac{\nu x}{b}}$$

$$= f(x,\nu) \left(\frac{1}{2}\right)^{\nu} b^{-\nu} e^{-\frac{\nu x}{b}} \propto \mathrm{Inverse\_Gamma}(\alpha, \beta)$$

Where: $\alpha = 1 + \nu,\ \beta = \nu x$

❖ Dirichlet distribution

Where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$

$$P(x|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\dots\Gamma(\alpha_K)} \prod_{k=1}^{K} x_k^{\alpha_k - 1} = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\dots\Gamma(\alpha_K)} \prod_{k=1}^{K} x_k^{-1} \prod_{k=1}^{K} x_k^{\alpha_k}$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\dots\Gamma(\alpha_K)} \prod_{k=1}^{K} x_k^{-1} \exp\left(\sum_{i=1}^{K} \alpha_k \ln x_k\right)$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\dots\Gamma(\alpha_K)} \prod_{k=1}^{K} x_k^{-1} \exp\left(\begin{bmatrix}\alpha_1\\ \vdots \\ \alpha_K\end{bmatrix}^T \begin{bmatrix}\ln x_1\\ \vdots \\ \ln x_K\end{bmatrix}\right)$$

$$\eta = \begin{bmatrix}\alpha_1\\ \vdots \\ \alpha_K\end{bmatrix}, \qquad u(x) = \begin{bmatrix}\ln x_1\\ \vdots \\ \ln x_K\end{bmatrix}, \qquad h(x) = \prod_{k=1}^{K} x_k^{-1}, \qquad g(\eta) = \frac{\Gamma(\sum_{k=1}^{K} \eta_k)}{\prod_{k=1}^{K} \Gamma(\eta_k)}$$

Conjugate prior:

$$p(\eta|x,\nu) = f(x,\nu) g(\eta)^\nu e^{\nu \eta^T x} = f(x,\nu) \left[\frac{\Gamma(\sum_{k=1}^{K} \eta_k)}{\prod_{k=1}^{K} \Gamma(\eta_k)}\right]^\nu \exp\left\{\nu \begin{bmatrix}\alpha_1\\ \vdots \\ \alpha_K\end{bmatrix}^T x\right\}$$

$$= f(x,\nu) \left[\frac{1}{B(\alpha)}\right]^\nu \exp\left(\sum_{k=1}^{K} \nu x_k \alpha_k\right) \propto CD(\alpha|x,\nu) \propto \left[\frac{1}{B(\alpha)}\right]^\nu \exp\left(\sum_{k=1}^{K} \nu x_k \alpha_k\right)$$

❖ Gamma distribution

$$P(x|a,b) = \frac{1}{\Gamma(a)} b^a x^{a-1} e^{-bx} = \frac{b^a}{\Gamma(a)} e^{(a-1)\ln x} e^{-bx} = \frac{b^a}{\Gamma(a)} \exp[(a-1)\ln x - bx]$$

$$= \frac{b^a}{\Gamma(a)} \exp\left(\begin{bmatrix}a-1\\ -b\end{bmatrix}^T \begin{bmatrix}\ln x\\ x\end{bmatrix}\right)$$

$$\eta = \begin{bmatrix}a-1\\ -b\end{bmatrix}, \qquad u(x) = \begin{bmatrix}\ln x\\ x\end{bmatrix}, \qquad h(x) = 1, \qquad g(\eta) = \frac{b^a}{\Gamma(a)}$$

Conjugate prior (with known shape a):

$$p(\eta|x,\nu) = f(x,\nu) g(\eta)^\nu e^{\nu \eta^T x} = f(x,\nu) \left[\frac{b^a}{\Gamma(a)}\right]^\nu \exp\left\{\nu \begin{bmatrix}a-1\\ -b\end{bmatrix}^T x\right\}$$

$$= f(x,\nu) \left[\frac{1}{\Gamma(a)}\right]^\nu b^{a\nu} e^{\nu(a-1)x - \nu b x} \propto b^{a\nu} e^{-\nu x b} \propto Gamma(\alpha,\beta)$$

