



数据预处理



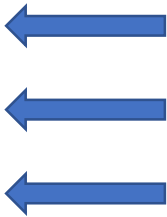
采集的原始数据里存在着各种不利于分析与建模工作的因素，比如数据不完整，格式不正确，数据之间存在矛盾，异常值等。这些因素不仅会影响建模的执行过程，更有甚者在不知不觉间给出错误的建模结果，这就使得数据的预处理显得尤为重要。



通过爬虫爬下来的二手房数据

```
{"name": ["宝盛北里 ", "清河"], "desp": ["2室2厅 | 86.42平米 | 南 北 | 精装 | 中楼层(共6层) | 2004年建 | 板楼"], "price": ["568"]}  
{"name": ["知春东里 ", "双榆树"], "desp": ["3室1厅 | 72.8平米 | 南 北 | 精装 | 高楼层(共6层) | 1987年建 | 板楼"], "price": ["755"]}  
{"name": ["宝盛里 ", "清河"], "desp": ["2室1厅 | 69.52平米 | 南 北 | 精装 | 高楼层(共6层) | 2000年建 | 板楼"], "price": ["460"]}  
{"name": ["魏公村8号院 ", "魏公村"], "desp": ["2室1厅 | 75.6平米 | 西南 | 简装 | 高楼层(共14层) | 1997年建 | 板塔结合"], "price": ["750"]}  
{"name": ["大钟寺甲8号院 ", "皂君庙"], "desp": ["3室1厅 | 101.9平米 | 南 北 | 简装 | 低楼层(共6层) | 1994年建 | 板楼"], "price": ["860"]}  
{"name": ["成府路20号院 ", "五道口"], "desp": ["2室1厅 | 67.71平米 | 南 北 | 简装 | 中楼层(共7层) | 2000年建 | 板楼"], "price": ["630"]}  
{"name": ["橡树湾 ", "清河"], "desp": ["4室2厅 | 187.16平米 | 南 北 | 精装 | 顶层(共16层) | 2016年建 | 板楼"], "price": ["1750"]}  
{"name": ["宝盛里 ", "清河"], "desp": ["1室1厅 | 53.09平米 | 南 | 简装 | 高楼层(共6层) | 2000年建 | 板楼"], "price": ["358"]}  
{"name": ["花园路18号院 ", "北太平庄"], "desp": ["2室1厅 | 66.3平米 | 南 北 西 | 简装 | 高楼层(共12层) | 1998年建 | 板塔结合"], "price": ["580"]}  
{"name": ["西单宿舍 ", "牡丹园"], "desp": ["2室1厅 | 54.1平米 | 南 北 | 精装 | 中楼层(共6层) | 1986年建 | 板楼"], "price": ["510"]}  
{"name": ["静淑苑 ", "学院路"], "desp": ["2室1厅 | 68.3平米 | 南 北 | 简装 | 中楼层(共7层) | 1996年建 | 板楼"], "price": ["599"]}  
{"name": ["铁东小区 ", "军博"], "desp": ["2室1厅 | 54.6平米 | 南 北 | 简装 | 中楼层(共6层) | 1990年建 | 板楼"], "price": ["478"]}  
{"name": ["锦秋知春 ", "知春路"], "desp": ["1室1厅 | 86.29平米 | 东 | 精装 | 中楼层(共9层) | 2003年建 | 板楼"], "price": ["780"]}
```

JSON文件->CSV文件

- 名字：有空格
 - 价格：字符串，不方便进行后续的计算
 - 描述部分：内容太杂
- 
- 去掉空格
 - 将其转换为数字
 - 分成多列，分别是房型，面积，朝向、装修情况等
 - 增加一列单价，并将数据按照单价倒排序



作业1：把通过爬虫爬下来的新房数据，进行预处理。

- 最终的csv文件，应包括以下字段：名称，地理位置（3个字段分别存储），房型（只保留最小房型），面积（按照最小值），总价（万元，整数），均价（万元，保留小数点后4位）；
- 对于所有字符串字段，要求去掉所有的前后空格；
- 如果有缺失数据，不用填充。





公园十七区

住宅

在售

关注

顺义 / 后沙峪 / 中央别墅区火沙路和裕庆路交汇口北500米

3室 / 4室

建面 90-144m²

55711 元/平(均价)

总价475万/套起

免费专车

明星户型

限竞房

多轨交汇

三甲医院



V7九间堂

别墅

在售

通州 / 潞苑 / 通燕高速耿庄桥北出口中化石油对面

4室 / 5室 / 6室 / 8室

建面 220-420m²

1700 万/套(总价)

总价1700万/套起

私属庭院

入户花园

环线房

近主干道

名称，地理位置（3个字段分别存储），房型（只保留最小房型），面积（按照最小值，整数），均价（元，整数），总价（万元，保留小数点后4位）。

公园十七区，顺义，后沙峪，中央别墅区火沙路和裕庆路交汇口北500米，3室，90，55711，501.3990。注：最后一项501.3990由面积乘以单价计算得出。

V7九间堂，通州，潞苑，通燕高速耿庄桥北出口中化石油对面，4室，220，45455，1700.0000。注：45455由1700万除以面积220得出，只保留整数。



