法律声明

- 本课件包括演示文稿、示例、代码、题库、视频和声音等内容，小象学院和主讲老师拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意及内容，我们保留一切通过法律手段追究违反者的权利。
- 课程详情请咨询
  - 微信公众号：小象
  - 新浪微博：ChinaHadoop

数据清洗和特征选择

小象学院 ChinaHadoop.cn 邹博

主要内容

- 内容
  - 庄家与赔率
  - Pandas数据读取和处理
  - Fuzzywuzzy字符串模糊查找
  - 数据清洗和校正
  - 特征提取主成分分析PCA
  - 手机用户流失率分析
  - One-hot编码
- 思考：
  - 字符串编辑距离
  - ROC与AUC
  - 分类器：随机森林、Logistic回归

本次说明

- 本PPT后面仅列举使用Python库的效果截图，详细内容请参考该PPT的配套代码。

赔率

- 赔率最早出现在赛马中,1790年由英国人奥格登发明。
  - 中国从2001年发行足彩开始引入赔率。
- 赔率的举例定义：
- 浔阳江畔艄公张横和张顺正进行400米自由泳比赛，宋江开赌场做庄，规定：张横赢赔率为3，张顺赢赔率为2。假定不存在平局。赌徒李逵为张横下注10两。
  - 比赛结束后，若最终张横赢，则宋江付赌徒李逵30两(10×3)，赌资10两归庄家宋江所有，即李逵赚20两。
  - 若张顺赢，赌资10两归庄家宋江所有，即李逵赔10两。

问题分析

X	张横	张顺
P	0.8	0.2
y	1.25	5

- 假定张横赢的概率为0.8，宋江给出的赔率为张横1.25，张顺5，则宋江的盈亏分析如下：
  - 为表述方便，张横赢简称“张横”，张顺赢简称“张顺”。
- 假定所有赌徒中，共有a元买张横，b元买张顺，则开赛前宋江收入为a+b元
- 开赛后的赔付期望为：

$$E(y)=\sum_i p_i y_i=0.8\times1.25\times a+0.2\times5\times b=a+b$$

X	张横	张顺
P	0.8	0.2
y	1.25	5

赔率分析

- 从上述结论知：使用y=1/p作为赔率，会使得庄家在期望上不赔不赚。
  - 这即“公平赔率”：y_{fair}
  - ——没有利润，这显然是庄家不希望看到的。
- 实际问题中，庄家总是会将公平赔率乘以某小于1的系数α，即得到真实赔率：

$$y = \alpha \cdot y_{fair} = \alpha / p$$

  - 庄家对于α取值不公开。

身边的故事

- 5月18日《机器学习 升级版V期》最终参团人数是素数还是合数？
- 我坐庄，我开出的赔率是：
  - 素数：5.5
  - 合数：1.1
- 以素数下注1元为例：
  - 若最终人数为素数，则返还5.5元，赌资1元为庄家所有。
  - 若最终人数是合数，则无返还，赌资1元为庄家所有。

计算赔率

```
p = np.array(filter(is_prime3, range(2, b+1)))
p = p[p >= a]
print p
p_rate = float(len(p)) / float(b-a+1)
print '素数的概率：', p_rate, '\t',
print '公正赔率：', 1/p_rate
print '合数的概率：', 1-p_rate, '\t',
print '公正赔率：', 1 / (1-p_rate)
```

```
Prime
[751 757 761 769 773 787 797 809 811 821 823 827 829 839 853 857 859 863 877 881 883 887]
素数的概率： 0.145695364238    公正赔率： 6.86363636364
合数的概率： 0.854304635762    公正赔率： 1.17054263566
```

- 拼团人数当时是603人，尚有两天结束，根据历史先验，假定1天参团人数为100人，则最终参团人数为803左右。结合业务逻辑，往往后期每日参团人数略多于前期，因此大体参团区间可能是[750,900]。
- 计算该区间的素数
  - [751 757 761 769 773 787 797 809 811 821 823 827 829 839 853 857 859 863 877 881 883 887]
  - 素数的概率：0.1457　　公正赔率：　6.8636
  - 合数的概率：0.8543　　公正赔率：　1.1705