Where: $\alpha = 1 + av$, $\beta = vx$

Table 1.1. Summary of exponential family distributions and their conjugate prior

Distribution	$\boldsymbol{\eta}$	$\boldsymbol{u}(\boldsymbol{x})$	$\boldsymbol{h}(\boldsymbol{x})$	$\boldsymbol{g}(\boldsymbol{\eta})$	**Conjugate prior**
Poisson	$\ln \lambda$	x	$\frac{1}{x!}$	$e^{-e^{\eta}}$	Gamma
Multinomial	$\begin{bmatrix} \ln \mu_1 \\ \vdots \\ \ln \mu_K \end{bmatrix}$	$\begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix}$	$\frac{N!}{x_1!\,x_2!\dots x_K!}$,	1	Dirichlet
Laplace	$-\frac{1}{b}$	$\lvert x-\mu \rvert$	1	$-\frac{\eta}{2}$	Inverse Gamma
Dirichlet	$\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{bmatrix}$	$\begin{bmatrix} \ln x_1 \\ \vdots \\ \ln x_K \end{bmatrix}$	$\prod_{k=1}^{K} x_k^{-1}$	$\frac{\Gamma(\sum_{k=1}^{K} \eta_k)}{\prod_{k=1}^{K} \Gamma(\eta_k)}$	Unknown
Gamma	$\begin{bmatrix} a-1 \\ -b \end{bmatrix}$	$\begin{bmatrix} \ln x \\ x \end{bmatrix}$	1	$\frac{b^a}{\Gamma(a)}$	Gamma

3. Decision Theory

$$y \sim Bernoulli\left(\frac{1}{2}\right) \Rightarrow P(y) = \left(\frac{1}{2}\right)^{y}\left(1-\frac{1}{2}\right)^{1-y} = \frac{1}{2} \Rightarrow P(y=1) = P(y \neq 1) = \frac{1}{2}$$

$$P(x|y=1) \sim Bernoulli(p); \quad P(x|y \neq 1) \sim Bernoulli(q)$$

$$\begin{aligned} P(x=1|D) &= P(x=1|y=1)P(y=1|D) + P(x=1|y \neq 1)P(y \neq 1|D) \\ &= P(x=1|y=1)P(y=1) + P(x=1|y \neq 1)P(y \neq 1) \\ &= \frac{1}{2}p^1(1-p)^0 + \frac{1}{2}q^1(1-q)^0 = \frac{1}{2}(p+q) \end{aligned}$$

$$P(x=0|D) = 1 - \frac{1}{2}(p+q)$$

Expected loss or risk based on zero-one loss:

$$R = \sum_{x=0}^{1} P(x|D)P(x \neq \omega_m|D) = P(x \neq \omega_m|D)\sum_{x=0}^{1} P(x|D)$$

$$= [1 - P(x=\omega_m|D)][P(x=0|D) + P(x=1|D)] = 1 - P(x=\omega_m|D)$$

Where: ω_m is the major probable class.

- ❖ Case 1: $p+q > 1 \Rightarrow P(x=1|D) > \frac{1}{2} \Rightarrow P(x=1|D) > P(x=0|D)$. The major probable class is $x = 1$.

Expected loss or risk:

$$R = 1 - P(x = \omega_m|D) = 1 - P(x = 1|D) = 1 - \frac{1}{2}(p + q)$$

❖ Case 2: $0 < p + q < 1 \Rightarrow P(x = 0|D) > \frac{1}{2} \Rightarrow P(x = 0|D) > P(x = 1|D)$. The most probable class is $x = 0$.

Expected loss or risk:

$$R = 1 - P(x = \omega_m|D) = 1 - P(x = 0|D) = \frac{1}{2}(p + q)$$

4. Implementing Backpropagation Algorithm (Matlab Programming Assignment)

Figure 4.1 and Figure 4.2 displays the test accuracy and the performance of each epoch. In the feed-forward process, the chosen activation function is sigmoid function. For the back-propagation, the derivative components are calculated in two for-loops which presents to layer. All weights of each layer are updated concurrently by matrix operations.

```
Command Window
>> main_nn
epoch: 1  iteration: 10000  test acc: 0.8678
epoch: 2  iteration: 20000  test acc: 0.8935
epoch: 2  iteration: 30000  test acc: 0.9049
epoch: 3  iteration: 40000  test acc: 0.9146
epoch: 4  iteration: 50000  test acc: 0.9186
epoch: 4  iteration: 60000  test acc: 0.9217
epoch: 5  iteration: 70000  test acc: 0.9207
```