雾霾指数数据分析

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
No	year	month	day	hour	season	PM_Taiyua	PM_US Po	PM_Xiaoh	DEWP	HUMI	PRES	TEMP	cbwd	lws	precipitatio	lprec
1	2010	1	1	0	4	NA	NA	NA	-26	69.79	1024	-22	NE	1.0289	NA	NA
2	2010	1	1	1	4	NA	NA	NA	-26	76.26	1024	-23	NE	2.5722	NA	NA
3	2010	1	1	2	4	NA	NA	NA	-27	69.56	1023	-23	NE	5.1444	NA	NA
4	2010	1	1	3	4	NA	NA	NA	-27	69.56	1023	-23	NE	7.7166	NA	NA
5	2010	1	1	4	4	NA	NA	NA	-27	69.56	1022	-23	NE	9.7744	NA	NA
6	2010	1	1	5	4	NA	NA	NA	-26	76.26	1022	-23	NE	11.8322	NA	NA
7	2010	1	1	6	4	NA	NA	NA	-25	76.46	1021	-22	NE	14.4044	NA	NA
8	2010	1	1	7	4	NA	NA	NA	-24	70.26	1021	-20	NE	16.9766	NA	NA
9	2010	1	1	8	4	NA	NA	NA	-23	70.49	1021	-19	NE	19.0344	NA	NA
10	2010	1	1	9	4	NA	NA	NA	-22	70.71	1021	-18	NE	21.6066	NA	NA
11	2010	1	1	10	4	NA	NA	NA	-20	77.39	1022	-17	NE	24.1788	NA	NA
12	2010	1	1	11	4	NA	NA	NA	-18	77.75	1021	-15	NE	27.2655	NA	NA
13	2010	1	1	12	4	NA	NA	NA	-17	77.92	1020	-14	NE	29.8377	NA	NA
14	2010	1	1	13	4	NA	NA	NA	-16	78.1	1019	-13	NE	32.9244	NA	NA
15	2010	1	1	14	4	NA	NA	NA	-15	84.87	1018	-12	NE	35.4066	NA	NA

- No: 行号
- year: 年份
- month: 月份
- day: 日期
- hour: 小时
- season: 季节
- PM: PM2.5浓度 (ug/m³)
- DEWP: 露点 (摄氏温度) 指在固定气压之下, 空气中所含的气态水达到饱和而凝结成液态水所需要降至的温度。
- TEMP: Temperature (摄氏温度)
- HUMI: 湿度 (%)
- PRES: 气压 (hPa)c
- cbwd: 组合风向
- lws: 累计风速 (m/s)
- precipitation: 降水量/时 (mm)
- lprec: 累计降水量 (mm) mm)



作业2：计算北京空气质量数据

1. 汇总计算PM指数年平均值的变化情况

2. 汇总计算每年中1-12月的PM指数数据变化情况

- No: 行号
- PM: PM2.5浓度 ($\mu\text{g}/\text{m}^3$)
- PRES: 气压 (hPa)
- year: 年份
- DEWP: 露点 (摄氏温度) 指在固定气压
- cbwd: 组合风向
- month: 月份
- 之下, 空气中所含的气态水达到饱和
- lws: 累计风速 (m/s)
- day: 日期
- 而凝结成液态水所需要降至的温度。
- precipitation: 降水量/时 (mm)
- hour: 小时
- TEMP: Temperature (摄氏温度)
- lprec: 累计降水量 (mm)
- season: 季节
- HUMI: 湿度 (%)



第1节 数据缺失值的处理

第2节 异常值的处理

第3节 数据归一化

第4节 数据连续属性离散化



- 1.忽略, 不参与计算
- 2.删除
- 3.插值



沈阳空气质量数据， 计算PM指数年平均值的变化情况

No	year	month	day	hour	season	PM_Taiyua	PM_US Pos	PM_Xiaoheyan	DEWP	HUMI	PRES	TEMP	cbwd	lws	precipitation	lprec
1	2010	1	1	0	4	NA	NA	NA	-26	69.79	1024	-22	NE	1.0289	NA	NA
2	2010	1	1	1	4	NA	NA	NA	-26	76.26	1024	-23	NE	2.5722	NA	NA
3	2010	1	1	2	4	NA	NA	NA	-27	69.56	1023	-23	NE	5.1444	NA	NA
4	2010	1	1	3	4	NA	NA	NA	-27	69.56	1023	-23	NE	7.7166	NA	NA
5	2010	1	1	4	4	NA	NA	NA	-27	69.56	1022	-23	NE	9.7744	NA	NA
6	2010	1	1	5	4	NA	NA	NA	-26	76.26	1022	-23	NE	11.8322	NA	NA
7	2010	1	1	6	4	NA	NA	NA	-25	76.46	1021	-22	NE	14.4044	NA	NA
8	2010	1	1	7	4	NA	NA	NA	-24	70.26	1021	-20	NE	16.9766	NA	NA
9	2010	1	1	8	4	NA	NA	NA	-23	70.49	1021	-19	NE	19.0344	NA	NA
10	2010	1	1	9	4	NA	NA	NA	-22	70.71	1021	-18	NE	21.6066	NA	NA
11	2010	1	1	10	4	NA	NA	NA	-20	77.39	1022	-17	NE	24.1788	NA	NA
12	2010	1	1	11	4	NA	NA	NA	-18	77.75	1021	-15	NE	27.2655	NA	NA
13	2010	1	1	12	4	NA	NA	NA	-17	77.92	1020	-14	NE	29.8377	NA	NA
14	2010	1	1	13	4	NA	NA	NA	-16	78.1	1019	-13	NE	32.9244	NA	NA
15	2010	1	1	14	4	NA	NA	NA	-15	84.87	1019	-13	NE	35.4966	NA	NA
16	2010	1	1	15	4	NA	NA	NA	-15	78.27	1019	-12	NE	38.5833	NA	NA
17	2010	1	1	16	4	NA	NA	NA	-15	78.27	1019	-12	NE	41.1555	NA	NA
18	2010	1	1	17	4	NA	NA	NA	-15	78.27	1020	-12	NE	43.2133	NA	NA
19	2010	1	1	18	4	NA	NA	NA	-16	78.1	1020	-13	NE	45.7855	NA	NA
20	2010	1	1	19	4	NA	NA	NA	-17	77.92	1021	-14	NE	48.3577	NA	NA
21	2010	1	1	20	4	NA	NA	NA	-17	84.62	1021	-15	NE	50.4155	NA	NA
22	2010	1	1	21	4	NA	NA	NA	-19	77.57	1022	-16	NE	51.9588	NA	NA
23	2010	1	1	22	4	NA	NA	NA	-20	77.39	1022	-17	NE	53.5021	NA	NA
24	2010	1	1	23	4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
25	2010	1	2	0	4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
26	2010	1	2	1	4	NA	NA	NA	-23	70.49	1023	-19	NE	1.5433	NA	NA
27	2010	1	2	2	4	NA	NA	NA	-24	70.26	1023	-20	NE	2.5722	NA	NA
28	2010	1	2	3	4	NA	NA	NA	-24	76.65	1023	-21	NE	3.6011	NA	NA