计算庄家的盈亏期望

- 实际给出的赔率为5.5和1.1，带入赔率公式 $y = \alpha / p$ 得到α分别是0.8013和0.9397，则庄家盈利期望为：

$$E = (a+b) - E(y) = a + b - (\alpha_1 \cdot a + \alpha_2 \cdot b)$$
$$= (1-\alpha_1) \cdot a + (1-\alpha_2) \cdot b$$

- 若假定a=b，则庄家盈利率为12.94%
  - 即：赌徒不分析场景，随机选边下注。
- 若假定a/b=0.1457/0.8543，则庄家盈利率为8.04%
  - 即：赌徒下注前经过了细致分析，下注比例与实际估计场景概率相符。
- 结论：无论如何，庄家肯定赚。
  - 最终报名人数为1257人，原区间估计结果略显保守。

Pandas

	A	B	C	D	E	F	G	H	I
1	account	name	street	city	state	postal-code	Jan	Feb	Mar
2	211829	Kerluke, Koepp and Hilpert	34456 Sean Highway	New Jaycob	Texas	28752	10000	62000	35000
3	320563	Walter-Trantow	1311 Alvis Tunnel	Port Khadijah	NorthCarolina	38365	95000	45000	35000
4	648336	Bashirian, Kunde and Price	62184 Schamberger Underpass Apt. 231	New Lilianland	Iowa	76517	91000	120000	35000
5	109996	D'Amore, Gleichner and Bode	155 Fadel Crescent Apt. 144	Hyattburgh	Maine	46021	45000	120000	10000
6	121213	Bauch-Goldner	7274 Marissa Common	Shanahanchester	California	49681	162000	120000	35000
7	132971	Williamson, Schumm and Hettinger	89403 Casimer Spring	Jeremieburgh	Arkansas	62785	150000	120000	35000
8	145068	Casper LLC	340 Consuela Bridge Apt. 400	Lake Gabriellaton	Mississipi	18008	62000	120000	70000
9	205217	Kovacek-Johnston	91971 Cronin Vista Suite 601	Deronville	RhodeIsland	53461	145000	95000	35000
10	209744	Champlin-Morar	26739 Grant Lock	Lake Juliannton	Pennsylvania	64415	70000	95000	35000
11	212303	Gerhold-Maggio	366 Maggio Grove Apt. 998	North Ras	Idaho	46308	70000	120000	35000
12	214098	Goodwin, Homenick and Jerde	649 Cierra Forks Apt. 078	Rosaberg	Tenessee	47743	45000	120000	55000
13	231907	Hahn-Moore	18115 Olivine Throughway	Norbertomouth	NorthDakota	31415	150000	10000	162000
14	242368	Frami, Anderson and Donnelly	182 Bertie Road	East Davian	Iowa	72686	162000	120000	35000
15	268755	Walsh-Haley	2624 Beatty Parkways	Goodwinmouth	RhodeIsland	31919	55000	120000	35000
16	273274	McDermott PLC	8917 Bergstrom Meadow	Kathryneborough	Delaware	27933	150000	120000	70000

