Homework 3

Exercise 18.1; Exercise 18.2; Exercise 18.5; Exercise 18.8; Exercise 18.11
Exercise 21.6; Exercise 21.10; Exercise 21.11
Exercise 13.2; Exercise 13.9
Exercise 14.2; Exercise 14.6;
Exercise 15.2
Exercise 16.3; Exercise 16.13; Exercise 16.17; Exercise 16.20

Exercise 18.1 (0.5')

What is the rank of the 3 × 3 diagonal matrix below?

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 2 & 1 \end{pmatrix}$$

Solution:
By applying Gauss elimination, we can get:

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 2 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Hence, the rank of this matrix is 2.

Exercise 18.2 (0.5')
Show that $\lambda = 2$ is an eigenvalue of

$$C = \begin{pmatrix} 6 & -2 \\ 4 & 0 \end{pmatrix}$$

Find the corresponding eigenvector.
Solution:
If $\lambda = 2$, then $\det(C - \lambda I) = \lambda^2 - 6\lambda + 8 = 0$. Hence, $\lambda = 2$ is an eigenvalue of C.
Suppose the corresponding eigenvector is $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, so we get

$$\begin{pmatrix} 6 & -2 \\ 4 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 2 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Solving the system of equations, we get $\mathrm{x}_2 = 2x_1$.
Hence, any vector $\begin{pmatrix} k \\ 2k \end{pmatrix}$ $(k \neq 0)$ is the corresponding eigenvector.

Exercise 18.5 (0.5')
Verify that the SVD of the matrix in Equation (18.12) is

$$\mathrm{U} = \begin{pmatrix} -0.816 & 0.000 \\ -0.408 & -0.707 \\ -0.408 & 0.707 \end{pmatrix}, \Sigma = \begin{pmatrix} 1.732 & 0.000 \\ 0.000 & 1.000 \end{pmatrix} \ and\ V^T = \begin{pmatrix} -0.707 & -0.707 \\ 0.707 & -0.707 \end{pmatrix},$$

by verifying all of the properties in the statement of Theorem 18.3.

Solution:

$$C^T C = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \times \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

$$\det(C^T C - \lambda I) = \lambda^2 - 4\lambda + 3 = 0$$

$$\lambda_1 = 3, \lambda_2 = 1$$

$$CC^T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \times \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

$$\det(CC^T - \lambda I) = det \begin{pmatrix} 2-\lambda & 1 & 1 \\ 1 & 1-\lambda & 0 \\ 1 & 0 & 1-\lambda \end{pmatrix} = \lambda(\lambda^2 - 4\lambda + 3) = 0$$

$$\lambda_1 = 3, \lambda_2 = 1, \lambda_3 = 0$$

It turns out that the first two largest eigenvalues of $C^T C$ are the same as those of CC^T.

Furthermore, $\Sigma_{11} \approx \sqrt{\lambda_1,}, \Sigma_{22} = \sqrt{\lambda_2,}.$

Exercise 18.8 (0.5')

Compute a rank 1 approximation *C*1 to the matrix *C* in Exercise 18.12, using the SVD as in Equation 18.13. What is the Frobenius norm of the error of this approximation?

Solution:

$$\Sigma = \begin{pmatrix} 1.732 & 0.000 \\ 0.000 & 1.000 \end{pmatrix}, \ \Sigma_1 = \begin{pmatrix} 1.732 & 0.000 \\ 0.000 & 0.000 \end{pmatrix}$$

$$\mathrm{C_1} = \mathrm{U\Sigma_1 V} = \begin{pmatrix} -0.816 & 0.000 \\ -0.408 & -0.707 \\ -0.408 & 0.707 \end{pmatrix} \begin{pmatrix} 1.732 & 0.000 \\ 0.000 & 0.000 \end{pmatrix} \begin{pmatrix} -0.707 & -0.707 \\ 0.707 & -0.707 \end{pmatrix}$$

