Fourth Home Assignment
Data Analysis

Nikolaj Dybdahl Rathcke (rfq695)

May 26, 2015

Question 1

(1)

We want to prove that

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y] \tag{1}$$

We know that from definition that

$$\mathbb{E}[X] = \sum_{x\in\mathcal{X}} x p_X(x)$$

This means we can write

$$\begin{aligned}\mathbb{E}[X+Y] &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}(x+y)p_{XY}(x,y)\\ &= \sum_{x\in\mathcal{X}} x \sum_{y\in\mathcal{Y}} p_{XY}(x,y) + \sum_{y\in\mathcal{Y}} y \sum_{\in\mathcal{X}} p_{XY}(x,y)\\ &= \sum_{x\in\mathcal{X}} xp(x) + \sum_{y\in\mathcal{Y}} yp(y)\\ &= \mathbb{E}[X] + \mathbb{E}[Y]\end{aligned}$$

Which means we have proven equation 1.

(2)

We want to show, when X and Y are independent, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \tag{2}$$

We have the following definition

$$\mathbb{E}[X] = \sum_{x\in\mathcal{X}} x p_X(x)$$

This means we can write the following

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xyp_{XY}(x,y)\\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xyp_X(x)p_Y(y) && \rightarrow \text{using independence assumption}\\ &= \sum_{x\in\mathcal{X}} xp_X(x) \sum_{y\in\mathcal{Y}} yp_Y(y)\\ &= \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

which proves equation 2.

(3)

As an example, let X and Y be two cards drawn from the same standard deck of cards (without replacement).
In this example, for both X and Y then $\mathbb{E}\{\text{card is red}\} = \mathbb{E}\{\text{card is black}\} = \frac{1}{2}$.
The joint distribution of X and Y is

X\Y	R	B
R	25/102	26/102
B	26/102	25/102

We get that

$$\begin{aligned}\mathbb{E}[X = R, Y = R] &= 25/102\\ \mathbb{E}[X = R]\mathbb{E}[Y = R] &= 1/2 \cdot 1/2 = 1/4\end{aligned}$$

which means

$$\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$$

because the random variables are not independent of eachother.

(4)

We want to prove

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \tag{3}$$

The expected value of a random variable will be just some constant c. So $\mathbb{E}[X] = c$. An expected value of a constant is the constant itself, so $\mathbb{E}[c] = c$, thus we get

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X]] &= \mathbb{E}[c]\\ &= c\end{aligned}$$

which proves equation 3.

Question 2

(1)

We want to show that

$$\mathbb{P}\{N \text{ green balls are pulled in a row}\} \leq e^{-N\varepsilon} \tag{4}$$

The probability of pulling N green balls out in a row is the same as the chance of not drawing red balls N times. The probability of drawing a red ball when there are $2N$ balls is $2N\varepsilon/2N$, thus the chance of drawing a green ball is $1-(2N\varepsilon/2N)$. Since we are not replacing the balls, the probability for drawing a red ball increases each time we pull out a green one because the total number of balls decreases and so does the number of green ones. It can be expressed as

$$\prod_{i=0}^{N-1}\left(1-\frac{2N\varepsilon}{2N-i}\right)$$

We can rewrite equation 4

$$\prod_{i=0}^{N-1}\left(1-\frac{2N\varepsilon}{2N-i}\right) \leq e^{-N\varepsilon}$$

We use the inequality $1+x \leq e^x$ to get

$$\begin{aligned}\prod_{i=0}^{N-1}\left(1-\frac{2N\varepsilon}{2N-i}\right) &\leq \prod_{i=0}^{N-1} e^{-\frac{2N\varepsilon}{2N-i}}\\ &= e^{\sum_{i=0}^{N-1} -\frac{2N\varepsilon}{2N-i}}\\ &= e^{-2N\varepsilon\sum_{i=0}^{N-1}\frac{1}{2N-i}}\end{aligned}$$

The sum will always provide values in $[\frac{1}{2},1)$. This means that

$$\mathbb{P}\{N \text{ green balls are pulled in a row}\} \leq e^{-2N\varepsilon\sum_{i=0}^{N-1}\frac{1}{2N-i}} \leq e^{-2N\varepsilon} \leq e^{-N\varepsilon}$$

And thus, equation 4 is shown.

