COMS 4771 Machine Learning (Spring 2015)
Problem Set #4

Jingwei Yang - jy2653@columbia.edu
Discussants: pp2526,yd2300

May 3, 2015

Problem 1

(a)

$$\mathbb{E}[(Y-\hat{Y})^2] = \mathbb{E}[(Y-\langle \boldsymbol{w}, \boldsymbol{X}\rangle)^2]$$

Take a gradient of $\boldsymbol{w}$

$$\nabla_{\boldsymbol{w}}\mathbb{E}[(Y-\langle \boldsymbol{w}, \boldsymbol{X}\rangle)^2] = \mathbb{E}[2(Y-\langle \boldsymbol{w}, \boldsymbol{X}\rangle)\boldsymbol{X}^T]$$

To minimize $\mathbb{E}[(Y-\hat{Y})^2]$, the $\nabla_{\boldsymbol{w}}\mathbb{E}[(Y-\langle \boldsymbol{w}, \boldsymbol{X}\rangle)^2]$ should equal to 0, then we have

$$\begin{aligned}\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{X}^T &= Y\boldsymbol{X}^T\\ \boldsymbol{w}^T &= (Y\boldsymbol{X}^T)(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\\ \boldsymbol{w} &= (\boldsymbol{X}\boldsymbol{X}^T)^{-1}(\boldsymbol{X}Y)\end{aligned}$$

(b)

$$\begin{aligned}\mathbb{E}[(Y-\langle \boldsymbol{w}, \boldsymbol{X}\rangle)\boldsymbol{X}] &= \mathbb{E}[(Y-\boldsymbol{w}^T\boldsymbol{X})\boldsymbol{X}]\\ &= \mathbb{E}[(Y-Y\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X})\boldsymbol{X}]\\ &= \mathbb{E}[(Y-Y\boldsymbol{X}^T(\boldsymbol{X}^T)^{-1}\boldsymbol{X}^{-1}\boldsymbol{X})\boldsymbol{X}]\\ &= \boldsymbol{0}\end{aligned}$$

(c)

$$\begin{aligned}\mathbb{E}[Z_i\boldsymbol{X}_{(-i)}] &= \mathbb{E}[X_i\boldsymbol{X}_{(-i)} - E(X_i\boldsymbol{X}_{(-i)})^T E(\boldsymbol{X}_{(-i)}\boldsymbol{X}_{(-i)}^T)^{-1}\boldsymbol{X}_{(-i)}\boldsymbol{X}_{(-i)}]\\ &= \mathbb{E}[X_i\boldsymbol{X}_{(-i)} - X_i\boldsymbol{X}_{(-i)}^T(\boldsymbol{X}_{(-i)}^T)^{-1}\boldsymbol{X}_{(-i)}^{-1}\boldsymbol{X}_{(-i)}\boldsymbol{X}_{(-i)}]\\ &= \boldsymbol{0}\end{aligned}$$

(d)

$$
\begin{aligned}
\mathbb{E}[Z_i^2] &= \mathbb{E}[Z_iX_i - Z_iX_i\boldsymbol{X}_{(-i)}^T(\boldsymbol{X}_{(-i)}\boldsymbol{X}_{(-i)}^T)^{-1}\boldsymbol{X}_{(-i)}] \\
&= \mathbb{E}[Z_iX_i - X_i\boldsymbol{X}_{(-i)}^T(\boldsymbol{X}_{(-i)}\boldsymbol{X}_{(-i)}^T)^{-1}Z_i\boldsymbol{X}_{(-i)}] \\
&= \mathbb{E}[Z_iX_i - X_i\boldsymbol{X}_{(-i)}^T(\boldsymbol{X}_{(-i)}\boldsymbol{X}_{(-i)}^T)^{-1}\mathbf{0}] \\
&= \mathbb{E}[Z_iX_i]
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[\langle \boldsymbol{w}, \boldsymbol{X}\rangle Z_i] &= \mathbb{E}[(\langle \boldsymbol{w}_{(-i)}, \boldsymbol{X}_{(-i)}\rangle + w_iX_i)Z_i] \\
&= \mathbb{E}[Z_i\boldsymbol{X}_{(-i)}\boldsymbol{w}_{(-i)}^T + w_iX_iZ_i] \\
&= \mathbb{E}[0 + w_iX_iZ_i] \\
&= w_i\mathbb{E}[Z_iX_i]
\end{aligned}
$$