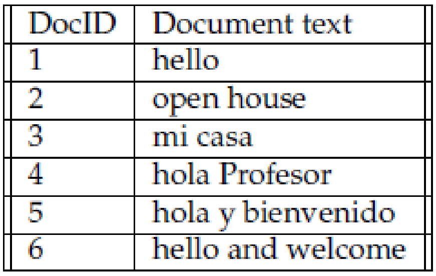
$$= \begin{pmatrix} 0.9992 & 0.9992 \\ 0.4996 & 0.4996 \\ 0.4996 & 0.4996 \end{pmatrix} \approx \begin{pmatrix} 1 & 1 \\ 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$$

$$\mathrm{X} = \mathrm{C} - \mathrm{C_1} = \begin{pmatrix} 0 & 0 \\ -0.5 & 0.5 \\ 0.5 & -0.5 \end{pmatrix}$$

$$\text{Frobenius norm } = (-0.5)^2 + (0.5)^2 + (0.5)^2 + (-0.5)^2 = 1$$

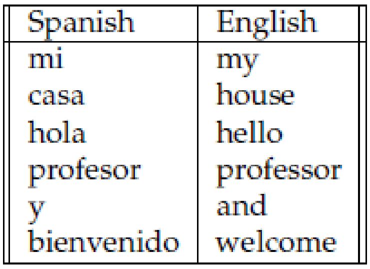
Exercise 18.11 (1')

Assume you have a set of documents each of which is in either English or in Spanish. The collection is given in Figure 18.4.



DocID	Document text
1	hello
2	open house
3	mi casa
4	hola Profesor
5	hola y bienvenido
6	hello and welcome

► **Figure 18.4** Documents for Exercise 18.11.



Spanish	English
mi	my
casa	house
hola	hello
profesor	professor
y	and
bienvenido	welcome

► **Figure 18.5** Glossary for Exercise 18.11.

Figure 18.5 gives a glossary relating the Spanish and English words above for your own information. This glossary is NOT available to the retrieval system:
1. Construct the appropriate term-document matrix C to use for a collection consisting of these documents. For simplicity, use raw term frequencies rather than normalized tf-idf weights. Make sure to clearly label the dimensions of your matrix.
2. Write down the matrices U2, Σ_2 and V2 and from these derive the rank 2 approximation
C2.
3. State succinctly what the (i, j) entry in the matrix $C^T C$ represents.
4. State succinctly what the (i, j) entry in the matrix C^T
2 C2 represents, and why it differs from that in $C^T C$.

Solution:

1

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
hello	1	0	0	0	0	1
open	0	1	0	0	0	0
house	0	1	0	0	0	0
mi	0	0	1	0	0	0
casa	0	0	1	0	0	0
hola	0	0	0	1	1	0
Profesor	0	0	0	1	0	0
y	0	0	0	0	1	0
bienvenido	0	0	0	0	1	0
and	0	0	0	0	0	1
welcome	0	0	0	0	0	1

$$C=\begin{pmatrix}1&0&0&0&0&1\\0&1&0&0&0&0\\0&1&0&0&0&0\\0&0&1&0&0&0\\0&0&1&0&0&0\\0&0&0&1&1&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&0&1&0\\0&0&0&0&0&1\\0&0&0&0&0&1\end{pmatrix}$$

C is an 11x6 matrix.

2

Σ_2 =

```
    1.9021         0
    0           1.8478
```

U_2 =

```
             0        0.7071
    0.0000             0
   -0.0000             0
    0.0000             0
   -0.0000             0
   -0.7236             0
   -0.2764             0
   -0.4472             0
   -0.4472             0
         0        0.5000
         0        0.5000
```

V_2 =

```
         0        0.3827
         0             0
         0             0
   -0.5257             0
   -0.8507             0
         0        0.9239
```

C2 =

```
    0.5000         0         0         0         0    1.2071
         0         0         0   -0.0000   -0.0000         0
         0         0         0    0.0000    0.0000         0
```

```
     0          0          0    -0.0000    -0.0000          0
     0          0          0     0.0000     0.0000          0
     0          0          0     0.7236     1.1708          0
     0          0          0     0.2764     0.4472          0
     0          0          0     0.4472     0.7236          0
     0          0          0     0.4472     0.7236          0
0.3536          0          0          0          0     0.8536
0.3536          0          0          0          0     0.8536
```

3. The (i, j) entry in the matrix C^TC represents the number of terms occurring in both document i and document j.

4. The (i, j) entry in the matrix $C_2^TC_2$ represents the similarity between document i and document j in the low dimensional space.

Exercise 21.6 (0.5’)

Consider a web graph with three nodes 1, 2 and 3. The links are as follows: 1 → 2, 3 → 2, 2 → 1, 2 → 3. Write down the transition probability matrices for the surfer’s walk with teleporting, for the following three values of the teleport probability: (a) *a* = 0; (b) *a* = 0.5 and (c) *a* = 1.