Figure 4.2. Test accuracy

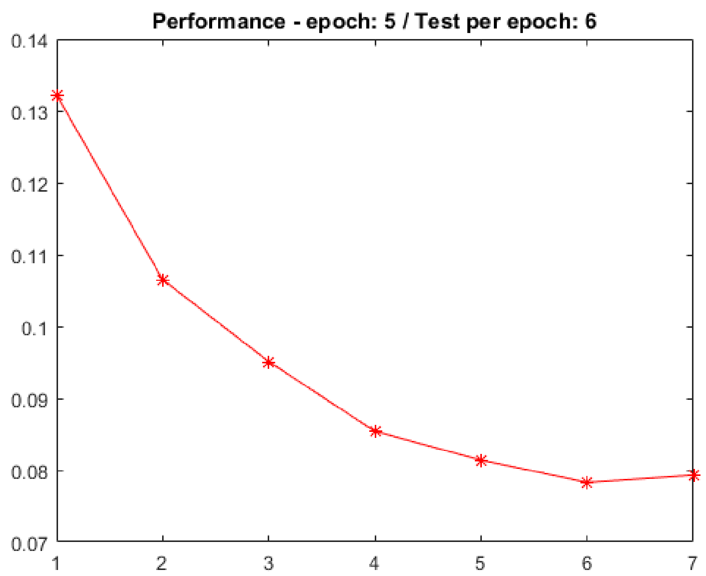

Figure 4.1. Performance each epoch

5. **Recurrent Neural Network**

i. **Problem 5-i**

$$h^{(1)} = f_\theta\left(h^{(0)}, x^{(1)}\right) = \tanh\left(Ux^{(1)} + Wh^{(0)} + b\right)$$

$$\begin{aligned} h^{(2)} = f_\theta\left(h^{(1)}, x^{(2)}\right) &= \tanh\left(Ux^{(2)} + Wh^{(1)} + b\right) \\ &= \tanh\left(Ux^{(2)} + W\tanh\left(Ux^{(1)} + Wh^{(0)} + b\right) + b\right) \end{aligned}$$

$$\begin{aligned} h^{(3)} = f_\theta\left(h^{(2)}, x^{(3)}\right) &= \tanh\left(Ux^{(3)} + Wh^{(2)} + b\right) \\ &= \tanh\left(Ux^{(3)} + W\tanh\left(Ux^{(2)} + Wh^{(1)} + b\right) + b\right) \\ &= \tanh\left(Ux^{(3)} + W\tanh\left(Ux^{(2)} + W\tanh\left(Ux^{(1)} + Wh^{(0)} + b\right) + b\right) \right. \\ &\quad \left. + b\right) \end{aligned}$$

…

$$\begin{aligned} h^{(\tau)} = \tanh&\left(Ux^{(\tau)}\right. \\ &+ W\tanh\left(Ux^{(\tau-1)}\right. \\ &+ W\tanh\left(Ux^{(\tau-2)}\right. \\ &+ W\tanh\left(Ux^{(\tau-3)} + \cdots + W\tanh\left(Ux^{(1)} + Wh^{(0)} + b\right) + \cdots + b\right) + b\right) \\ &\left.\left. + b\right) + b\right) \end{aligned}$$

ii. **Problem 5-ii**

Cross entropy:

$$L = -\sum_t \sum_i y_i^{(t)} \log \hat{y}_i^{(t)}$$

Where:

$$\hat{y}^{(t)} = \operatorname{softmax}\left(o^{(t)}\right); \quad o^{(t)} = Vh^{(t)} + c; \quad h^{(t)} = \tanh\left(Ux^{(t)} + Wh^{(t-1)} + b\right)$$

Let $L^{(t)} = y_i^{(t)} \log \hat{y}_i^{(t)}$

$$\begin{aligned} \nabla_V L = \frac{\partial L}{\partial V} = \frac{\partial}{\partial V}\left(-\sum_t \sum_i y_i^{(t)} \log \hat{y}_i^{(t)}\right) &= -\sum_t \sum_i \frac{\partial}{\partial V}\left(y_i^{(t)} \log \hat{y}_i^{(t)}\right) \\ &= -\sum_t \sum_i \frac{\partial L}{\partial o_i^{(t)}} \frac{\partial o_i^{(t)}}{\partial V} \end{aligned}$$