- Fuzzywuzzy - Levenshtein distance
- 模糊查询与替换

	A	B	C	D	E	F	G	H	I	J	K
1	**account**	**name**	**street**	**city**	**state**	**SC**	**postal-code**	**Jan**	**Feb**	**Mar**	**total**
2	211829	Kerluke, Koepp and Hilpert	34456 Sean Highway	New Jaycob	Texas	TX	28752	10000	62000	35000	107000
3	320563	Walter-Trantow	1311 Alvis Tunnel	Port Khadijah	North Carolina	NC	38365	95000	45000	35000	175000
4	648336	Bashirian, Kunde and Price	62184 Schamberger Underpass Apt. 231	New Lilianland	Iowa	IA	76517	91000	120000	35000	246000
5	109996	D'Amore, Gleichner and Bode	155 Fadel Crescent Apt. 144	Hyattburgh	Maine	ME	46021	45000	120000	10000	175000
6	121213	Bauch-Goldner	7274 Marissa Common	Shanahanchester	California	CA	49681	162000	120000	35000	317000
7	132971	Williamson, Schumm and Hettinger	89403 Casimer Spring	Jeremieburgh	Arkansas	AR	62785	150000	120000	35000	305000
8	145068	Casper LLC	340 Consuela Bridge Apt. 400	Lake Gabriellaton	Mississippi	MS	18008	62000	120000	70000	252000
9	205217	Kovacek-Johnston	91971 Cronin Vista Suite 601	Deronville	Rhode Island	RI	53461	145000	95000	35000	275000
10	209744	Champlin-Morar	26739 Grant Lock	Lake Juliannton	Pennsylvania	PA	64415	70000	95000	35000	200000
11	212303	Gerhold-Maggio	366 Maggio Grove Apt. 998	North Ras	Idaho	ID	46308	70000	120000	35000	225000
12	214098	Goodwin, Homenick and Jerde	649 Cierra Forks Apt. 078	Rosaberg	Tennessee	TN	47743	45000	120000	55000	220000
13	231907	Hahn-Moore	18115 Olivine Throughway	Norbertomouth	North Dakota	ND	31415	150000	10000	162000	322000
14	242368	Frami, Anderson and Donnelly	182 Bertie Road	East Davian	Iowa	IA	72686	162000	120000	35000	317000
15	268755	Walsh-Haley	2624 Beatty Parkways	Goodwinmouth	Rhode Island	RI	31919	55000	120000	35000	210000
16	273274	McDermott PLC	8917 Bergstrom Meadow	Kathryneborough	Delaware	DE	27933	150000	120000	70000	340000
17	0	0	0	0	0		0	1462000	1507000	717000	3686000