Question 3

(1)

We want to prove that

$$m_{\mathcal{H}}(2N) \leq m_{\mathcal{H}}(N)^2 \tag{5}$$

We let

$$m_{\mathcal{H}}(N) = c \tag{6}$$

If we partition a set of $2N$ points into sets of N points, then that means we get c dichotomies in each of these sets. These two sets can, at maximum produce dichotomies corresponding to the cross product of the two sets of dichotomies. That means the LHS of equation 5 has the upper bound

$$\begin{aligned} m_{\mathcal{H}}(2N) &\leq c \cdot c \\ &= c^2 \\ &= m_{\mathcal{H}}(N)^2 \end{aligned}$$

Last step using equation 6, and thus equation 5 is proved.

(2)

Equation 2.12 in [LFD] states that

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

Substituting the upper bound we found in (1) for $m_{\mathcal{H}}(2N)$, we can write a generalization bound that does not include this factor

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(N)^2}{\delta}}$$

This upper bound is a bit looser.

Question 4

(1)

We want to prove that

$$\sum_{i=0}^{d} \binom{N}{i} \leq N^d + 1 \tag{7}$$

when $d \leq N$ We begin by setting $d = 0$ and get

$$\begin{aligned} \sum_{i=0}^{0} \binom{N}{i} &\leq N^0 + 1 && \Leftrightarrow \\ 1 &\leq 1 + 1 \end{aligned}$$

which holds. Now we do our induction step, assuming equation 7 (hypothesis) holds for d we want to prove that it holds for $d+1$

$$\sum_{i=0}^{d+1} \binom{N}{i} \leq N^{d+1} + 1$$

for $d \geq 1$. We can rewrite this to

$$\begin{aligned} \sum_{i=0}^{d} \binom{N}{i} + \binom{N}{d+1} &\leq N^{d+1} + 1 \\ &= N^d + (N-1)N^d + 1 \end{aligned}$$

Using our hypothesis we can write

$$\begin{aligned} N^d + 1 + \binom{N}{d+1} &\leq N^d + (N-1)N^d + 1 && \Leftrightarrow \\ \binom{N}{d+1} &\leq (N-1)N^d \end{aligned}$$

The right hand side will always be larger since we are only looking at $d \geq 1$. Therefore equation 7 holds.

(2)

From [LFD] equation (2.9):

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d} \binom{N}{i}$$

This means we can derive the following bound

$$m_{\mathcal{H}}(N) \leq N^d + 1$$

using our result from (1).

(3)

Equation 2.12 in [LFD]:

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

Substituting the result in the equation yields

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4(2N)^d + 1)}{\delta}}$$

The coefficient d should be relatively small compared to N so the bound is not too loose. The smaller d compared to N, the tighter the bound.

Question 5

(1)

We implement k-means clustering in Matlab, one thing differing to the standard implementation of this is that, instead of seeing if the centroids are exactly alike, we see if the sum of their differences is sufficiently small > 0.001, seeing if the centroids have changed by doing a comparison will fail due to rounding errors. This is applied to the first 100 instances of each digit in the MNINST training set, we set $k = 2$ to find two clusters as specified. Now a plot of E_{in} as a a funtion of the number of iterations of the k-means algorithm, this plot can be seen in figure 1.