$$\frac{\partial L}{\partial o_i^{(t)}} = \frac{\partial L}{\partial L^{(t)}} \times \frac{\partial L^{(t)}}{\partial o_i^{(t)}} = 1 \times y_i^{(t)}\left(\hat{y}_i^{(t)} - 1\right) = y_i^{(t)}\left(\hat{y}_i^{(t)} - 1\right)$$

$$\frac{\partial o_i^{(t)}}{\partial V} = \frac{\partial}{\partial V}\left(Vh_i^{(t)} + c\right) = h_i^{(t)}$$

$$\frac{\partial L}{\partial V} = -\sum_t \sum_i y_i^{(t)} \left(\hat{y}_i^{(t)} - 1\right) h_i^{(t)} = \sum_t \sum_i y_i^{(t)} \left(1 - \hat{y}_i^{(t)}\right) h_i^{(t)}$$
$$= \sum_t y^{(t)} \operatorname{diag}\left(1 - \hat{y}^{(t)}\right) \left(h^{(t)}\right)^T = \sum_t \left(\nabla_{o^{(t)}} L\right) \left(h^{(t)}\right)^T$$

$$\nabla_W L = \frac{\partial L}{\partial W} = \sum_t \sum_i \frac{\partial L}{\partial h_i^{(t)}} \frac{\partial h_i^{(t)}}{\partial W} = \sum_t \operatorname{diag}\left(\mathbb{I} - h^{(t)} \left(h^{(t)}\right)^T\right) \left(\nabla_{h^{(t)}} L\right) \left(h^{(t-1)}\right)^T$$

iii. Problem 5-iii

$$f_\theta\left(h^{(0)}, x^{(1)}\right) = \tanh\left(Ux^{(1)} + Wh^{(0)} + b\right)$$

$$f_\theta\left(h^{(1)}, x^{(2)}\right) = \tanh\left(Ux^{(2)} + Wh^{(1)} + b\right)$$
$$= \tanh\left(Ux^{(2)} + W \tanh\left(Ux^{(1)} + Wh^{(0)} + b\right) + b\right)$$

$$f_\theta\left(h^{(2)}, x^{(3)}\right) = \tanh\left(Ux^{(3)} + Wh^{(2)} + b\right)$$
$$= \tanh\left(Ux^{(3)} + W \tanh\left(Ux^{(2)} + Wh^{(1)} + b\right) + b\right)$$
$$= \tanh\left(Ux^{(3)} + W \tanh\left(Ux^{(2)} + W \tanh\left(Ux^{(1)} + Wh^{(0)} + b\right) + b\right)$$
$$+ b\right)$$

…

$$f_\theta\left(h^{(\tau)}, x^{(\tau)}\right) = \tanh\left(Ux^{(\tau)}\right.$$
$$+ W \tanh\left(Ux^{(\tau-1)}\right.$$
$$+ W \tanh\left(Ux^{(\tau-2)}\right.$$
$$+ W \tanh\left(Ux^{(\tau-3)} + \cdots + W \tanh\left(Ux^{(1)} + Wh^{(0)} + b\right) + \cdots + b\right) + b\right)$$
$$\left. + b\right) + b\right)$$

II. PART 2

The results of PCA (Matlab Programming Assignment) are shown in Table 2.1 – 2.2 and Figure 2.1 – 2.4. Table 2.1 presents reconstruction error in each case k = 5, 50, 200 and 500. The top 5 principal components (PCs) are displayed from Figure 2.1 to Figure 2.4. Table 2.2 presents specifically the prediction about emotion of each test image in 4 cases of k.

Table 2.1. Reconstruction Error

k	5	50	200	500
Reconstruction Error	2.1290×10^{6}	2.5866×10^{5}	4.0246×10^{-22}	3.9890×10^{-22}