手机用户流失率分析

1	region	tenure	age	marital	address	income	ed	employ	retire	gender	reside	tollfree	equip	callcard	wireless	longmon	tollmon	equipmon
2	Zone 2	13	44	Married	9	64	College degree	5	No	Male	2	No	No	Yes	No	3.7	0	0
3	Zone 3	11	33	Married	7	136	Post-undergraduate degree	5	No	Male	6	Yes	No	Yes	Yes	4.4	20.75	0
4	Zone 3	68	52	Married	24	116	Did not complete high school	29	No	Female	2	Yes	No	Yes	No	18.15	18	0
5	Zone 2	33	33	Unmarriec	12	33	High school degree	0	No	Female	1	No	No	No	No	9.45	0	0
6	Zone 2	23	30	Married	9	30	Did not complete high school	2	No	Male	4	No	No	No	No	6.3	0	0
7	Zone 2	41	39	Unmarriec	17	78	High school degree	16	No	Female	1	Yes	No	Yes	No	11.8	19.25	0
8	Zone 3	45	22	Married	2	19	High school degree	4	No	Female	5	No	No	Yes	No	10.9	0	0
9	Zone 2	38	35	Unmarriec	5	76	High school degree	10	No	Male	3	Yes	Yes	Yes	Yes	6.05	45	50.1
10	Zone 3	45	59	Married	7	166	College degree	31	No	Male	5	Yes	No	Yes	No	9.75	28.5	0
11	Zone 1	68	41	Married	21	72	Did not complete high school	22	No	Male	3	No	No	Yes	No	24.15	0	0
12	Zone 2	5	33	Unmarriec	10	125	College degree	5	No	Female	1	No	Yes	No	No	4.85	0	26.15
13	Zone 3	7	35	Unmarriec	14	80	High school degree	15	No	Female	1	Yes	No	Yes	No	7.1	22	0
14	Zone 1	41	38	Married	8	37	High school degree	9	No	Female	3	No	No	Yes	No	8.55	0	0
15	Zone 2	57	54	Married	30	115	College degree	23	No	Female	3	Yes	Yes	No	Yes	15.6	46.25	46.7
16	Zone 2	9	46	Unmarriec	3	25	Did not complete high school	8	No	Female	2	No	No	No	No	4.4	0	0
17	Zone 1	29	38	Married	12	75	Post-undergraduate degree	1	No	Male	4	No	Yes	Yes	No	5.1	0	30.25
18	Zone 3	60	57	Unmarriec	38	162	High school degree	30	No	Male	1	Yes	Yes	Yes	No	16.15	29.75	31.3
19	Zone 3	34	48	Unmarriec	3	49	High school degree	6	No	Female	3	Yes	No	No	No	6.65	18.5	0
20	Zone 2	1	24	Unmarriec	3	20	Did not complete high school	3	No	Male	1	No	No	No	No	1.05	0	0
21	Zone 1	26	29	Married	3	77	College degree	2	No	Male	4	No	Yes	Yes	Yes	6.7	0	48.1
22	Zone 3	6	30	Unmarriec	7	16	Some college	1	No	Female	1	No	Yes	No	Yes	3.75	0	33.8
23	Zone 1	68	52	Married	17	120	Did not complete high school	24	No	Male	2	No	No	Yes	No	20.7	0	0
24	Zone 3	53	33	Unmarriec	10	101	Post-undergraduate degree	4	No	Female	2	No	Yes	Yes	Yes	5.3	0	49.6
25	Zone 3	55	48	Married	19	67	Did not complete high school	25	No	Male	3	No	No	Yes	No	15.05	0	0
26	Zone 3	14	43	Married	18	36	Did not complete high school	5	No	Male	5	Yes	No	Yes	No	12.5	19.75	0
27	Zone 2	1	21	Unmarriec	0	33	High school degree	0	No	Female	3	Yes	No	Yes	No	2.2	20.75	0
28	Zone 2	42	40	Unmarriec	7	37	High school degree	8	No	Female	1	Yes	Yes	Yes	Yes	8.25	23.5	36.9
29	Zone 3	25	33	Married	11	31	Did not complete high school	5	No	Male	4	No	No	No	No	9.1	0	0
30	Zone 1	9	21	Married	1	17	High school degree	2	No	Female	3	No	No	No	No	2.9	0	0
31	Zone 2	13	33	Married	9	19	College degree	0	No	Female	2	No	Yes	No	No	5.55	0	27.35
32	Zone 1	56	37	Married	6	36	Did not complete high school	13	No	Female	2	No	Yes	Yes	No	10.6	0	31.1
33	Zone 1	71	53	Married	27	155	Post-undergraduate degree	12	No	Male	2	Yes	No	Yes	Yes	21	56	0
34	Zone 1	35	50	Married	26	140	High school degree	21	No	Female	4	Yes	No	Yes	No	6.5	27.5	0
35	Zone 1	11	27	Married	8	55	Post-undergraduate degree	0	No	Male	3	No	Yes	No	No	4.8	0	19.55
36	Zone 2	60	46	Married	13	163	Some college	24	No	Male	2	Yes	Yes	Yes	Yes	33.9	38.25	44.65
37	Zone 3	20	35	Married	11	52	College degree	0	No	Male	2	No	Yes	No	No	4.25	0	30.55
38	Zone 2	54	60	Unmarriec	38	211	College degree	25	No	Male	1	Yes	Yes	Yes	Yes	21.15	39.25	46.35
39	Zone 1	44	57	Married	1	186	High school degree	17	No	Male	2	Yes	No	Yes	Yes	9.8	33.5	0
40	Zone 1	11	41	Married	0	39	Did not complete high school	1	No	Female	2	Yes	No	Yes	No	6.55	29.25	0
41	Zone 2	72	57	Unmarriec	34	22	High school degree	35	Yes	Female	1	Yes	No	Yes	No	41.75	49	0
42	Zone 3	10	41	Unmarriec	7	30	Did not complete high school	7	No	Male	1	Yes	No	Yes	No	2.5	19.25	0
43	Zone 2	15	28	Unmarriec	0	29	High school degree	4	No	Female	1	Yes	No	Yes	No	4.25	30	0
44	Zone 2	27	28	Married	4	23	High school degree	8	No	Male	5	No	No	No	No	6.2	0	0
45	Zone 1	9	36	Married	14	62	College degree	10	No	Male	6	No	Yes	No	Yes	5.65	0	46.75