<u>Solution:</u>

(i)

$$\begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{pmatrix}$$

(ii)

$$\begin{pmatrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{pmatrix}$$

(iii)

$$\begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$$

Exercise 21.10 (0.5’)

Show that the PageRank of every page is at least *a*/*N*. What does this imply about the difference in PageRank values (over the various pages) as *a* becomes close to 1?

<u>Solution:</u>

According to the definition of P_{ji}, we can find $P_{ji} \geq \alpha/\mathrm{N}$. Hence,

$$\vec{x}_i = \sum_{j=1}^{N} (\vec{x}_j P_{ji}) \geq \sum_{j=1}^{N} (\vec{x}_j \alpha/\mathrm{N}) = \left(\frac{\alpha}{\mathrm{N}}\right) \sum_{j=1}^{N} \vec{x}_j = \alpha/\mathrm{N}$$

So the PageRank of every page is at least α/N.
As α becomes closer to 1, the impact of the link structure of the web graph gets smaller. Hence, the difference in PageRank values over various pages will get smaller.

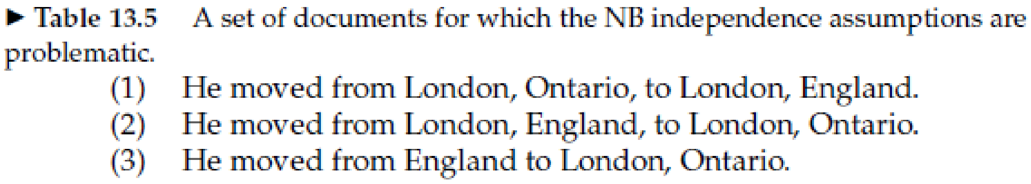
Exercise 21.11 (0.5')
For the data in Example 21.1, write a small routine or use a scientific calculator to compute the PageRank values stated in Equation (21.6).
<u>Solution 1:</u>
x*P = x, x = [0.05 0.04 0.11 0.25 0.21 0.035 0.31].

<u>Solution 2:</u>
The matlab code is as follows:

```
P = [0.02 0.02 0.88 0.02 0.02 0.02 0.02;

     0.02 0.45 0.45 0.02 0.02 0.02 0.02;

     0.31 0.02 0.31 0.31 0.02 0.02 0.02;

     0.02 0.02 0.02 0.45 0.45 0.02 0.02;

     0.02 0.02 0.02 0.02 0.02 0.02 0.88;

     0.02 0.02 0.02 0.02 0.02 0.45 0.45;

     0.02 0.02 0.02 0.31 0.31 0.02 0.31;];

[W D] = eig(P');

x = W(:, 1)';

x = x / sum(x);
```

We can get x = [0.05 0.04 0.11 0.25 0.21 0.035 0.31].

Exercise 13.2 (0.5')
Which of the documents in Table 13.5 have identical and different bag of words representations for (i) the Bernoulli model (ii) the multinomial model? If there are differences, describe them.



▶ **Table 13.5** A set of documents for which the NB independence assumptions are problematic.

(1) He moved from London, Ontario, to London, England.
(2) He moved from London, England, to London, Ontario.
(3) He moved from England to London, Ontario.

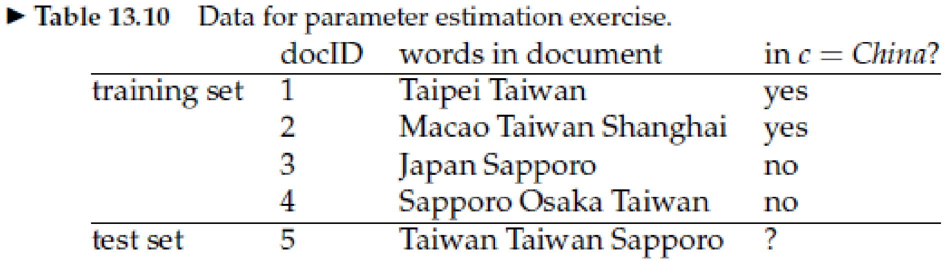
Solution:

(i) For the Bernoulli model, the 3 documents are identical.

(ii) For the multinomial model, documents 1 and 2 are identical and they are different from document 3, because the term London occurs twice in documents 1 and 2, but occurs once in document 3.