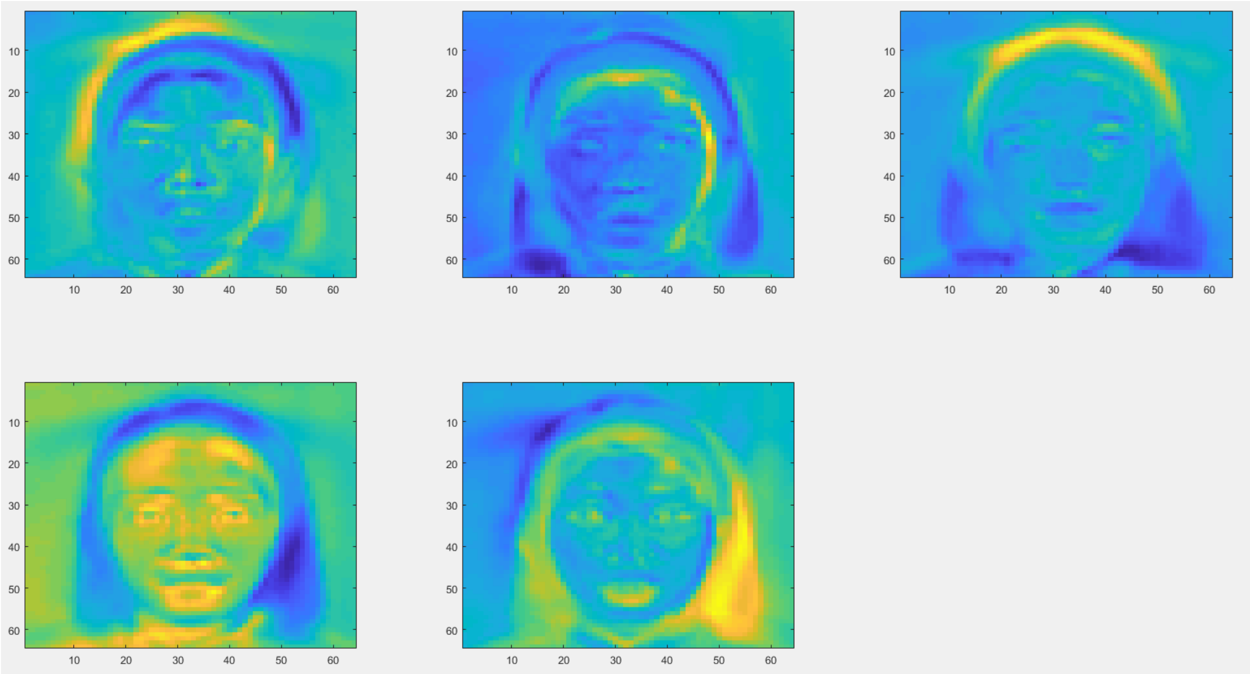

Figure 2.1. The top 5 PCs with k = 5

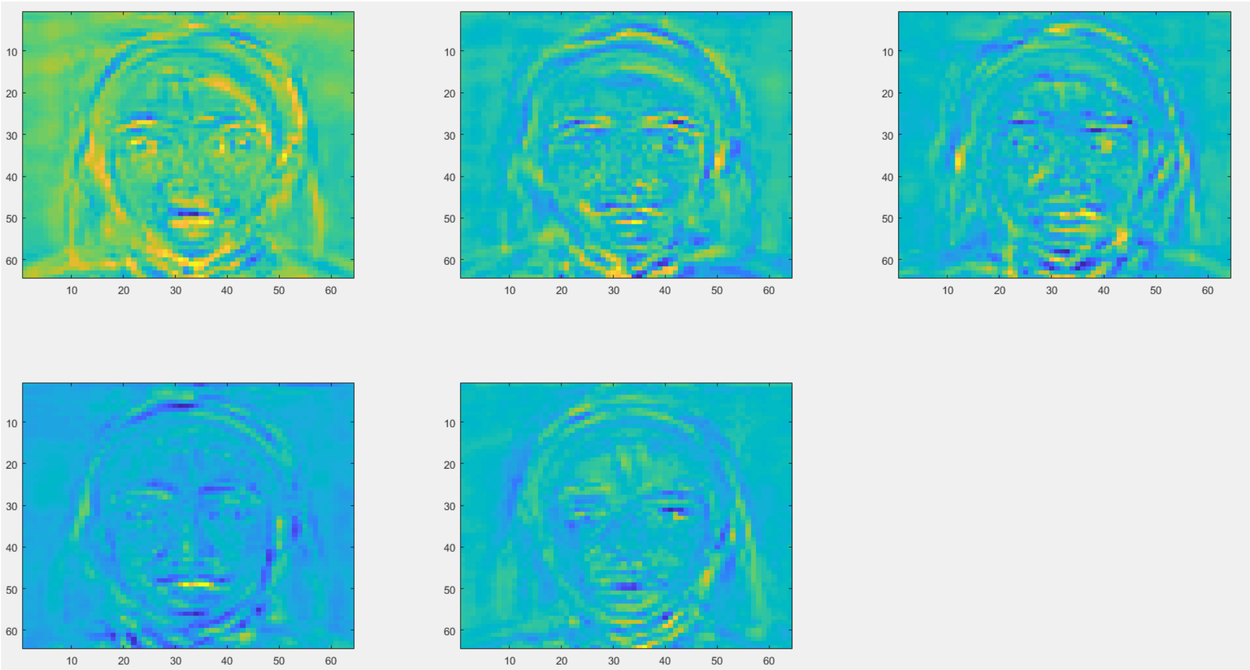

Figure 2.2. The top 5 PCs with k = 50

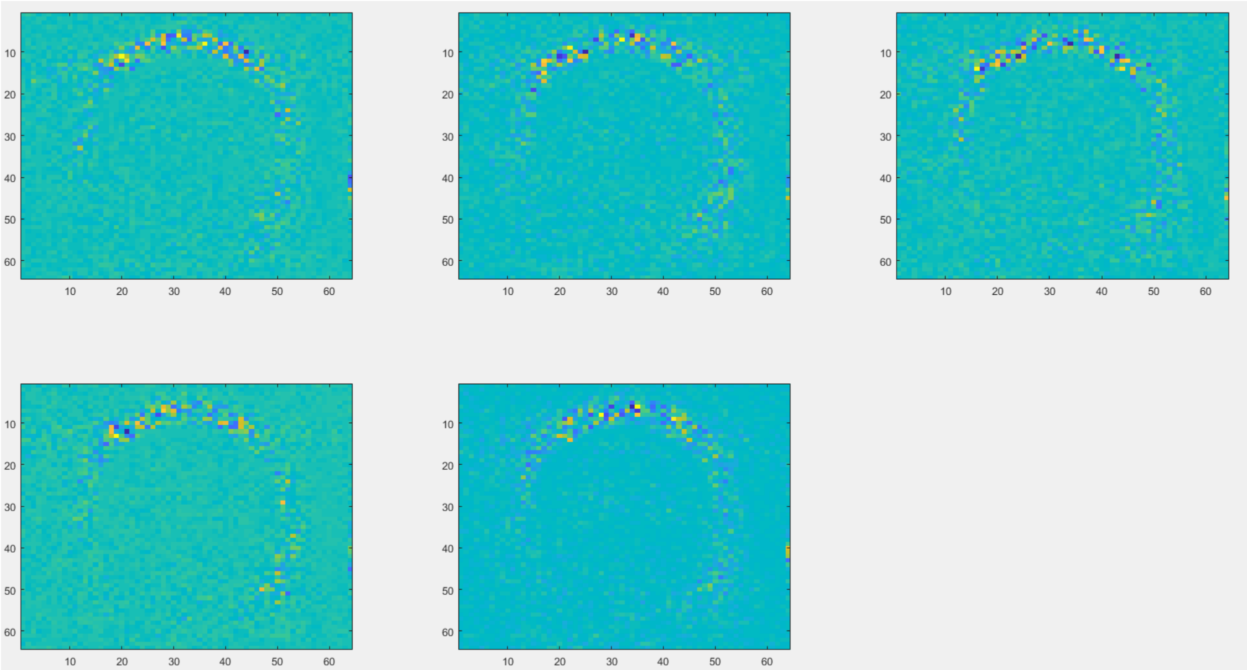

Figure 2.3. The top 5 PCs with k = 200

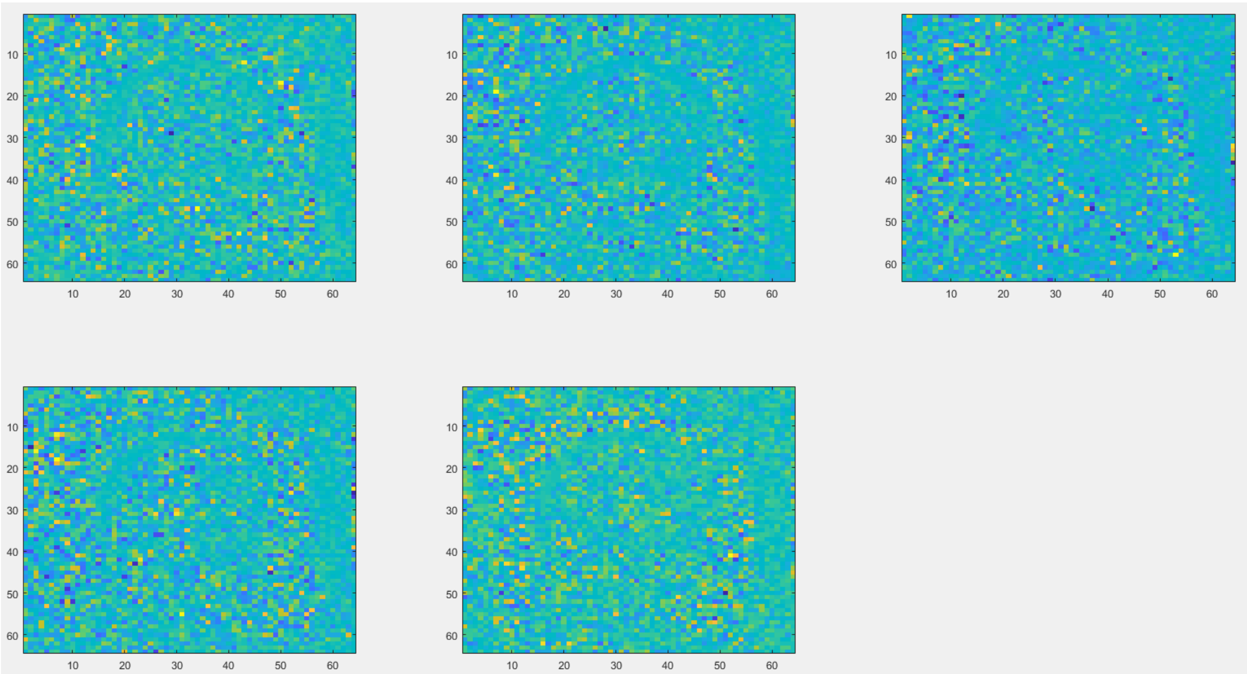

Figure 2.4. The top 5 PCs with k = 500

Table 2.2. The prediction results about emotion

ID of test image	k				Human eyes	Incorrect
	5	50	200	500		
1	7	18	18	18	Not exist	All
2	3	3	3	3		
3	5	5	5	5		
4	9	9	9	9		

ID of test image	k				Human eyes	Incorrect
	5	**50**	**200**	**500**		
5	14	14	14	14		
6	19	19	19	19		
7	22	21	22	22	21, 22	
8	28	27	27	27	27, 28	
9	30	30	30	30		
10	35	35	35	35		
11	36	36	36	36		