$$
\mathbb{E}[(Y-\hat{Y})Z_i] = \mathbb{E}[(Y-\hat{Y})X_i - (Y-\hat{Y})X_i\boldsymbol{X}_{(-i)}^T(\boldsymbol{X}_{(-i)}\boldsymbol{X}_{(-i)}^T)^{-1}\boldsymbol{X}_{(-i)}]
$$

According to part (b), we know $\mathbb{E}[(Y-\hat{Y})\boldsymbol{X}] = \mathbf{0}$, thus $\mathbb{E}[(Y-\hat{Y})X_i = 0$. Thus we have

$$
\mathbb{E}[(Y-\hat{Y})Z_i] = 0
$$

Since

$$
\mathbb{E}[YZ_i] = \mathbb{E}[\langle \boldsymbol{w}, \boldsymbol{X}\rangle Z_i + (Y-\hat{Y})Z_i]
$$

thus

$$
\mathbb{E}[YZ_i] = w_i\mathbb{E}[Z_iX_i] = \mathbb{E}[Z_i^2]w_i
$$

Problem 2

(a) As we can see from following comparision, the quantized image at k = 64 has better approximated performance than that at k = 8. Cause when k = 64, we could have more representative patches to choose from, thus the approximated performance is more refined.

image 24, k = 8

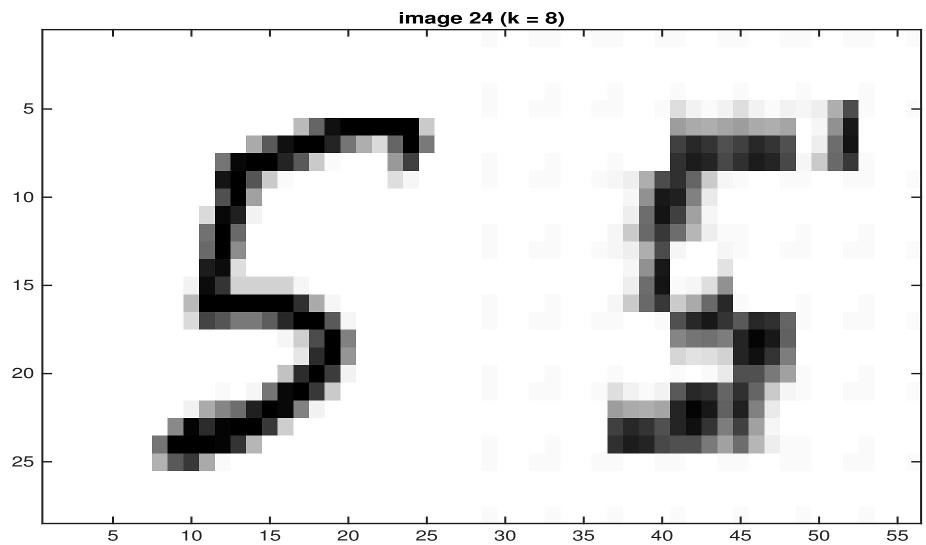


image 24, k = 64

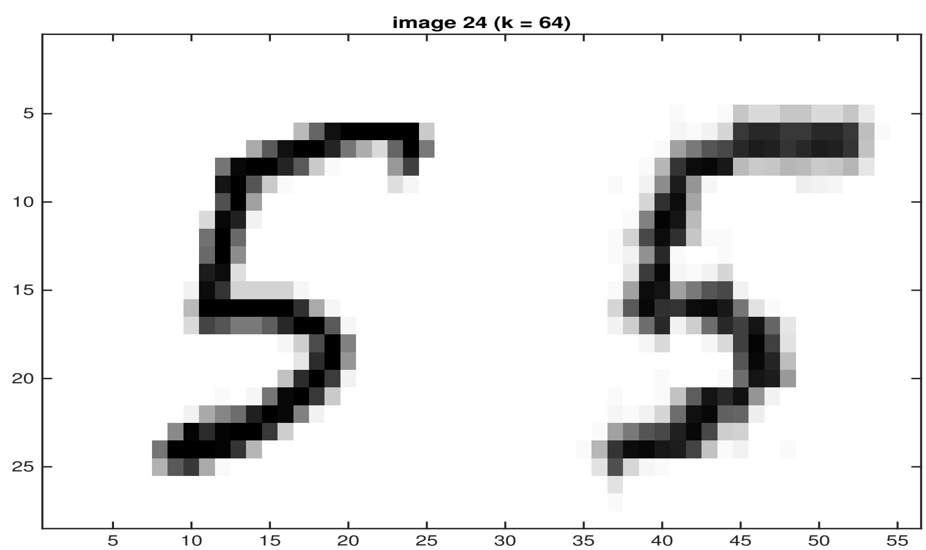


image 100, k = 8

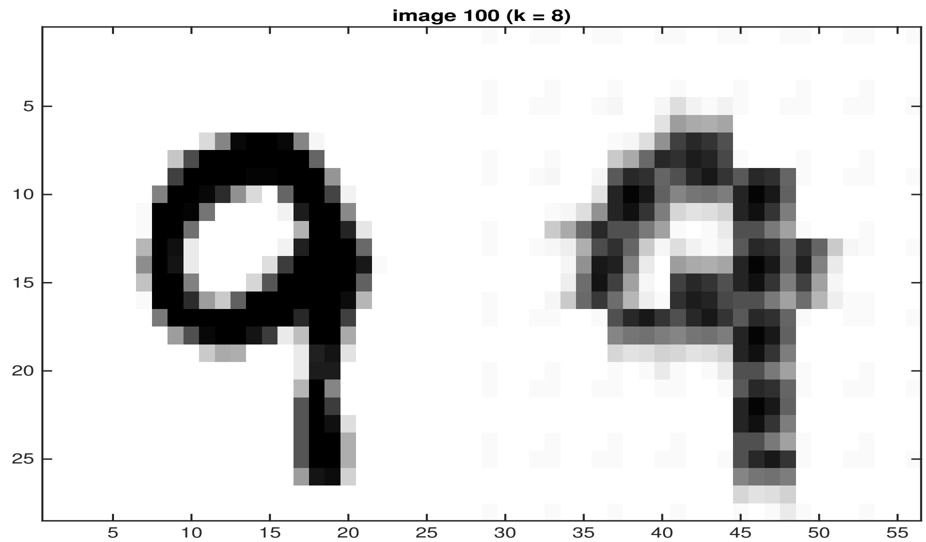


image 100, k = 64

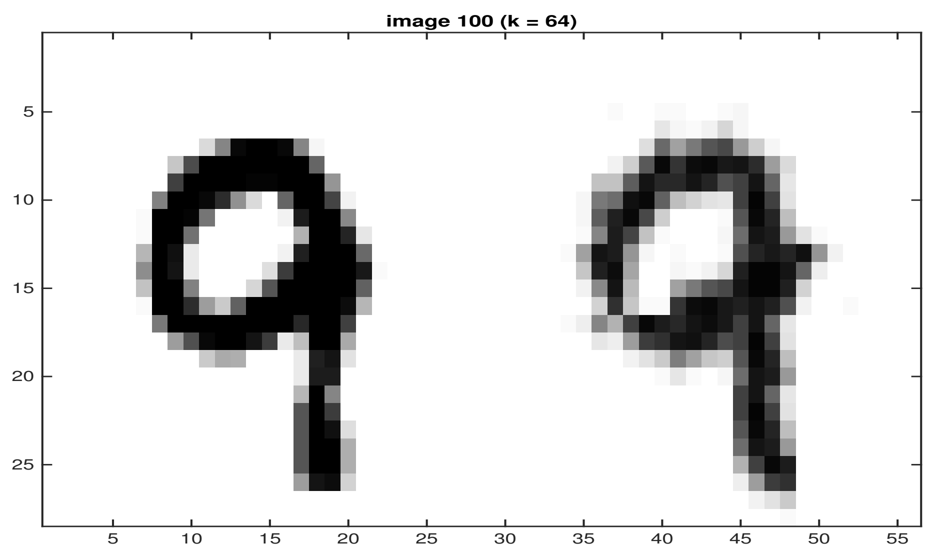


image 5000, k = 8

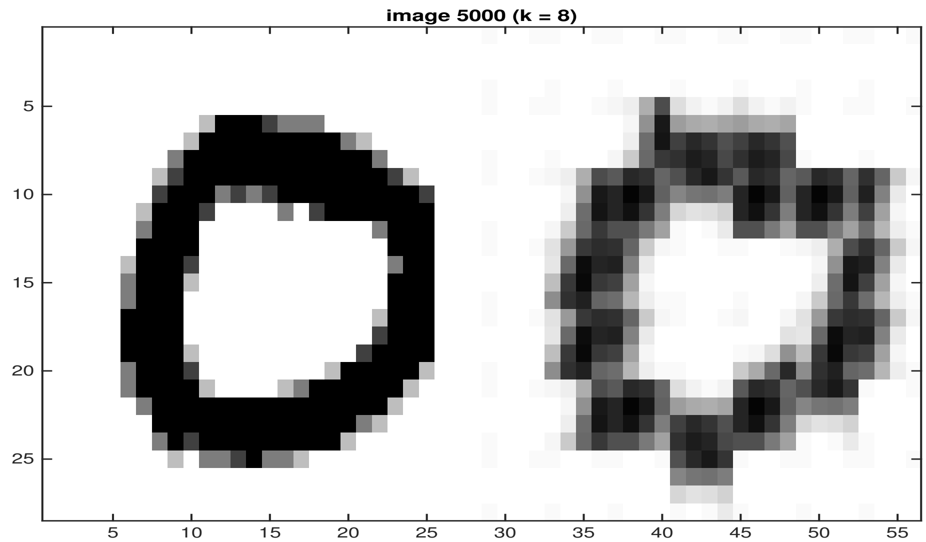
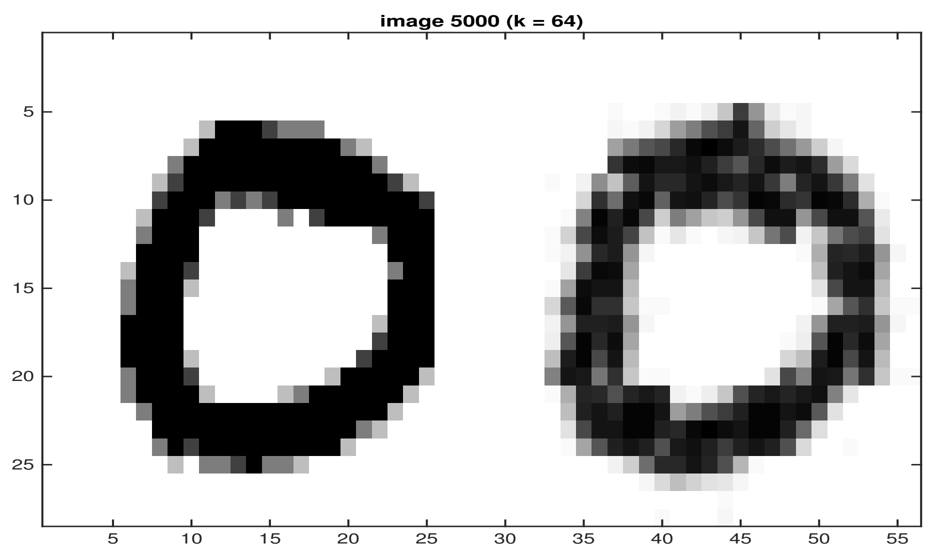


image 5000, k = 64



(b)

$f(k) = (16 * k + 10000 * 49) * 64$

Assume we use 64-bit integer numbers to record the indexes of each representative for each image.