Exercise 13.9 (1')

Based on the data in Table 13.10, (i) estimate a multinomial Naive Bayes classifier, (ii) apply the classifier to the test document, (iii) estimate a Bernoulli NB classifier, (iv) apply the classifier to the test document. You need not estimate parameters that you don't need for classifying the test document.



▶ **Table 13.10** Data for parameter estimation exercise.

	docID	words in document	in $c = China$?
training set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
test set	5	Taiwan Taiwan Sapporo	?

Solution:

(Multinomial model)

(i)

$$p(c) = \frac{1}{2}$$

$$P(Taiwan|c) = \frac{2+1}{5+7} = \frac{1}{4}$$

$$P(Sapporo|c) = \frac{0+1}{5+7} = \frac{1}{12}$$

$$P(\bar{c}) = \frac{1}{2}$$

$$P(Taiwan|\bar{c}) = \frac{1+1}{5+7} = \frac{1}{6}$$

$$P(Sapporo|\bar{c}) = \frac{2+1}{5+7} = \frac{1}{4}$$

(ii)

$$P(c|d5) \propto p(c)p(Taiwan|c)^2 p(Sapporo|c) = \frac{1}{2}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{12} \approx 0.0026$$

$$P(\bar{c}|d5) \propto p(\bar{c})p(Taiwan|\bar{c})^2 p(Sapporo|\bar{c}) = \frac{1}{2}\cdot\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{4} \approx 0.0035$$

So document 5 is not in Class China.

(Bernoulli model)

(iii)

$$P(c) = 1/2$$

$$P(Taiwan|c) = \frac{2+1}{2+2} = \frac{3}{4}$$

$$P(Taipei|c) = P(Macro|c) = P(Shanghai|c) = \frac{1+1}{2+2} = \frac{1}{2}$$

$$P(Japan|c) = (Sapporo|c) = P(Osaka|c) = \frac{0+1}{2+2} = \frac{1}{4}$$

$$P(\bar{c}) = \frac{1}{2}$$

$$P(Taiwan|\bar{c}) = P(Japan|\bar{c}) = P(Osaka|\bar{c}) = \frac{1+1}{2+2} = \frac{1}{2}$$

$$P(Sapporo|\bar{c}) = \frac{2+1}{2+2} = \frac{3}{4}$$

$$P(Taipei|\bar{c}) = P(Macro|\bar{c}) = P(Shanghai|\bar{c}) = \frac{0+1}{2+2} = \frac{1}{4}$$

(iv)

$$P(c|d5)$$

$$\propto p(c)p(Taiwan|c)p(Sapporo|c)\big(1-p(Macro|c)\big)\big(1-p(Shanghai|c)\big)$$

$$\big(1-p(Taipei|c)\big)\big(1-p(Japan|c)\big)\big(1-p(Oskta|c)\big)$$

$$= \frac{1}{2}\cdot\frac{3}{4}\cdot\frac{1}{4}\cdot(1-\frac{1}{2})\cdot\left(1-\frac{1}{2}\right)\cdot\left(1-\frac{1}{2}\right)\cdot\left(1-\frac{1}{4}\right)\cdot\left(1-\frac{1}{4}\right) \approx 0.0066$$

$$P(\bar{c}|d5)$$

$$\propto p(c)p(Taiwan|\bar{c})p(Sapporo|\bar{c})\big(1-p(Macro|\bar{c})\big)\big(1-p(Shanghai|\bar{c})\big)$$

$$\big(1-p(Taipei|\bar{c})\big)\big(1-p(Japan|\bar{c})\big)\big(1-p(Osaka|\bar{c})\big)$$

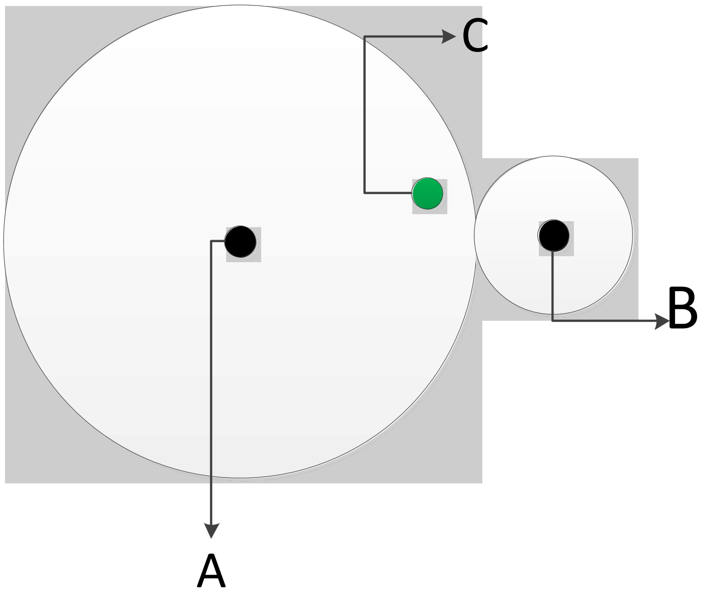
$$= \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{3}{4}\cdot(1-\frac{1}{4})\cdot(1-\frac{1}{4})\cdot(1-\frac{1}{4})\cdot(1-\frac{1}{2})\cdot(1-\frac{1}{2}) \approx 0.0198$$

So document 5 is not in Class China.

Exercise 14.2 (0.5')

Show that Rocchio classification can assign a label to a document that is different from its training set label.

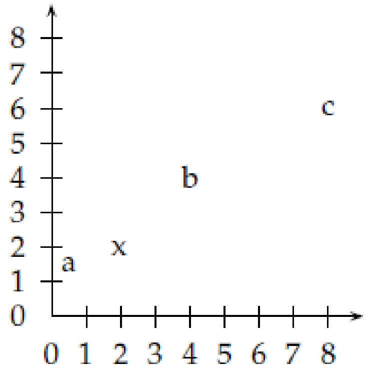
<u>**Solution:**</u>



Take the above picture as an example. There are 2 classes in the plane, with the left one being much bigger than the right one. Then a large part of the left circle will be misclassified like the point C. C is a document belonging to the A class in the training set, but it is closer to B than A. So it will be labeled as the B class.

Exercise 14.6 (0.5')

In Figure 14.14, which of the three vectors a, b, and c is (i) most similar to x according to dot product similarity, (ii) most similar to x according to cosine similarity, (iii) closest to x according to Euclidean distance?



► **Figure| 14.14** Example for differences between Euclidean distance, dot product similarity and cosine similarity. The vectors are $\vec{a} = (0.5 \ \ 1.5)^T$, $\vec{x} = (2 \ \ 2)^T$, $\vec{b} = (4 \ \ 4)^T$, and $\vec{c} = (8 \ \ 6)^T$.

Solution:

(i) <a, x> = 4, <b, x> = 16, <c, x> = 28.

So a is most similar to x.

(ii) $\frac{<a,x>}{|a||x|} = 0.8944$, $\frac{<b,x>}{|b||x|} = 1$, $\frac{<c,x>}{|c||x|} = 0.9899$

So b is most similar to x.

(iii) d(x, a) = 1.5811, d(x, b) = 2.8284, d(x, c) = 7.2111

So a is most similar to x.

Exercise 15.2 (0.5’)

The basis of being able to use kernels in SVMs (see Section 15.2.3) is that the classification function can be written in the form of Equation (15.9) (where, for large problems, most α_i are 0). Show explicitly how the classification function could be written in this form for the data set from Example 15.1. That is, write f as a function where the data points appear and the only variable is x.

Solution:

Assume the three points: $\vec{x}_1 = \begin{pmatrix}1\\1\end{pmatrix}, y_1 = -1,\ \vec{x}_2 = \begin{pmatrix}2\\0\end{pmatrix}, y_2 = -1,\ \vec{x}_3 = \begin{pmatrix}2\\3\end{pmatrix}, y_3 = 1$

$\alpha_2 = 0$

$-\alpha_1 < \begin{pmatrix}1\\1\end{pmatrix}, \vec{x} > +\alpha_2 < \begin{pmatrix}2\\3\end{pmatrix}, \vec{x} > = < \begin{pmatrix}2/5\\4/5\end{pmatrix}, \vec{x} >$

$\alpha_1 = \alpha_2 = 2/5$

$\mathrm{f(x)} = sgn(-\frac{2}{5} < \begin{pmatrix}1\\1\end{pmatrix}, \vec{x} > + \frac{2}{5} < \begin{pmatrix}2\\3\end{pmatrix}, \vec{x} > - \frac{11}{5})$

Exercise 16.3 (1’)

Replace every point *d* in Figure 16.4 with two identical copies of *d* in the same class. (i) Is it less difficult, equally difficult or more difficult to cluster this set of 34 points as opposed to the 17 points in Figure 16.4? (ii) Compute purity, NMI, RI, and *F*5 for the clustering with 34 points. Which measures increase and which stay the same after doubling the number of points? (iii) Given your assessment in (i) and the results in (ii), which measures are best suited to compare the quality of the two clusterings?

Solution:

(i) It is equally difficult.

(ii) $\text{purity} = (1/34) \times (10 + 8 + 6) \approx 0.71$

NMI = 0.36

TP = 97, FP = 80, FN = 96, TN = 288

P = 0.55, R = 0.50

RI = 0.686

F5 = 0.5

Purity and NMI stay the same, while RI and F5 increase.

(iii) Purity and NMI

Exercise 16.13 (0.5’)

Prove that RSSmin(*K*) is monotonically decreasing in *K*.

Solution:

In a clustering with i clusters, take a cluster with non-identical vectors. Splitting this cluster in two will lower RSS.

Exercise 16.17 (0.5’)

Perform a *K*-means clustering for the documents in Table 16.3. After how many iterations does *K*-means converge? Compare the result with the EM clustering in Table 16.3 and discuss the differences.

Solution:

After 2 iterations K-means converges.

The K-means clustering converges faster than EM. One possible reason is that the data might satisfy the hard clustering condition rather than the soft clustering condition, which can be seen from the final results of EM.

Exercise 16.20 (0.5')

The *within-point scatter* of a clustering is defined as

$\sum_k \frac{1}{2} \sum_{\overrightarrow{x_i} \in w_k} \sum_{\overrightarrow{x_j} \in w_k} |\overrightarrow{x_i} - \overrightarrow{x_j}|^2$.

Show that minimizing RSS and minimizing within-point scatter are equivalent.

Solution:

$$RSS_k = \sum_{\vec{x} \in w_k} |\vec{x} - \vec{u}(w_k)|^2$$

$$= \sum_{\vec{x} \in w_k} |\vec{x}|^2 - \sum_{\vec{x} \in w_k} 2 < \vec{x}, \vec{u}(w_k) > + \sum_{\vec{x} \in w_k} |\vec{u}(w_k)|^2$$

$$= \sum_{\vec{x} \in w_k} |\vec{x}|^2 - 2 \sum_{\vec{x} \in w_k} < \vec{x}, \frac{1}{|w_k|} \sum_{\vec{x} \in w_k} \vec{x} > + |W_k| \frac{1}{|w_k|^2} \left| \sum_{\vec{x} \in w_k} \vec{x} \right|^2$$

$$= \sum_{\vec{x} \in w_k} |\vec{x}|^2 - \frac{1}{|w_k|} \sum_{\vec{x}_i \in w_k} \sum_{\vec{x}_j \in w_k} < \overrightarrow{x_i}, \overrightarrow{x_j} >$$

$$\frac{1}{2} \sum_{\overrightarrow{x_i} \in w_k} \sum_{\overrightarrow{x_j} \in w_k} \left| \overrightarrow{x_i} - \overrightarrow{x_j} \right|^2$$

$$= \frac{1}{2} \left(\sum_{\overrightarrow{x_i} \in w_k} |\overrightarrow{x_i}|^2 + \sum_{\overrightarrow{x_j} \in w_k} |\overrightarrow{x_j}|^2 - \sum_{\overrightarrow{x_i} \in w_k} \sum_{\overrightarrow{x_j} \in w_k} < \overrightarrow{x_i}, \overrightarrow{x_j} > \right)$$

$$= |\mathrm{w}_k| \sum_{\vec{x} \in w_k} |\vec{x}|^2 - \sum_{\overrightarrow{x_i} \in w_k} \sum_{\overrightarrow{x_j} \in w_k} < \overrightarrow{x_i}, \overrightarrow{x_j} >$$

$$RSS_K = \frac{1}{|\mathrm{w}_k|} \frac{1}{2} \sum_{\overrightarrow{x_i} \in w_k} \sum_{\overrightarrow{x_j} \in w_k} |\overrightarrow{x_i} - \overrightarrow{x_j}|^2$$

So minimizing RSS and minimizing within-point scatter are equivalent.