12	44	44	44	44		
13	46	46	46	46		
14	48	49	49	49	49	k = 5
15	52	52	52	52		
16	55	55	55	55		
17	56	55	55	55	55, 56	
18	57	57	57	57		
19	67	67	67	67		
20	57	67	67	67	67	k = 5
21	68	68	68	68		
22	68	68	68	68		
23	78	79	79	79	78, 79	
24	77	81	81	81	81	k = 5
25	73	83	83	83	83	k = 5
26	85	85	85	85		
27	94	94	94	94	92	All
28	94	92	92	92	92	k = 5
29	93	94	94	94	93, 94	
30	98	96	96	96	96	k = 5
31	100	100	100	100		
32	100	100	100	100		
33	101	101	101	101		
34	113	113	113	113		
35	103	108	108	108	108	k = 5
36	116	116	115	115	115	k = 5, 50
37	118	118	118	118		
38	119	119	119	119	119, 120	
39	125	125	124	124	124, 125	
40	129	129	129	129		
41	137	134	134	134	134	k = 5
42	144	140	140	140	140	
43	144	144	144	144		
44	144	144	144	144		
45	145	145	145	145		
46	148	148	148	148		

ID of test image	k				Human eyes	Incorrect
	5	**50**	**200**	**500**		
47	154	149	149	149	149	k = 5
48	154	154	154	154	150	All
49	151	151	152	152	152	k = 5, 50
50	154	154	154	154		

Generally, all cases predict correctly about the face of person in the test images. However, the emotion prediction of the case k = 200 and 500 are the best. While the case k = 5 gives us the worst emotion prediction. There are 3 test images that all cases of k predicts emotion wrong and their ID equals to 1, 27 and 48. Conclusion, the PCA algorithm operates very well on face recognition but in emotion detection, it required improvement.