针对单元格数据的插值方法：

- interpolate：线性插值
- ffill：前向填充
- bfill：后向填充



第1节 数据缺失值的处理

第2节 异常值的处理

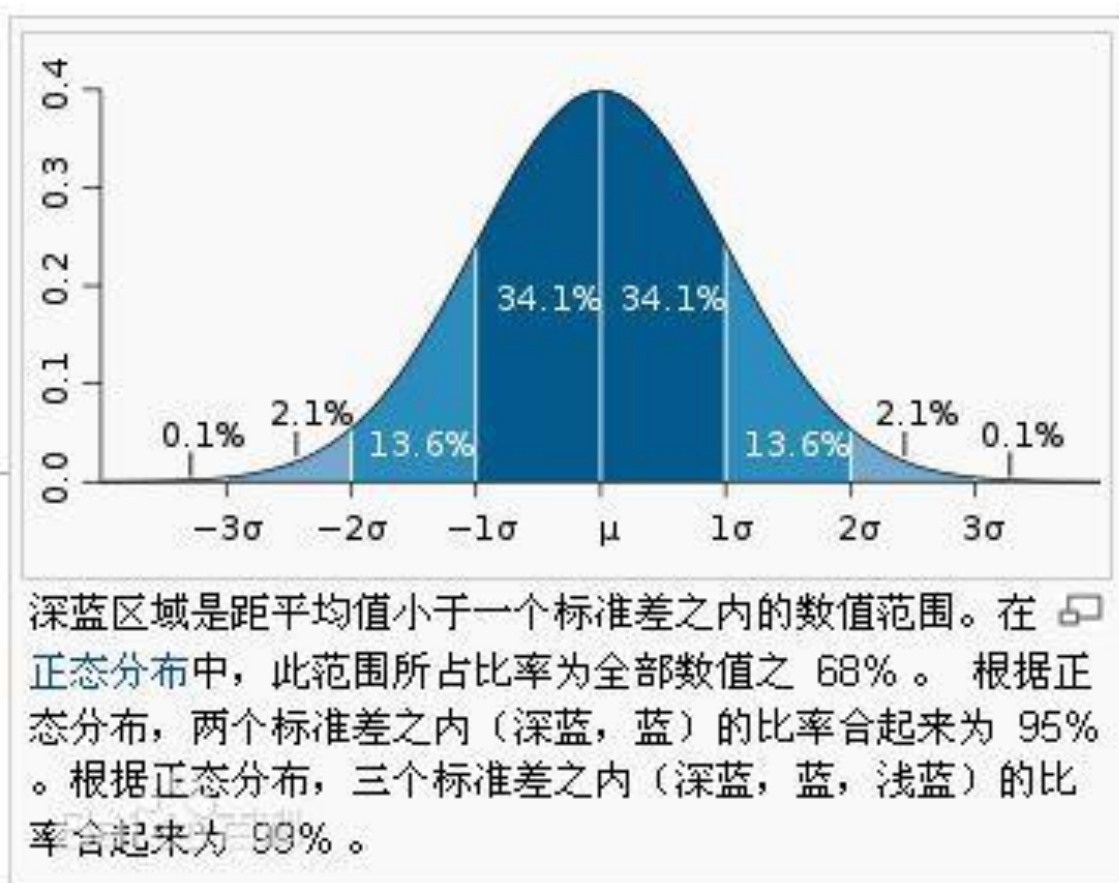
第3节 数据归一化

第4节 数据连续属性离散化



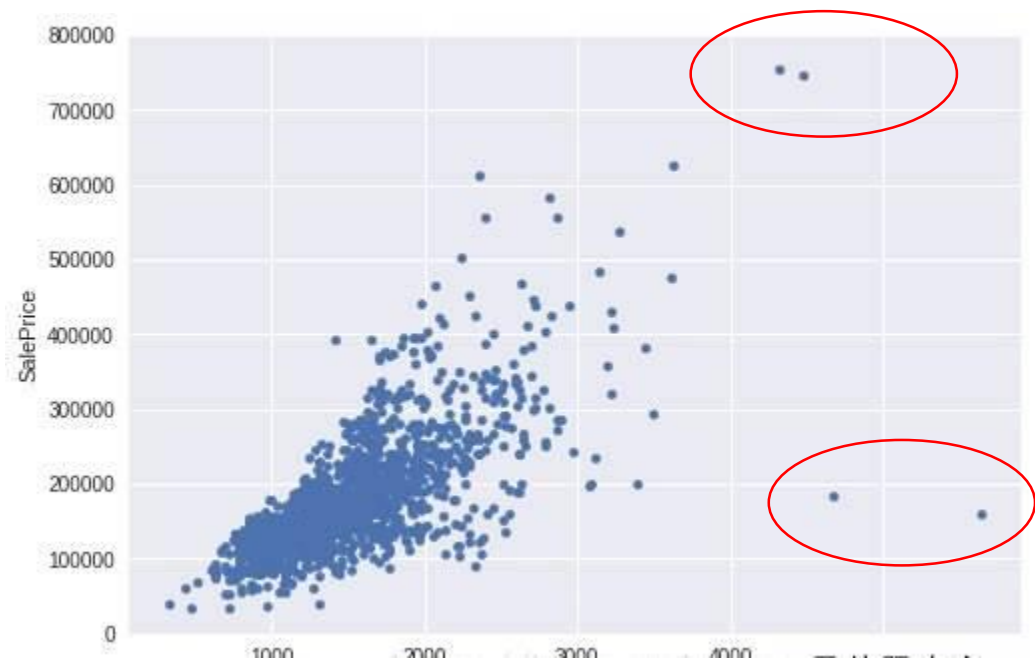
2.数据异常值的处理

- 异常值是指一组测定值中与平均值的偏差超过两倍标准差的测定值;
- 与平均值的偏差超过三倍标准差的测定值, 称为高度异常的异常值。



2.数据异常值的处理

- 发现异常值
 - 观察df的统计信息，使用describe和info函数，查看平均值、最小值和最大值，是否有明显的错误
 - 计算两倍标准差和三倍标准差
 - 使用图形化方式对数据进行展示和分析
- 异常值的处理
 - 直接替换为合理的数据
 - 先置为空，再使用插值的方法进行填充