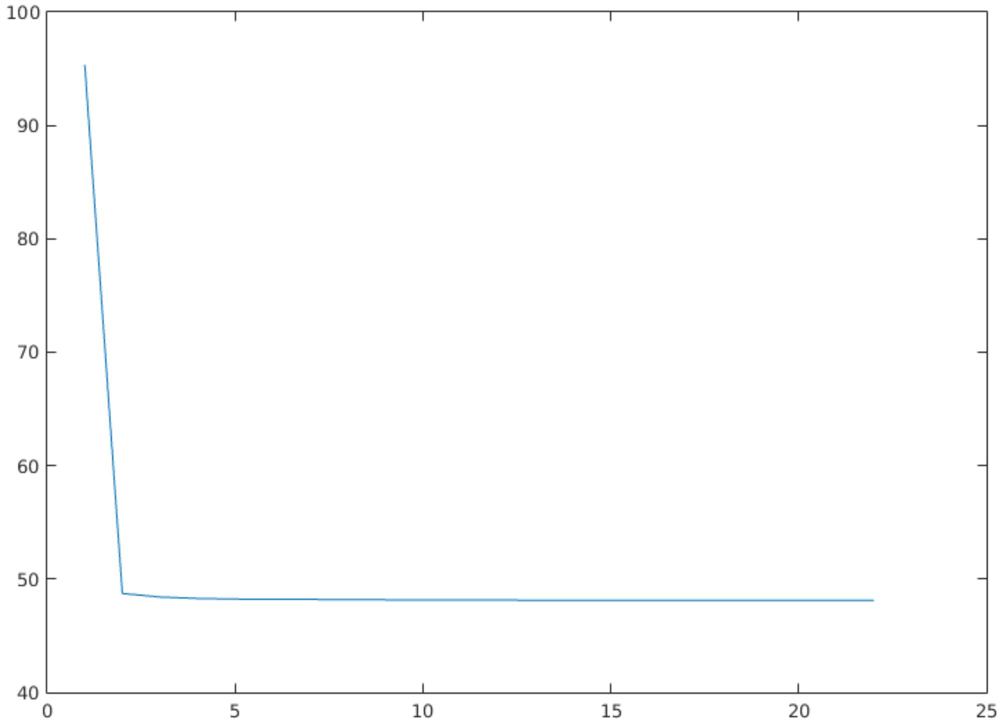

Figure 1:

It is verified that E_{in} monotonically decreases, thus the algorithm is not obviously faulty implemented.

(2)

The same experiment is done with $k = 4$, the plot for this can be seen in figure 2.

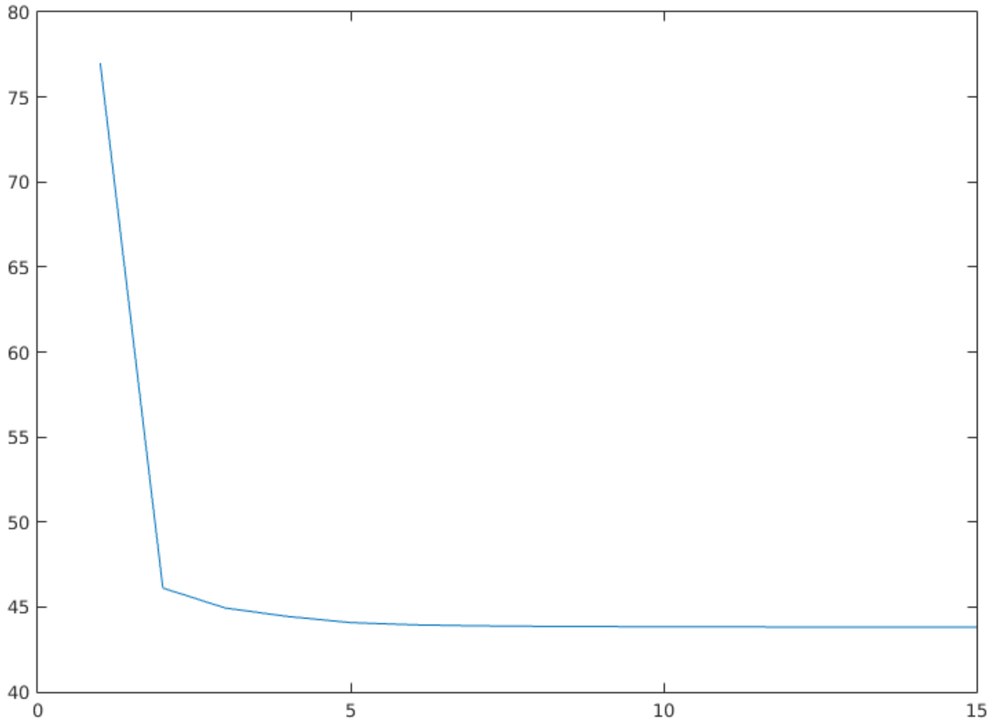

Figure 2:

(3-4)

k-means is now applied to cluster the dataset into $k = [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]$ clusters, for each k we apply 5 random initializations of the algorithm and pick the best outcome.

(5)

For the data calculated in (3-4) we plot E_{in} as a function of k. This plot can be seen below in figure 3.