Reasoning:

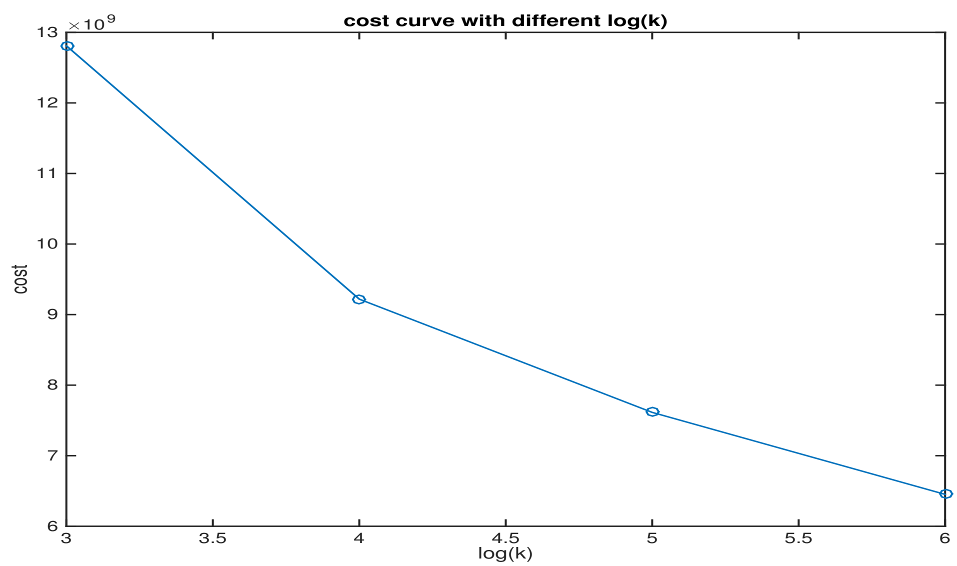
Each patch is a $4 * 4$ matrix, thus we need to use 16 double to record a patch (16 double-precision floating point numbers), $16 * k$ is the number of double-precision floating point numbers needed for recording all representative patches.

After quantization, we only need to record a $7 * 7$ matrix for each image, which is 49 64-bit integer numbers. Since we have 10000 test images in total, we need to use $10000 * 49$ 64-bit integer numbers to record all quantized images.

Since each double-precision floating point number use 64 bits, the total bits needed is $f(k) = (16 * k + 10000 * 49) * 64$.

(c)

cost curve for online K-means algorithm



Problem 3

sfdsdf
$\ell_{sq}(z) = (1 - z)^2$
$\ell_{log}(z) = \ln(1 + \exp(-z))$
$\ell_{\exp}(z) = \ln(\exp(-z))$

(a)
Since Y in $\{-1, +1\}$, $\ell_{sq}(Y\hat{y}) = (1 - Y\hat{y})^2$. When $Y = +1$, $\ell_{sq}(Y\hat{y}) = (1 - \hat{y})^2$, When $Y = -1$, $\ell_{sq}(Y\hat{y}) = (1 + \hat{y})^2$. And $Pr(Y = +1) = \eta$, $Pr(Y = -1) = 1 - \eta$. We have

$$\mathbb{E}[\ell_{sq}(Y\hat{y})] = \eta(1 - \hat{y})^2 + (1 - \eta)(1 + \hat{y})^2$$

Take the gradient of $\hat{y}$, we have

$$\begin{aligned}\nabla_{\hat{y}}\mathbb{E}[\ell_{sq}(Y\hat{y})] &= 2\eta(\hat{y} - 1) + 2(1 - \eta)(1 + \hat{y}) \\ &= 2\eta\hat{y} - 2\eta + 2 + 2\hat{y} - 2\eta - 2\eta\hat{y} \\ &= 2 - 4\eta + 2\hat{y}\end{aligned}$$

Since $\ell_{sq}(Y\hat{y}) = (1-Y\hat{y})^2$ is a convex function, to minimize $\mathbb{E}[\ell_{sq}(Y\hat{y})]$, we assign $\nabla_{\hat{y}}\mathbb{E}[\ell_{sq}(Y\hat{y})] = 0$, thus we have $\hat{y} = 2\eta - 1$