作业3：处理北京空气质量数据

1. 对HUMI、PRES、TEMP三列，进行线性插值处理。并对其中超过3倍标准差的高度异常数据，修改为3倍标准差的数值。
2. 假设PM指数最高为500，对PM_Dongsi、PM_Dongsihuan、PM_Nongzhanguan三列中超过500的数据，修改为500PM指数进行异常值的处理。
3. 修改cbwd列中值为“cv”的单元格，其值用后项数据填充。



第1节 数据缺失值的处理

第2节 异常值的处理

第3节 数据归一化

第4节 数据连续属性离散化



3.数据归一化

长度1000cm

直径1cm



- 一个钢筋的样品，直径是0.95cm，长度是1010cm
- 直径和标准的差距是-0.05，取平方后是0.0025
- 长度和标准的差距是10，取平方后是100
- 直径的残差被忽略，而长度的残差会带来极大的影响
- 需要统一量纲



3.数据归一化

量纲单位不同

序号	直径	长度
1	0.95	1010
2	0.98	998
3	1.02	1005
...		
...		
...		
N	1.03	996

量纲不同

序号	身高	体重
1	1.78	81
2	1.85	102
3	1.79	75
...		
...		
...		
N	1.68	69

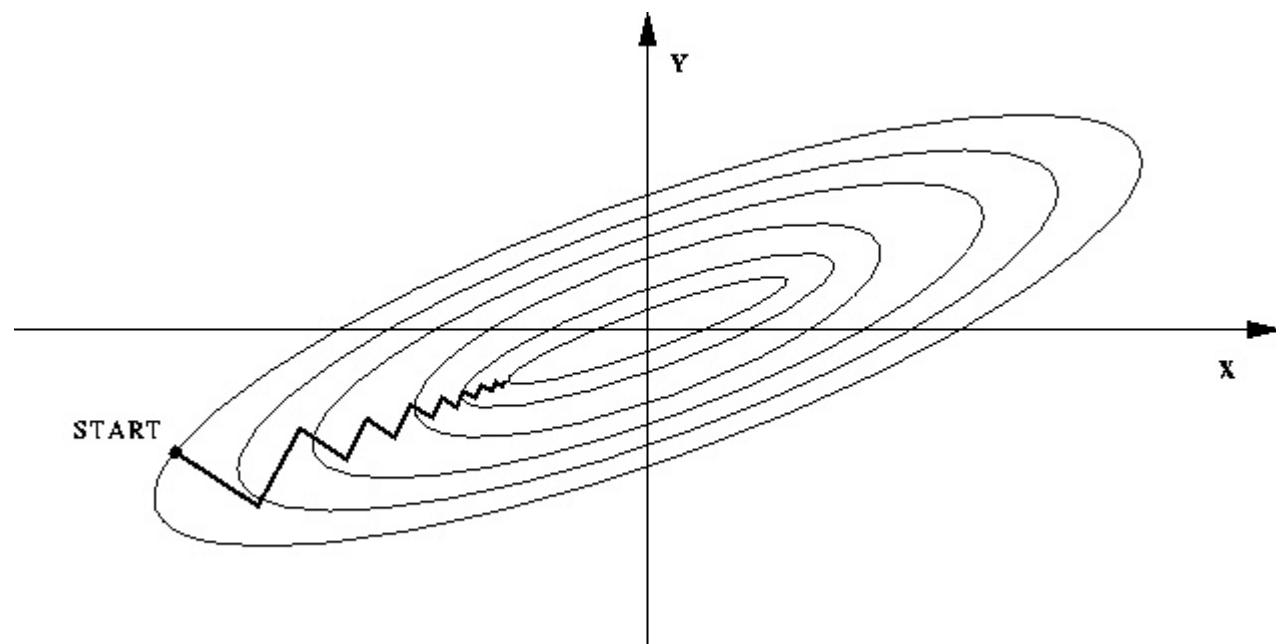


3.数据归一化

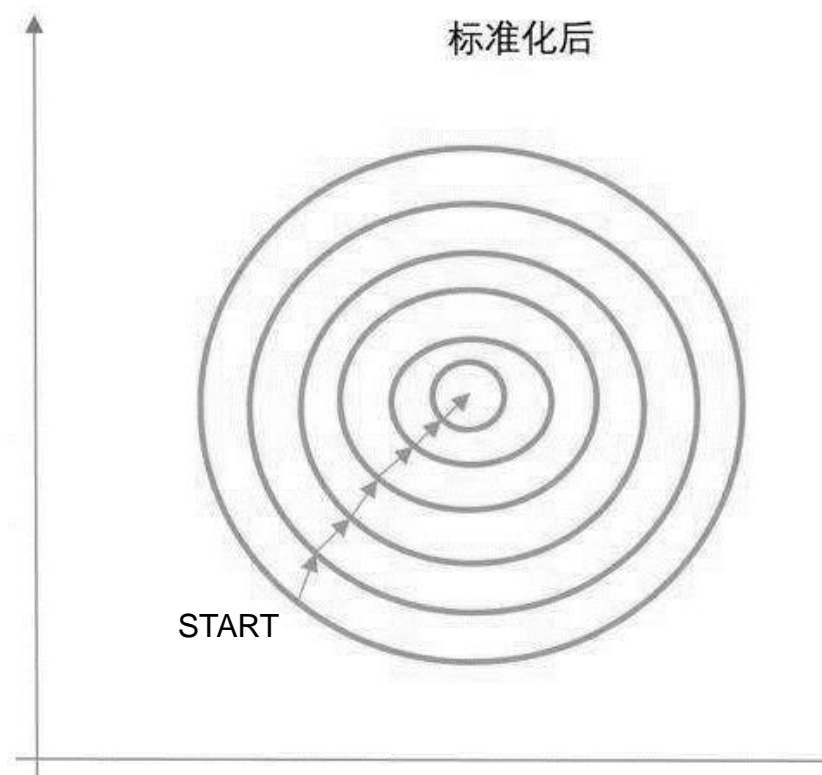
优点：

- (1) 归一化后加快了梯度下降求最优解的速度。
- (2) 归一化有可能提高精度（归一化是让不同维度之间的特征在数值上有一定的比较性）。

标准化前



标准化后



Rescaling (Min-Max归一化，最大最小标准化，离差标准化):