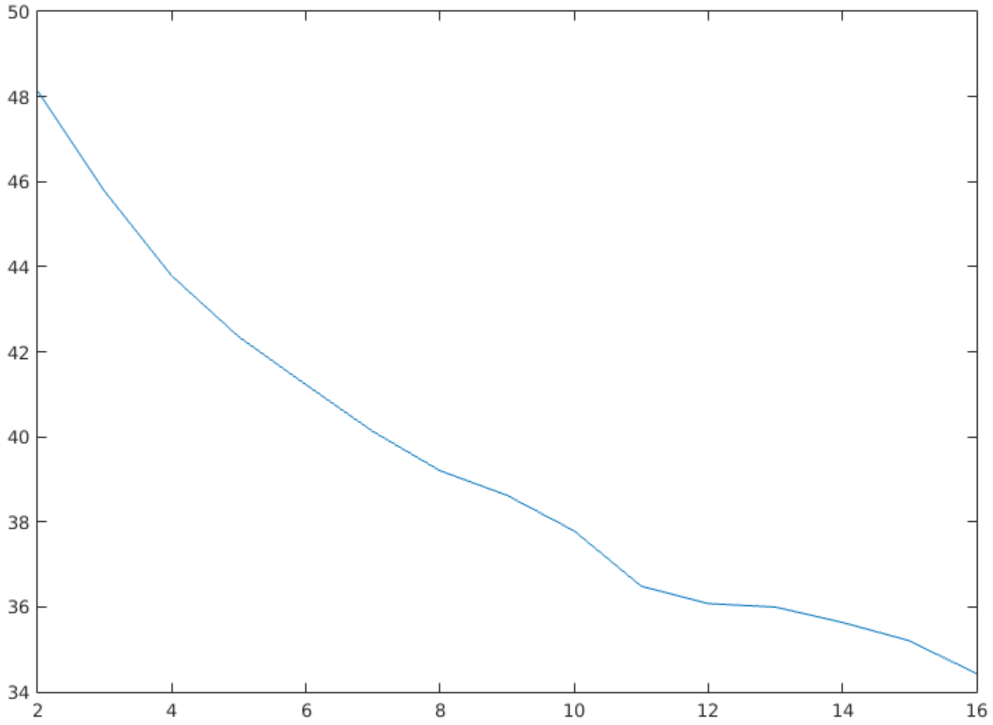

Figure 3:

It can easily be seen that E_{in} monotonically decreases, this makes sense as the more clusters we have the smaller distance there will be to their closest centroid.

(6)

For $k = 10$ we want to visualize the centroids of the clusters, these can be seen in figure 4.

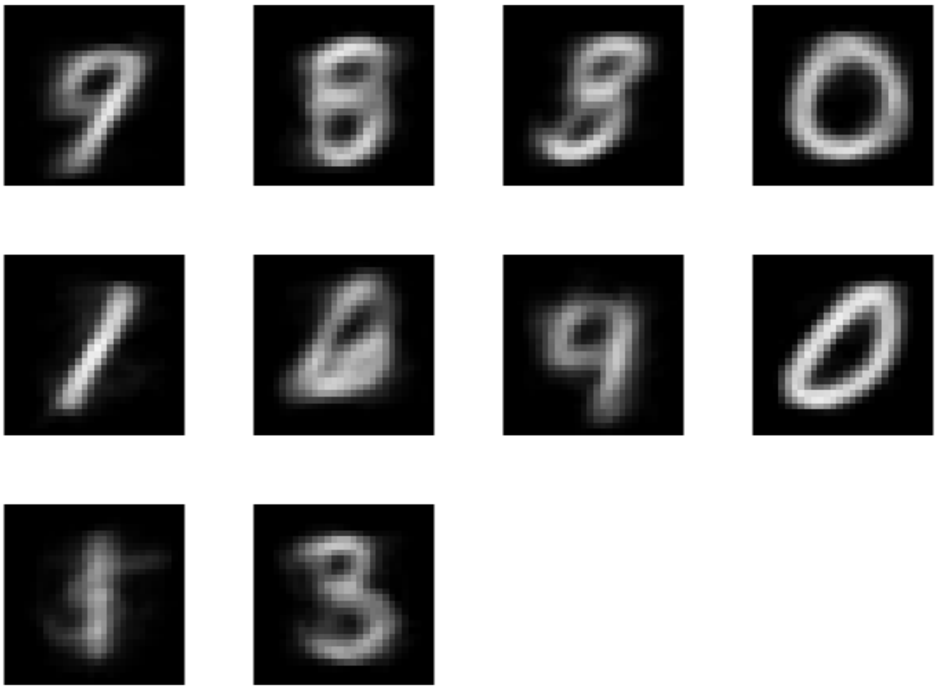

Figure 4:

(7)

We check the coherence of each cluster and find the weighted error, this is found to be:

$$err = 0.5420$$

As it can be seen in the visualizations as well as the weighted error, the clustering is not that successful in identifying the digits, it is mostly taking the wrong guess in our case.

