一、1. 计算X的均值，Q1，中值。

x	6	-1	3	0	1	4	1	8	3
y	8	-1	5	4	1	6	2	7	2

2. 用最小-最大规范化将y规范到区间[0, 1]

3. 假设上表的每一列代表二维平面上一个点，求Manhattan距离最大的点及其对应的距离，求Euclidean距离最小的点及其对应的距离

(1). 均值: $\bar{x}=(6+(-1)+3+0+1+4+1+8+3)/9=\frac{25}{9}$

对X值进行排序. -1 0 1 1 3 3 4 6 8

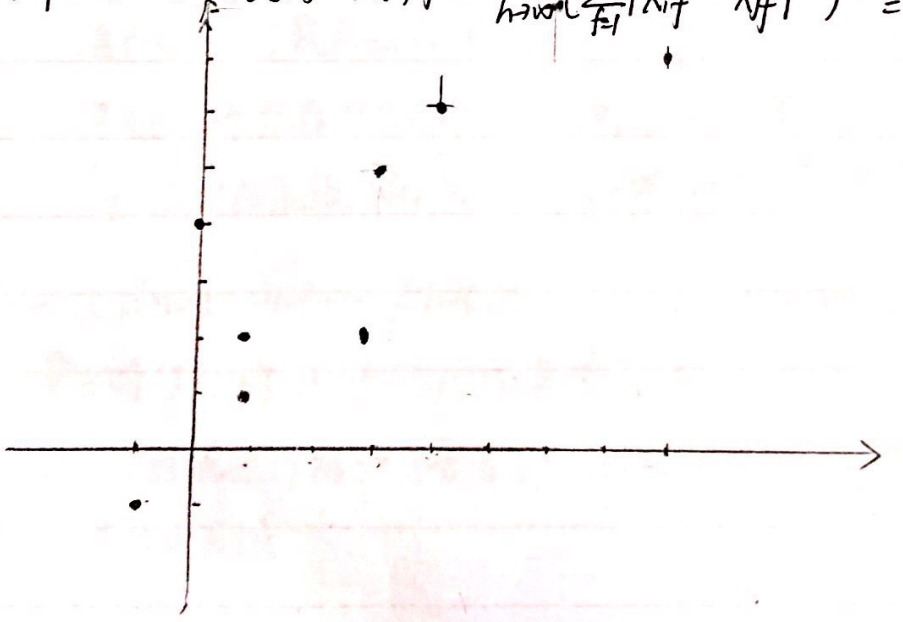
中值为3　　Q1为1.0.　　Q3为4　　(中列数=(最大值+最小值)/2)

(3) ~~闵可夫斯基~~欧氏距离 $d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{ip}-x_{jp})^2}$

曼哈顿距离 $d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\cdots+|x_{ip}-x_{jp}|$.

闵可夫斯基距离. $d(i,j)=\sqrt[h]{|x_{i1}-x_{j1}|^h+|x_{i2}-x_{j2}|^h+\cdots+|x_{ip}-x_{jp}|^h}$ (h范数, p=h)

上确界(切比雪夫)距离 $d(i,j)=\lim_{h\to\infty}\left(\sum_{f=1}^{p}|x_{if}-x_{jf}|^h\right)^{\frac{1}{h}}=\max_{f}^{p}|x_{if}-x_{jf}|$



Euclidean距离最小应该是(1,1)和(1,2). $d=\sqrt{(1-1)^2+(1-2)^2}=1$

Manhattan距离最大应该是 (-1,-1)和(6,8)或(-1,-1)和(8,7).

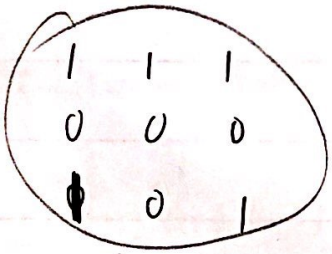
$d_1=7+9=16$　$d_2=9+8=17$

∴ (-1,-1)和(8,7).

1) 解释卷积神经网络中"卷积核"和"池化"的概念

2) 将图1中的卷积运算补充完整.

1x1	1x1	1x1	0	0	4 3 3
0x0	1x0	1x0	1	0	3 4 3
0x1	0x0	1x1	1	1	2 3 4
0	0	1	1	0	
0	1	1	0	0	



— 卷积核

卷积核: ~~假设108个参数需要变成100个参数，这~~

108个参数需要变成了100个参数需要确定，这种技术被称为"权值共享".

其中这100个参数构成的神经元处理方式称为一个滤波器/filter或卷积核

即不同的卷积核对应不同的图像特征(特征提取方法).

每个卷积核作用在输入图像上，就得到一个映射特征(可视为一种图像变换)

池化: (down-pooling /下采样) (对卷积结果进行处理).

聚合特征、降维 达到减少运算量的目的.

对一块数据进行抽样或聚合，例如选择该区域的最大值(或平均值)

取代该区域

(2). 最小-最大规范化：对属性A，从$[V_{min}, V_{max}]$映射到$[V'_{min}, V'_{max}]$

$$V' = \frac{V - V_{min}}{V_{max} - V_{min}}(V'_{max} - V'_{min}) + V'_{min}$$

Z分数规范化（0均值规范化）：设数据集在属性A上的均值为μ_A，方差为σ_A，则对任意值V映射为$V' = \frac{V - \mu_A}{\sigma_A}$.

小数定标规范化：$V' = \frac{V}{10^j}$，j是使$Max(|V'|) < 1$的最小j

对V排序：-1, 1, 2, 2, 4, 5, 6, 7, 8　　此时$V_{max} - V_{min} = 9$.

$$V_1' = \frac{8 - (-1)}{9}(1 - 0) + 0 = 0$$

$$V_2' = \frac{-1 - (-1)}{9}(1 - 0) + 0 = 0$$

$$V_3' = \frac{5 - (-1)}{9}(1 - 0) + 0 = \frac{6}{9}$$

1. 价格. 11 85 10 296 38 102 123 203 65 225

销量. 3 3 4 35 9 5 1 26 8 42

9 164 102 125 73 615 125 242 441 325

5 13 6 11 12 32 12 48 55 78

(1) 对销量排序: 1 3 3 4 (5) 5 6 8 9 (11) 12 12 13 26 (32) 35 42 48 55 78

中位数 11. Q1: 5 Q3: 32.

均值 $\bar{x} = (\Sigma 销量)/20$.

对价格排序: 9 10 11 38 (65) 73 85 102 102 (123) 125 125 164

203 (225) 242 296 325 441 615

中位数 123. Q1: 65 Q3: 225.

(数据的模态: 有几个众数就叫几模, 五数概括: Q1、中值、Q3、最大值、最小值.

(3) 正分数规范化: $v' = \frac{v - \mu_A}{\sigma_A}$.

(4) Pearson 相关系数

2. 设最小支持度为0.4，给定交易数据集，如下表

t1	b d f g l
t2	f g h l m n
t3	b f h k m
t4	a f h j m
t5	d f g j m.

(1) Apriori： 绝对支持度 = 5×0.4=2

C1： 项集 支持度计数. 与最小支持度比较 L1：

项集	支持度计数
a	1
b	2
d	2
f	5
g	3
h	3
j	2
k	1
l	2
m	4
n	1

项集	支持度计数
b	2
d	2
f	5
g	3
h	3
j	2
l	2
m	4

由L1生成C2： 项集 支持度计数.

{b,d}	1	{d,m}	1	{h,j}	1
{b,f}	2	{f,g}	3	{h,l}	1
{b,g}	1	{f,h}	3	{h,m}	3
{b,h}	1	{f,j}	2	{j,m}	2
~~{b,j}~~ {b,l}	1	{f,l}	2	{l,m}	1.
{b,m}	1	{f,m}	4		
{d,f}	2	{g,h}	1		
{d,g}	2	{g,l}	2		
{d,j}	1	{g,m}	1		
{d,l}	1				

由C2生成L2：

{b,f} 2
{d,f} 2
{d,g} 2
{f,g} 3
{f,h} 3
{f,j} 2
{f,l} 2
{f,m} 4
{g,l} 2
{h,m} 3
{j,m} 2.

~~由C3 L3生成C4~~

~~{f,g,l,m}~~

由L2生成C3.

{d,f,g} 2.
{f,g,l} 2
{f,g,m} 2
{f,j,m} 2

由C3生成L3

{d,f,g} 2
{f,g,l} 2
{f,g,m} 2
{f,j,m} 2

∴ 算法终止，找到了所有的频繁项集L1, L2, L3.

决策树

(1). 信息熵

$$H=-\sum_i D_i \log D_i = -\left(\frac{5}{8}\log\frac{5}{8}+\frac{3}{8}\log\frac{3}{8}\right)$$

$$=-\log\frac{1}{2}=\log 2.$$

(2) 通过信息增益来选择分割属性 A1 或 A2.

① 对 A1 的每个类别考察 ~~+和-~~ + 和 - 的分布.

$$Info(A1)=\frac{3}{8}\times\left(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}\right)$$

$$+\frac{5}{8}\times\left(-\frac{1}{5}\log_2\frac{1}{5}-\frac{4}{5}\log_2\frac{4}{5}\right)$$

$$=\left(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}\right)\times 2$$

$$Info(A2)=\frac{2}{8}\left(-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}\right)$$

$$+\frac{6}{8}\left(-\frac{2}{6}\log\frac{2}{6}-\frac{2}{6}\log\frac{2}{6}\right)$$

$$=\frac{1}{3}\log 2+\frac{2}{3}\log 2=\log 2$$

~~Info(A1)~~ $Info(D)-Info(A1)=\sim$

$Info(D)-Info(A2)=0$ (信息增益小)

所以分割 ~~A2~~ A1 (选增益大的).

球	网	打
风力		
湿度		
温度		
天气		
弱	小	N
强	小	N
弱	大	Y
弱	小	Y
强	大	Y
弱	小	N
弱	大	Y
强	小	N
弱	大	Y
弱	大	Y
强	大	Y
强	小	Y
弱	大	Y
强	小	N
高	热	晴
高	热	晴
高	热	阴
高	中	雨
正常	冷	雨
正常	冷	晴
高	中	雨
正常	冷	晴
正常	中	阴
正常	中	晴
高	中	阴
正常	热	阴
高	中	雨

A1	A2	C类别
T	T	+
T	T	+
T	F	-
F	F	+
F	T	-
F	T	-

$-\frac{3}{8}\log\frac{3}{8}+$

下: $\frac{2}{3}$

(1)、用朴素贝叶斯分类器算法求解该问题，给出网络结构图.

(2)、给出朴素贝叶斯分类器(晴、后、高强)的计算过程和结果

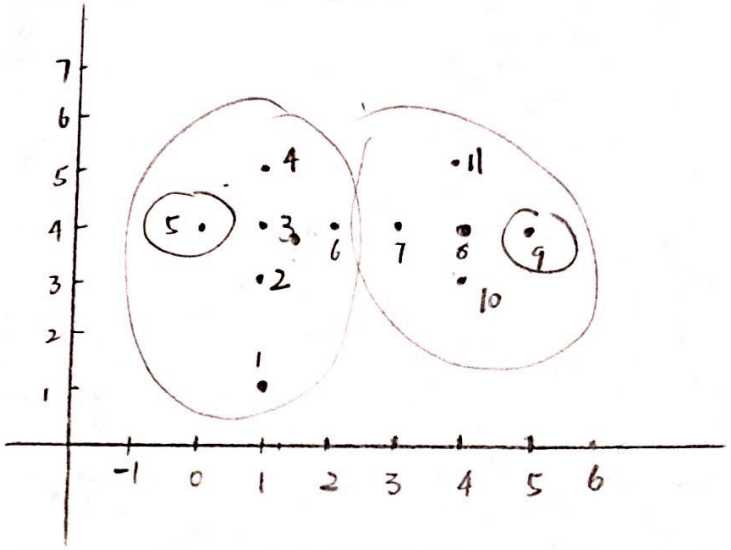
(3)解释最大似然和最大后验概率的联系和区别

共有11个数据点，坐标都是整数，初始选择点"5"和"9"，令cluster数目 k=2，执行k-means算法：

1. 给出第一次和第二次调整后的中心点。

2. 给出最终的k-means算法聚类结果，并指出"1"属于哪个cluster。

(1). 1(1,2) 2(1,3) 3(1,4)
4(1,5) 5(0,4) 6(2,4)
7(3,4) 8(4,4), 9(5,4)
10(4,3) 11(4,5)



分配到簇时：计算欧式距离

更新中心：每一簇所有点的中心

第一次：5簇：$x=\frac{1+1+1+1+1+0+2}{6}=\frac{7}{6}$ $y=\frac{2+3+4+5+4+4}{6}=\frac{22}{6}$ $(\frac{7}{6},\frac{22}{6})$

9簇：$x=\frac{3+4+5+4+4}{5}=\frac{20}{5}=4$ $y=\frac{4+4+4+3+5}{5}=4$ $(4,4)$

第2次：不变 $(\frac{7}{6},\frac{22}{6})$，$(4,4)$

(2). cluster1：1，2，3，4，5，6

cluster2：7，8，9，10，11

3. 若1是噪声点，设计能抵抗噪声"1"的影响的算法，并阐述基本思想

K中心点算法