这是一种最简单的归一化，将特征线性映射到[0,1]的范围。

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$



Standardization (Z-score归一化，标准化):

在这种归一化中，对特征进行缩放，使其均值为零，方差为1。

$$x' = \frac{x - \text{average}(x)}{\sigma}$$



3.数据归一化

Rescaling (Min-Max归一化):

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

梯度下降收敛效果好

Standardization (Z-score归一化, 标准化)

$$x' = \frac{x - \text{average}(x)}{\sigma}$$

保留了样本原来的分布

VS



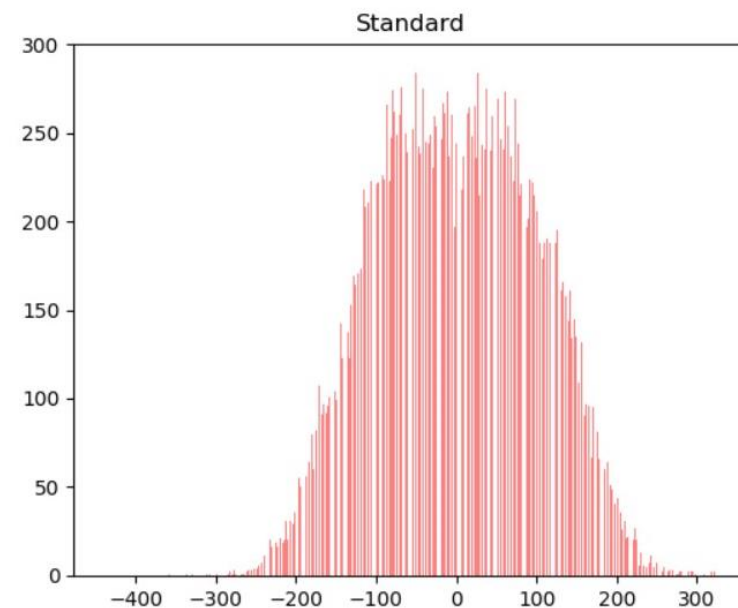
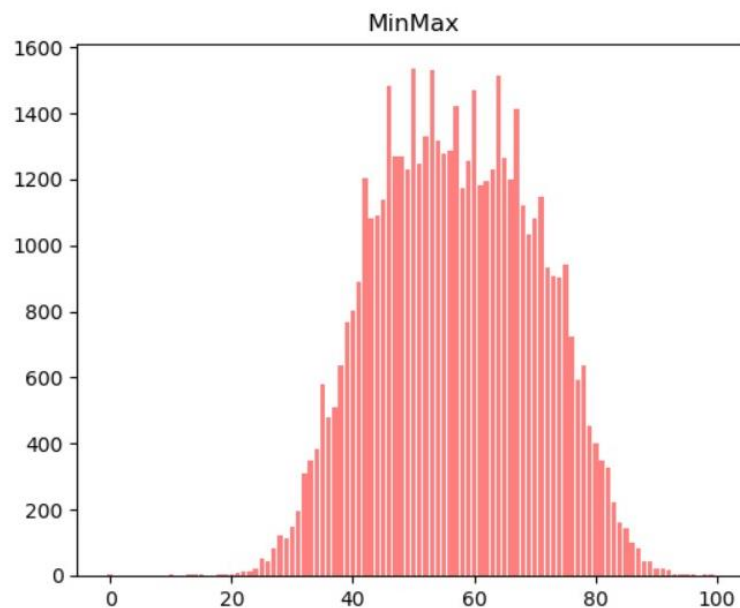
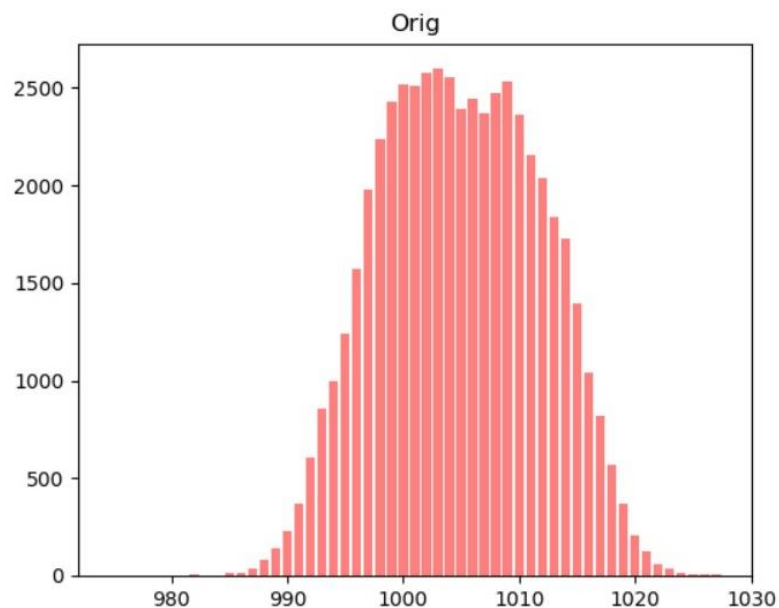
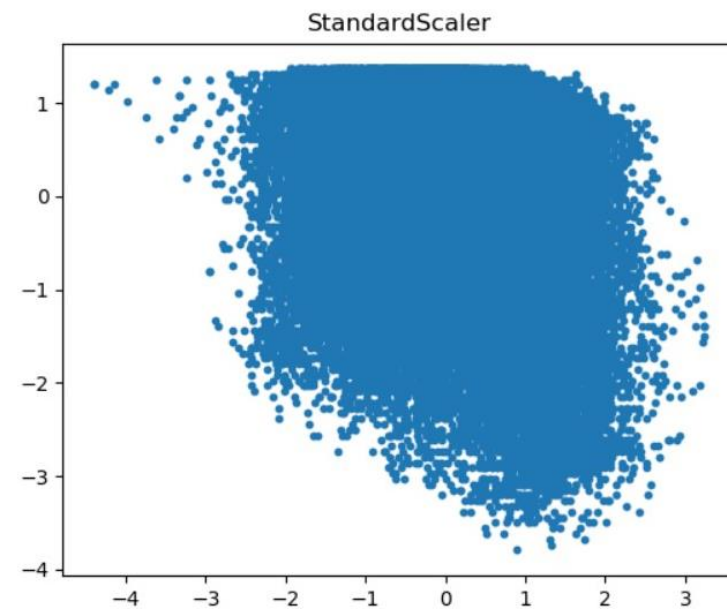
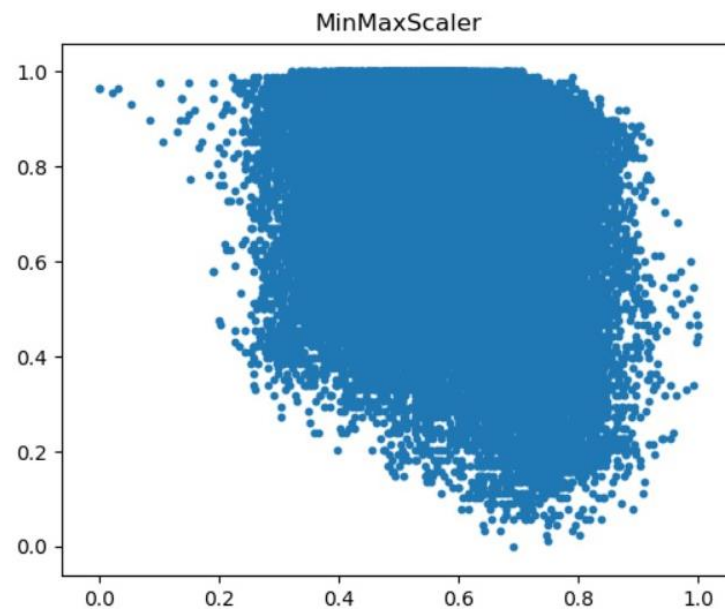
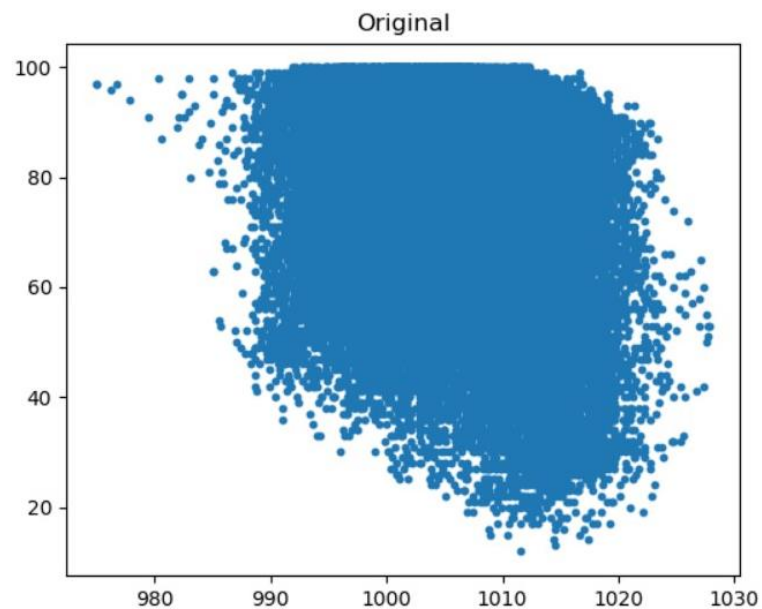
3.数据归一化

```
from sklearn.preprocessing import MinMaxScaler  
scaler = MinMaxScaler()  
#x是df中的某一行，即series对象。  
x_reshape = x.values.reshape(-1, 1) #变成n行1列的二维矩阵形式  
x2 = scaler.fit_transform(x_reshape) #调用MinMaxScaler的fit_transform转换方法，  
进行归一化处理
```

```
from sklearn.preprocessing import StandardScaler  
scaler_std = StandardScaler()  
x_reshape = x.values.reshape(-1, 1) #变成n行1列的二维矩阵形式  
x3 = scaler_std.fit_transform(x_reshape) #调用StandardScaler的fit_transform转换  
方法，进行归一化处理
```



3.数据归一化



第1节 数据缺失值的处理

第2节 异常值的处理

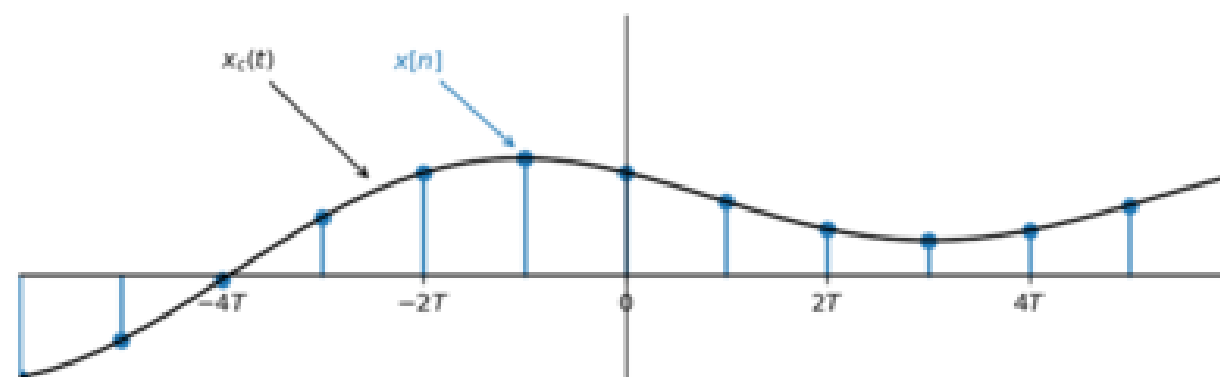
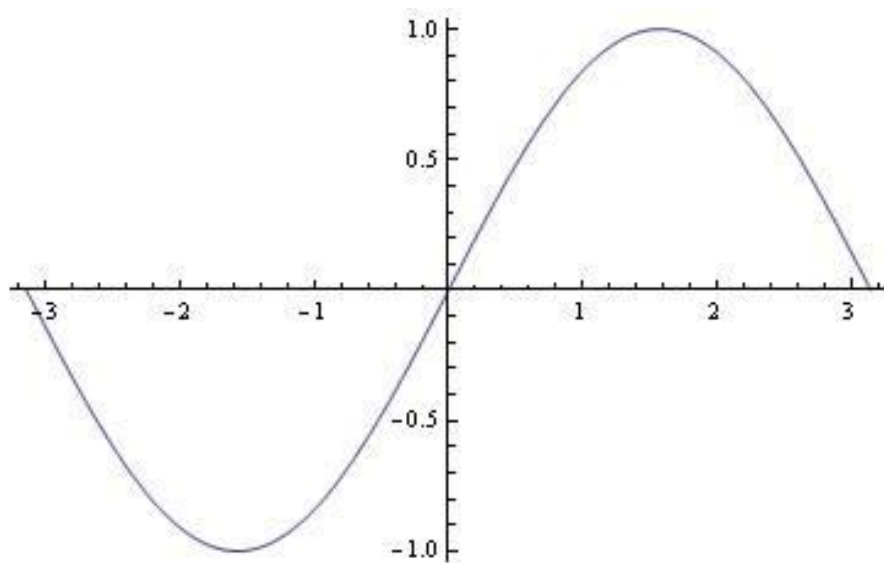
第3节 数据归一化

第4节 数据连续属性离散化



4.数据连续属性离散化

数据的属性分为连续和离散两大类。



离散属性比连续属性更接近于知识级的表达。通过对数据连续属性的离散化，数据可以被减少并被简化。对用户而言,离散的数据更易理解、使用和解释。

4.数据连续属性离散化

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	No	year	month	day	hour	season	PM_Taiyuanjie	PM_US Post	PM_Xiaoheyuan	DEWP	HUMI	PRES	TEMP	cbd	lws	precipitation	lprec	sum	count	ave	DEWP_new	HUMI_new	PRES_new	
2	26304	26305	2013	1	1	0	4	145		148	-17	66.23	1016	-12 SE	24		0	0	293	2	146.5	-17	66.23	1016
3	26305	26306	2013	1	1	1	4	150		133	-16	72.02	1016	-12 SE	26		0	0	283	2	141.5	-16	72.02	1016
4	26307	26308	2013	1	1	3	4	142		121	-14	78.44	1016	-11 cv	1	0.1	0.1	263	2	131.5	-14	78.44	1016	
5	26308	26309	2013	1	1	4	4	105		110	-16	78.1	1016	-13 NW	2		0	0	215	2	107.5	-16	78.1	1016
6	26309	26310	2013	1	1	5	4	154		107	-16	84.74	1016	-14 NE	1		0	0	261	2	130.5	-16	84.74	1016
7	26310	26311	2013	1	1	6	4	176		123	-16	84.74	1016	-14 SW	3		0	0	299	2	149.5	-16	84.74	1016
8	26311	26312	2013	1	1	7	4	140		111	-18	71.6	1018	-14 NW	5		0	0	251	2	125.5	-18	71.6	1018
9	26312	26313	2013	1	1	8	4	93		76	-18	71.6	1019	-14 NW	9		0	0	169	2	84.5	-18	71.6	1019
10	26313	26314	2013	1	1	9	4	53		56	-18	71.6	1020	-14 NW	12		0	0	109	2	54.5	-18	71.6	1020
11	26314	26315	2013	1	1	10	4	23		29	-22	54.98	1021	-15 NW	19		0	0	52	2	26	-22	54.98	1021
12	26315	26316	2013	1	1	11	4	29		20	-24	46.01	1021	-15 NW	24		0	0	49	2	24.5	-24	46.01	1021
13	26317	26318	2013	1	1	13	4	13		7	-24	46.01	1022	-15 NW	33		0	0	20	2	10	-24	46.01	1022
14	26318	26319	2013	1	1	14	4	8		6	-25	42.03	1022	-15 NW	37		0	0	14	2	7	-25	42.03	1022
15	26319	26320	2013	1	1	15	4	9		8	-26	38.37	1023	-15 NW	41		0	0	17	2	8.5	-26	38.37	1023
16	26320	26321	2013	1	1	16	4	14		9	-27	38.03	1024	-16 NW	44		0	0	23	2	11.5	-27	38.03	1024
17	26321	26322	2013	1	1	17	4	24		14	-27	41.36	1024	-17 NW	47		0	0	38	2	19	-27	41.36	1024
18	26322	26323	2013	1	1	18	4	34		24	-27	45.02	1025	-18 NW	49		0	0	58	2	29	-27	45.02	1025
19	26323	26324	2013	1	1	19	4	37		20	-27	45.02	1026	-18 NW	52		0	0	57	2	28.5	-27	45.02	1026
20	26324	26325	2013	1	1	20	4	25		17	-28	44.68	1026	-18 NW	53	截图[Alt + A]	0	0	42	2	21	-28	44.68	1026
21	26325	26326	2013	1	1	21	4	18		25	-28	53.13	1026	-21 SE	2		0	0	43	2	21.5	-28	53.13	1026
22	26326	26327	2013	1	1	22	4	21		77	-29	57.71	1026	-23 SE	4		0	0	98	2	49	-29	57.71	1026
23	26327	26328	2013	1	1	23	4	30		59	-30	62.84	1027	-25 cv	1		0	0	89	2	44.5	-30	62.84	1027
24	26328	26329	2013	1	2	0	4	43		26	-30	62.84	1027	-25 SW	2		0	0	69	2	34.5	-30	62.84	1027
25	26330	26331	2013	1	2	2	4	25		23	-30	62.84	1028	-25 SW	5		0	0	48	2	24	-30	62.84	1028
26	26331	26332	2013	1	2	3	4	19		13	-31	68.59	1028	-27 SW	7		0	0	32	2	16	-31	68.59	1028
27	26332	26333	2013	1	2	4	4	36		12	-31	62.56	1029	-26 SW	8		0	0	48	2	24	-31	62.56	1029
28	26333	26334	2013	1	2	5	4	22		9	-31	68.59	1029	-27 SW	10		0	0	31	2	15.5	-31	68.59	1029
29	26334	26335	2013	1	2	6	4	16		8	-31	68.59	1030	-27 SW	12		0	0	24	2	12	-31	68.59	1030
30	26335	26336	2013	1	2	7	4	19		10	-31	68.59	1031	-27 SE	1		0	0	29	2	14.5	-31	68.59	1031
31	26336	26337	2013	1	2	8	4	24		13	-30	68.83	1032	-26 SW	2		0	0	37	2	18.5	-30	68.83	1032
32	26337	26338	2013	1	2	9	4	30		14	-28	63.39	1033	-23 SW	4		0	0	44	2	22	-28	63.39	1033