(8)

We want to do a scatter plot of the images with their respective labels shown as shapes and colours, this can be done by applying PCA to the dataset, when this is completed we can plot the first two principle components, this can be seen in figure 5

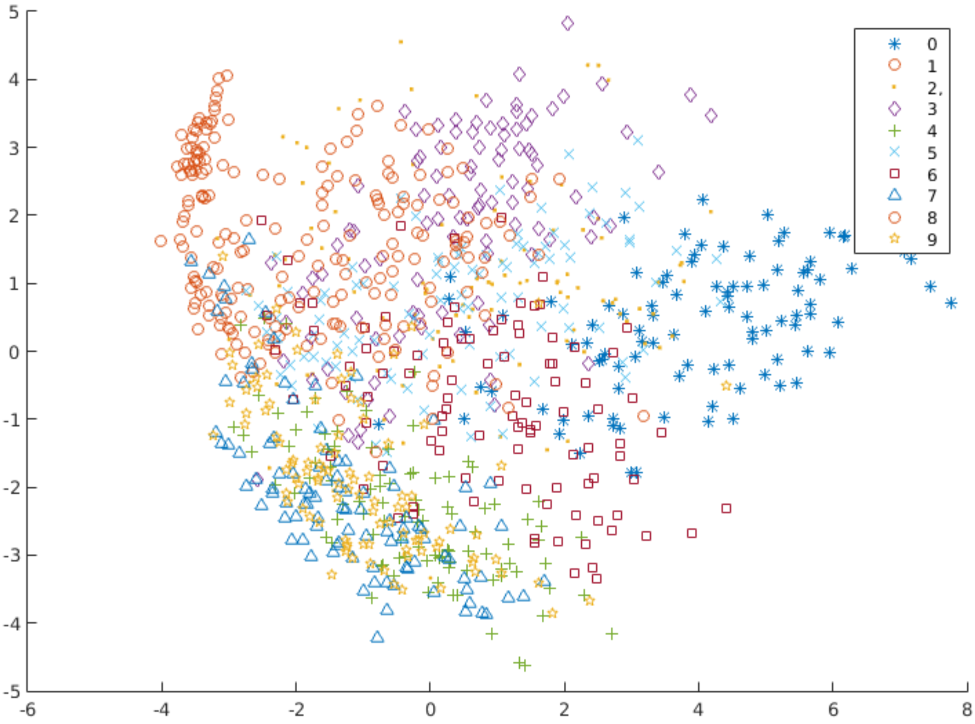

Figure 5:

(9)

We now want to do the same plot as in (7), but this time with the labels according to the centroids for $k = [2, 4, 8, 10, 16]$, these plots can be seen in figure 6.

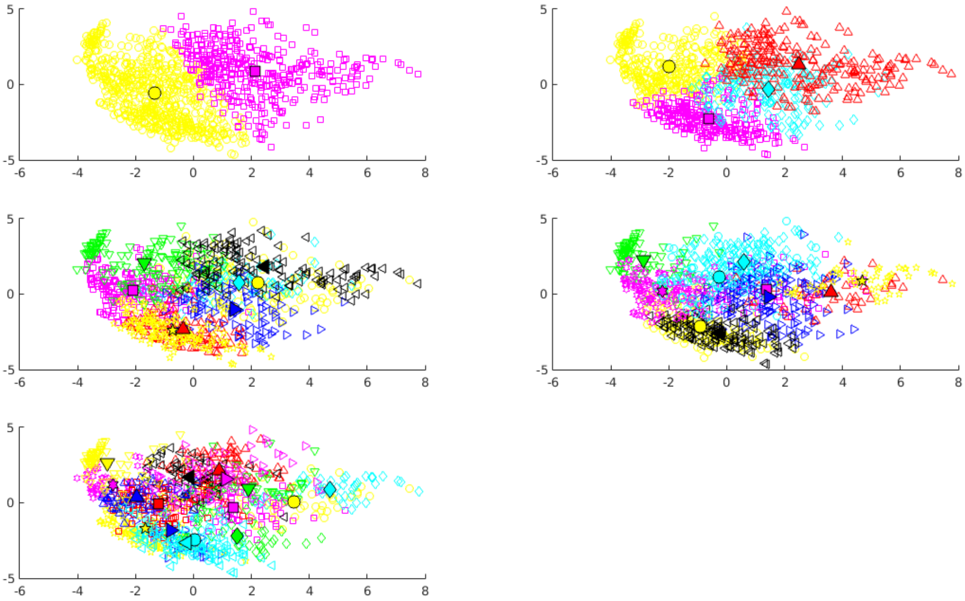

Figure 6: