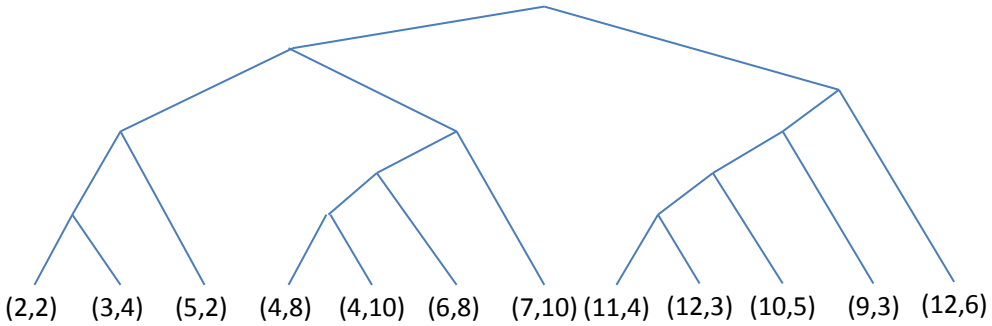
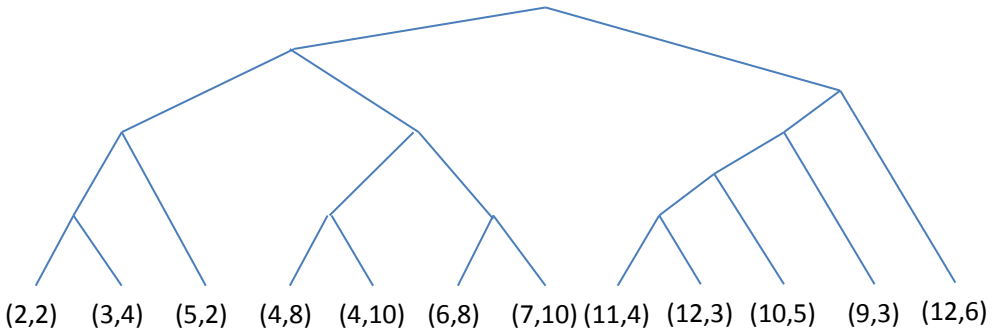
Solutions for Homework 3

Chapter 7 of MMDS Textbook:
Page 233 --- Exercise 7.2.2
Page 242 --- Exercise 7.3.4
Page 242 --- Exercise 7.3.5
Chapter 12 of MMDS Textbook:
Page 413 --- Exercise 12.2.1: (c), (d)
Page 425 --- Exercise 12.3.2: (a)
Page 433 --- Exercise 12.4.3: (a), (b)
Chapter 3 of MMDS Textbook:
Page 59 --- Exercise 3.1.1
Page 62 --- Exercise 3.2.1
Page 68 --- Exercise 3.3.3
Page 73 --- Exercise 3.4.1

Page 233 --- Exercise 7.2.2
(a)



(b)



Page 242 --- Exercise 7.3.4
(a)

cluster	points	N	SUM	SUMSQ
1	(4, 8) (4, 10) (6, 8) (7, 10)	4	(21, 36)	(117, 328)
2	(9, 3) (10, 5) (11, 4) (12, 3) (12, 6)	5	(54, 21)	(590, 95)
3	(2, 2) (3, 4) (5,2)	3	(10, 8)	(38, 24)

(b)

The variance in the i-th dimension is $SUMSQ_i/N - (SUM_i/N)^2$. The standard deviation in each dimension is the square root of the variance.

	cluster 1		cluster 2		cluster 3	
	x	y	x	y	x	y
variance	$\frac{27}{16}$	1	$\frac{34}{25}$	$\frac{34}{25}$	$\frac{14}{9}$	$\frac{8}{9}$
standard deviation	$\frac{3}{4}\sqrt{3}$	1	$\frac{\sqrt{34}}{5}$	$\frac{\sqrt{34}}{5}$	$\frac{\sqrt{14}}{3}$	$\frac{2}{3}\sqrt{2}$

Page 242 --- Exercise 7.3.5

Then the Mahalanobis distance between p(1,-3,4) and c(0,0,0) is

$$\sqrt{\sum_{i=1}^{3}(\frac{p_i - c_i}{\sigma_i})^2} = \sqrt{(\frac{1}{2})^2 + (-\frac{3}{3})^2 + (\frac{4}{5})^2} = \frac{3}{10}\sqrt{21} \approx 1.37$$

Page 413 --- Exercise 12.2.1

The training data is modified to be:

	and	viagra	the	of	nigeria	θ	y
a	1	1	0	1	1	-1	+1
b	0	0	1	1	1	-1	-1
c	0	1	1	0	0	-1	+1
d	1	0	0	1	0	-1	-1
e	1	0	1	0	1	-1	+1
f	1	0	1	1	0	-1	-1

(c)

Use learning rate c=1/2. Begin with w=[0,0,0,0,0,0] and compute w.a=0.
Since 0 is not positive, w:=w+(1/2)(+1)a=[1/2, 1/2, 0, 1/2, 1/2, -1/2].

Consider w.b=3/2 which is not negative. w:=w+(1/2)(-1)b=[1/2, 1/2, 0, 1/2, 1/2, -1/2] – [0, 0, 1/2, 1/2, 1/2, -1/2]=[1/2, 1/2, -1/2, 0, 0, 0].

Compute w.c=0, which is not positive. w:=w+(1/2)(+1)c=[1/2, 1/2, -1/2, 0, 0, 0] + [0, 1/2, 1/2, 0, 0, -1/2]=[1/2, 1, 0, 0, 0, -1/2].

w.d=1, not negative. w:=w+(1/2)(-1)d=[1/2, 1, 0, 0, 0, -1/2] - [1/2, 0, 0, 1/2, 0, -1/2]=[0, 1, 0, -1/2, 0, 0].

w.e=0, not positive. w:=w+(1/2)(+1)e=[0, 1, 0, -1/2, 0, 0]+[1/2, 0, 1/2, 0, 1/2, -1/2]=[1/2, 1, 1/2, -1/2, 1/2, -1/2].

w.f=1, not negative. w:=w+(1/2)(-1)= [1/2, 1, 1/2, -1/2, 1/2, -1/2] - [1/2, 0, 1/2, 1/2, 0, -1/2]=[0, 1, 0, -1, 1/2, 0].

Check a through f, and this w correctly classifies all of them. Thus, we have converged to a perceptron w=[0, 1, 0, -1, 1/2, 0]

(d)

x	y	w.x	OK?	and	viagra	the	of	nigeria	θ
				1	1	1	1	1	1
a	+1	3	yes						
b	-1	2	no	1	1	1/2	1/2	1/2	2
c	+1	-1/2	no	1	2	1	1/2	1/2	1
d	-1	1/2	no	1/2	2	1	1/4	1/2	2
e	+1	0	no	1	2	2	1/4	1	1
f	-1	9/4	no	1/2	2	1	1/8	1	2
a	+1	13/8	yes						
b	-1	1/8	no	1/2	2	1/2	1/16	1/2	4
c	+1	-3/2	no	1/2	4	1	1/16	1/2	2
d	-1	-23/16	yes						
e	+1	0	no	1	4	2	1/16	2	1
f	-1	33/16	no	1/2	4	1	1/32	2	2
a	+1	4+17/32	yes						
b	-1	33/32	no	1/2	4	1/2	1/64	1	4
c	+1	1/2	yes						
d	-1	33/64-4	yes						
e	+1	-2	no	1	4	1	1/64	2	2
f	-1	1/64	no	1/2	4	1/2	1/128	2	4
a	+1	2+65/128	yes						
b	-1	65/128-2	yes						
c	+1	1/2	yes						
d	-1	65/128-4	yes						
e	+1	-1	no	1	4	1	1/128	4	2
f	-1	1/128	no	1/2	4	1/2	1/256	4	4
a	+1		yes						
b	-1	129/256	no	1/2	4	1/4	1/512	2	8
c	+1	1/4 - 4	no	1/2	8	1/2	1/512	2	4

d	-1		yes						
e	+1	-1	no	1	8	1	1/512	4	2
f	-1	1/512	no						

Finally, w does not converge.

Page 425 --- Exercise 12.3.2

(a)

We optimize the following question

$$\min\left|\left|w\right|\right|^2$$

$$\text{s.t. } \mathrm{y}_i(\mathrm{w}\cdot\mathrm{x}_i+\mathrm{b})\geq 1$$

and we get $\mathrm{w}=[0,0.5,0.5],\mathrm{b}=-3.5$

We calculate the distances between the six points and the classification plane.

The support vectors are [3,4,5], [2,7,2], [1,2,3], [3,3,2], [2,4,1].

Page 433 --- Exercise 12.4.3

(a)

$$\mathrm{f(q)}=\begin{cases}1, x\leq 1.5\\ 2, 1.5<x\leq 3\\ 3, 3<x\leq 6\\ 4, 6<x\leq 12\\ 5, 12<x\leq 24\\ 6, x>24\end{cases}$$

(b)

$$\mathrm{f(q)}=\begin{cases}1.5, x\leq 2.5\\ 2.5, 2.5<x\leq 5\\ 3.5, 5<x\leq 10\\ 4.5, 10<x\leq 20\\ 5.5, x>20\end{cases}$$

Page 59 --- Exercise 3.1.1

A={1, 2, 3, 4}, B={2, 3, 5, 7}, and C={2, 4, 6}

SIM (A, B) = 2/6 = 1/3.

SIM (A, C) = 2/5.

SIM (B, C) = 1/6.

Page 62 --- Exercise 3.2.1

The set of the first 10 3-shingles is {“The”, “he “, “e m”, “ mo”, “mos”, “ost”, “st “, “t e”, “ ef”, “eff” }.

Or {“The most effective”, “most effective way”, “effective way to”, “way to

represent", "to represent documents", "represent documents as", "documents as sets", "as sets for", "sets for purpose", "for purpose of"}

Page 68 --- Exercise 3.3.3

(a)

Element	S1	S2	S3	S4	2x+1 mod 6	3x+2 mod 6	5x+2 mod 6
0	0	1	0	1	1	2	2
1	0	1	0	0	3	5	1
2	1	0	0	1	5	2	0
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	1	0	0	0	5	5	3

Minhash Signature Matrix

	S1	S2	S3	S4
h1(0)	\	1	\	1
h2(0)	\	2	\	2
h3(0)	\	2	\	2
h1(1)	\	1	\	1
h2(1)	\	2	\	2
h3(1)	\	1	\	2
h1(2)	5	1	\	1
h2(2)	2	2	\	2
h3(2)	0	1	\	0
h1(3)	5	1	1	1
h2(3)	2	2	5	2
h3(3)	0	1	5	0
h1(4)	5	1	1	1
h2(4)	2	2	2	2
h3(4)	0	1	4	0
h1(5)	5	1	1	1
h2(5)	2	2	2	2
h3(5)	0	1	4	0

The final minhash signature matrix is:

S1	S2	S3	S4
5	1	1	1
2	2	2	2
0	1	4	0

(b)

Only function h3 is a true permutation.

(c)

similarities	1-2	1-3	1-4	2-3	2-4	3-4
col/col	0	0	0.25	0	0.25	0.25
sig/sig	0.33	0.33	0.67	0.67	0.67	0.67

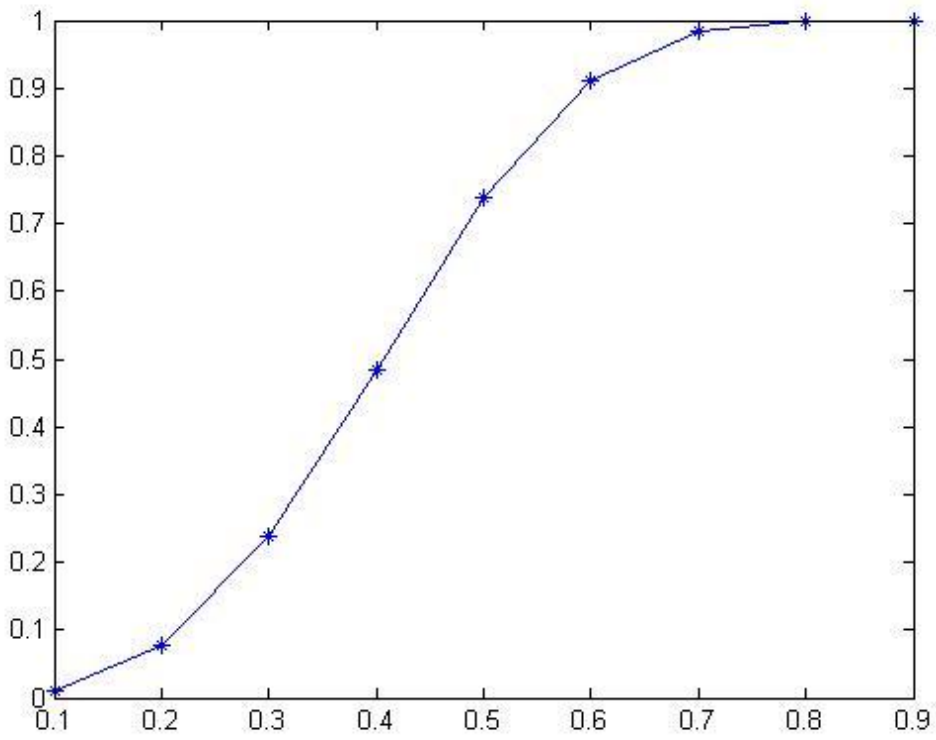
The estimated Jaccard similarities are not close to the true ones at all.

Page 73 --- Exercise 3.4.1

Values of the S-curve for b=10 and r=3

s	$1-(1-s^r)^b$
0.1	0.0100
0.2	0.0772
0.3	0.2394
0.4	0.4839
0.5	0.7369
0.6	0.9123
0.7	0.9850
0.8	0.9992
0.9	1.0000

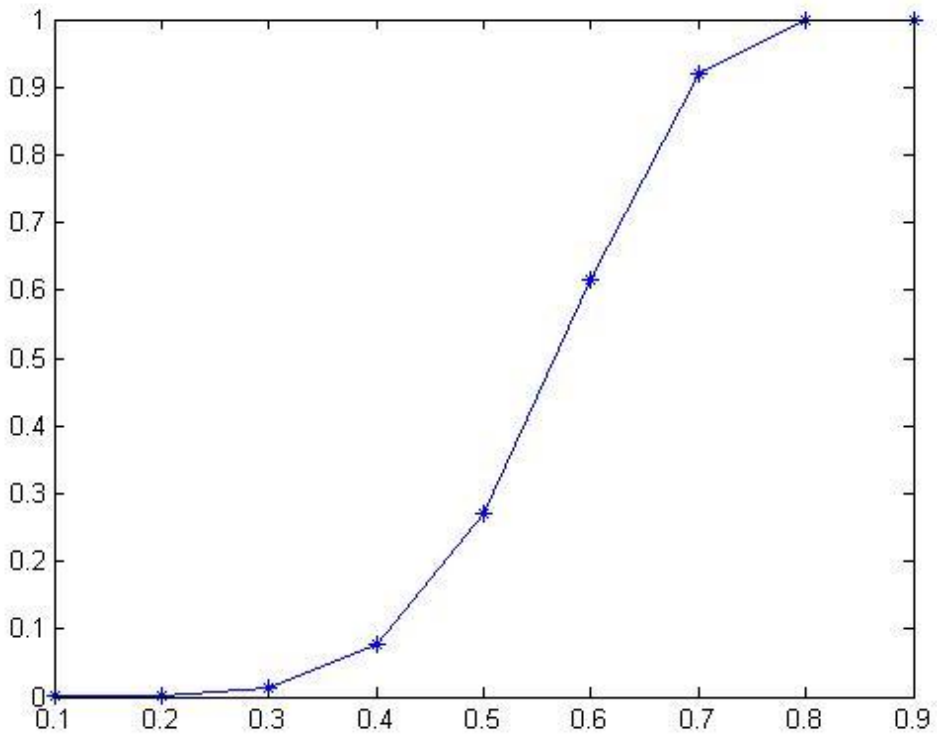
The figure is as follows:



Values of the S-curve for b=20 and r=6

s	$1-\ （1-s^r）^{\ b}$
0.1	0.0000
0.2	0.0013
0.3	0.0145
0.4	0.0788
0.5	0.2702
0.6	0.6154
0.7	0.9182
0.8	0.9977
0.9	1.0000

The figure is as follows:



Values of the S-curve for b=50 and r=5

s	$1-\ （1-s^r）^{\ b}$
0.1	0.0005
0.2	0.0159
0.3	0.1145
0.4	0.4023
0.5	0.7956
0.6	0.9825
0.7	0.9999

0.8	1.0000
0.9	1.0000