实时空气质量指数分级相关信息

指数值	指数等级 及表征颜色	对健康影响情况	建议采取的措施
0~50	优 绿色	空气质量令人满意,基本无空气污染	各类人群可正常活动
51~100	良 黄色	空气质量可接受,但某些污染物可能对极少数异常敏感人群健康有较弱影响	极少数异常敏感人群应减少户外活动
101~150	轻度污染 橙色	易感人群症状有轻度加剧,健康人群出现刺激症状	儿童、老年人及心脏病、呼吸系统疾病患者应减少长时间、高强度的户外锻炼
151~200	中度污染 红色	进一步加剧易感人群症状,可能对健康人群心脏、呼吸系统有影响	儿童、老年人及心脏病、呼吸系统疾病患者避免长时间、高强度的户外锻炼,一般人群适量减少户外运动
201~300	重度污染 紫色	心脏病和肺病患者症状显著加剧,运动耐力降低,健康人群普遍出现症状	儿童、老年人和心脏病、肺病患者应停留在室内,停止户外运动,一般人群减少户外运动
>300	严重污染 褐红色	健康人运动耐力降低,有明显强烈症状,提前出现某些疾病	儿童、老年人和病人应当停留在室内,避免体力消耗,一般人群应避免户外活动



cut方法：按值切割，根据数据值的大小范围分成n组，落入这个范围的分别进入到该组。

- 设定区间的个数，每个区间的间距相等
- 也可自定义每个区间的长度

```
pandas.cut(x, bins, right=True, labels=None, retbins=False, precision=3,  
include_lowest=False, duplicates='raise')
```

x：数据集，这里一般是pandas的Series

bins：为一个整数或数组，代表切割成几组或者具体的切割方式

labels：代表切割后的分组名称

Right：表示区间右端点的数据是否包含在内，默认为包含



qcut方法：按个数切割，使得每个区间里的元素个数基本相同

```
pandas.qcut(x, q, labels=None, retbins=False,  
precision=3, duplicates='raise')
```

x: 数据集，这里一般是pandas的Series

q: 为一个整数或分位数数组

labels: 代表切割后的分组名称



cut方法：按值切割

- 设定区间的个数，每个区间的间距相等
- 自定义每个区间的长度

3	1	2	9	5	10	6	4	0	8	7
---	---	---	---	---	----	---	---	---	---	---

5等份：0-2-4-6-8-10 : (0,1,2),(3,4),(5,6),(7,8),(9,10)

4等份：0-2.5-5.0-7.5-10 : (0,1,2),(3,4,5),(6,7),(8,9,10)

指定区间：0,2,7,10 : (1,2),(3,4,5,6,7),(8,9,10)



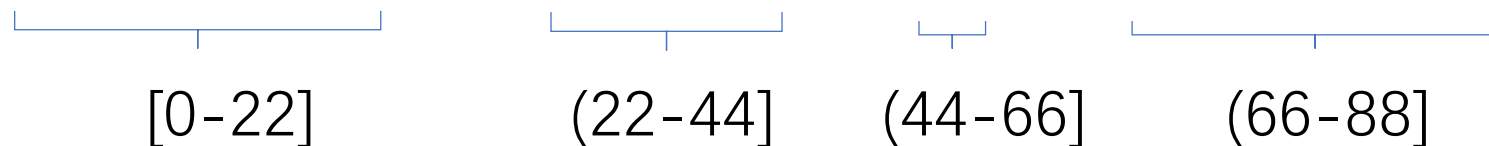
4.数据连续属性离散化

data

11	40	88	50	18	73	23	0	69
----	----	----	----	----	----	----	---	----

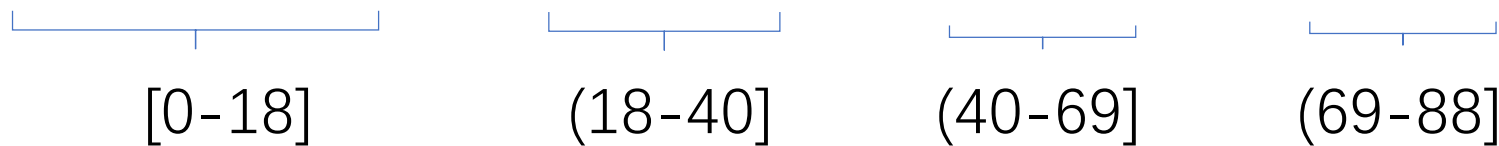
cut(data,4)

0	11	18	23	40	50	69	73	88
---	----	----	----	----	----	----	----	----



qcut (data,4)

0	11	18	23	40	50	69	73	88
---	----	----	----	----	----	----	----	----



4.数据连续属性离散化

```
sections = [0,50,100,150,200,300,1200] #划分为不同长度的区间
section_names=["green","yellow","orange","red","purple",
"Brownish red"] #设置每个区间的标签
result = pd.cut(df.ave,sections,labels=section_names)
print(pd.value_counts(result))
```

```
----- result count -----
green      11587
yellow     8006
orange     3190
red        1518
purple     1121
Brownish red  463
Name: ave, dtype: int64
```



作业4：处理北京空气质量数据

1. 对DEWP和TEMP两列，进行0-1归一化及Z-Score归一化处理。

结果使用散点图的形式表示（参考PPT第24页图形上半部分的表现形式）。

2. 将北京的空气质量数据进行离散化，按照空气质量指数分级标准，计算出每个级别（或颜色值）对应的天数各有多少。