(b)
Since Y in $\{-1, +1\}$, $\ell_{log}(Y\hat{y}) = \ln(1 + \exp(-Y\hat{y}))$. When $Y = +1$, $\ell_{log}(Y\hat{y}) = \ln(1 + \exp(-y))$. When $Y = -1$, $\ell_{log}(Y\hat{y}) = \ln(1 + \exp(y))$. And $Pr(Y = +1) = \eta$, $Pr(Y = -1) = 1 - \eta$. We have

$$\mathbb{E}[\ell_{log}(Y\hat{y})] = \eta \ln(1 + \exp(-\hat{y})) + (1 - \eta) \ln(1 + \exp(\hat{y}))$$

Take the gradient of $\hat{y}$, we have

$$\nabla_{\hat{y}}\mathbb{E}[\ell_{log}(Y\hat{y})] = \eta\frac{-\exp(-\hat{y})}{1 + \exp(-\hat{y})} + (1 - \eta)\frac{\exp(\hat{y})}{1 + \exp(\hat{y})}$$

Since $\ell_{log}(Y\hat{y}) = \ln(1 + \exp(-Y\hat{y}))$ is a convex function , to minimize $\mathbb{E}[\ell_{log}(Y\hat{y})]$, we assign $\nabla_{\hat{y}}\mathbb{E}[\ell_{log}(Y\hat{y})] = 0$, thus we have

$$\frac{\eta \exp(-\hat{y})}{1 + \exp(-\hat{y})} = \frac{(1 - \eta) \exp(\hat{y})}{1 + \exp(\hat{y})}$$

$\Longrightarrow$

$$\eta \exp(-\hat{y}) + \eta = \exp(\hat{y}) - \eta \exp(\hat{y}) + 1 - \eta$$

$\Longrightarrow$

$$\frac{\eta \exp(\hat{y}) + \eta}{\exp(\hat{y})} = (\exp(\hat{y}) + 1)(1 - \eta)$$

$\Longrightarrow$

$$\eta = \exp(\hat{y})(1 - \eta)$$

$\Longrightarrow$

$$\hat{y} = \ln \frac{\eta}{1 - \eta}$$

Thus when $\hat{y} = \ln \frac{\eta}{1-\eta}$, $E[\ell_{log}(Y\hat{y})]$ is minized.

(c)
Since Y in $\{-1, +1\}$, $\ell_{\exp}(Y\hat{y}) = \ln(\exp(-Y\hat{y}))$. When $Y = +1$, $\ell_{log}(Y\hat{y}) = \ln(\exp(-\hat{y}))$. When $Y = -1$, $\ell_{log}(Y\hat{y}) = \ln(\exp(\hat{y}))$. And $Pr(Y = +1) = \eta$, $Pr(Y = -1) = 1 - \eta$. We have

$$\mathbb{E}[\ell_{\exp}(Y\hat{y})] = \eta \exp(-\hat{y}) + (1 - \eta) \exp(\hat{y})$$

Take the gradient of $\hat{y}$, we have

$$\nabla_{\hat{y}} \mathbb{E}[\ell_{\exp}(Y\hat{y})] = (1 - \eta) \exp(\hat{y}) - \eta \exp(-\hat{y})$$

Since $\ell_{\exp}(Y\hat{y}) = \ln(\exp(-Y\hat{y}))$ is a convex function, to minimize $\mathbb{E}[\ell_{\exp}(Y\hat{y})]$, we assign $\nabla_{\hat{y}} \mathbb{E}[\ell_{\exp}(Y\hat{y})] = 0$, thus we have

$$\frac{\eta}{\exp(\hat{y})} = \exp(\hat{y})(1 - \eta)$$

$\Longrightarrow$

$$\exp(2\hat{y}) = \frac{\eta}{1 - \eta}$$

$\Longrightarrow$

$$2\hat{y} = \ln \frac{\eta}{1 - \eta}$$

$\Longrightarrow$

$$\hat{y} = \frac{1}{2} \ln \frac{\eta}{1 - \eta}$$

Thus when $\hat{y} = \frac{1}{2} \ln \frac{\eta}{1-\eta}$, $E[\ell_{\exp}(Y\hat{y})]$ is minized.