1	cardmon	wiremon	longten	tollten	equipten	cardten	wireten	multline	voice	pager	internet	callwait	forward	confer	ebill	lninc	custcat	churn
2	7.5	0	37.45	0	0	110	0	No	No	No	No	No	Yes	No	No	4.16	Basic service	Yes
3	15.25	35.7	42	211.45	0	125	380.35	No	Yes	Yes	No	Yes	Yes	Yes	No	4.91	Total service	Yes
4	30.25	0	1,300.60	1,247.20	0	2,150.00	0	No	No	No	No	Yes	No	Yes	No	4.75	Plus service	No
5	0	0	288.8	0	0	0	0	No	No	No	No	No	No	No	No	3.5	Basic service	Yes
6	0	0	157.05	0	0	0	0	No	No	No	No	No	Yes	Yes	No	3.4	Plus service	No
7	13.5	0	487.4	798.4	0	570	0	No	No	No	No	Yes	No	No	No	4.36	Plus service	No
8	8.75	0	504.5	0	0	415	0	Yes	No	No	Yes	Yes	No	No	Yes	2.94	E-service	Yes
9	23.25	64.9	239.55	1,873.05	1,820.90	880	2,256.70	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	4.33	Total service	No
10	12	0	449.05	1,240.15	0	505	0	Yes	No	No	No	Yes	Yes	Yes	No	5.11	Plus service	No
11	16.5	0	1,659.70	0	0	1,155.00	0	Yes	No	No	No	No	No	No	No	4.28	E-service	No
12	0	0	17.25	0	110.1	0	0	No	No	Yes	Yes	No	No	No	Yes	4.83	Basic service	Yes
13	23.75	0	47.45	166.1	0	145	0	No	Yes	No	No	Yes	Yes	Yes	No	4.38	Plus service	No
14	41.75	0	308.7	0	0	1,650.00	0	No	No	No	No	No	No	No	No	3.61	Basic service	No
15	0	61.05	825.35	2,624.25	2,590.95	0	3,348.85	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	4.74	Total service	Yes
16	0	0	36.8	0	0	0	0	No	No	No	No	No	No	No	No	3.22	Basic service	No
17	11.25	0	146.25	0	780.8	295	0	Yes	No	No	Yes	No	No	No	No	4.32	E-service	No
18	30	0	946.9	1,767.60	1,788.95	1,795.00	0	Yes	No	No	No	Yes	Yes	Yes	No	5.09	Plus service	No
19	0	0	230.8	614.3	0	0	0	No	No	No	No	Yes	Yes	Yes	No	3.89	Plus service	No
20	0	0	1.05	0	0	0	0	No	No	No	No	No	No	No	No	3	Basic service	No
21	24.25	38.3	140.95	0	1,132.80	610	910.1	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	4.34	Total service	Yes
22	0	18.7	25.65	0	175.3	0	78.2	Yes	No	Yes	Yes	Yes	No	No	No	2.77	E-service	Yes
23	22	0	1,391.05	0	0	1,505.00	0	No	No	No	No	No	Yes	No	No	4.79	Basic service	No
24	26.75	51.4	253.35	0	2,499.80	1,340.00	2,645.15	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	4.62	Total service	No
25	27.25	0	810.45	0	0	1,390.00	0	No	No	No	No	No	No	No	No	4.2	Basic service	No
26	18	0	153.75	273.75	0	215	0	No	No	No	No	Yes	Yes	Yes	No	3.58	Plus service	No
27	40.5	0	2.2	20.75	0	40.5	0	No	No	No	No	Yes	Yes	Yes	No	3.5	Plus service	No
28	28	37.4	399.15	950.65	1,532.90	1,190.00	1,562.00	Yes	Yes	No	Yes	Yes	No	Yes	Yes	3.61	Total service	No
29	0	0	234.95	0	0	0	0	Yes	No	No	No	No	Yes	No	No	3.43	Plus service	No
30	0	0	25.25	0	0	0	0	No	No	No	No	No	No	No	No	2.83	Basic service	No
31	0	0	75.25	0	330.65	0	0	Yes	No	No	Yes	No	No	Yes	Yes	2.94	E-service	No
32	18.25	0	582.6	0	1,756.35	1,005.00	0	Yes	No	No	Yes	No	No	No	Yes	3.58	E-service	No
33	34	49.95	1,519.20	4,064.30	0	2,345.00	3,646.90	Yes	Yes	No	Yes	No	Yes	Yes	Yes	5.04	Total service	No
34	35	0	247.55	1,068.25	0	1,215.00	0	No	No	No	No	Yes	Yes	Yes	No	4.94	Plus service	No
35	0	0	54.1	0	220.4	0	0	Yes	No	No	Yes	No	No	No	No	4.01	E-service	No
36	13.75	55.25	1,947.95	2,326.35	2,645.85	840	3,125.95	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	5.09	Total service	No
37	0	0	82.7	0	547.15	0	0	Yes	No	No	Yes	No	No	No	Yes	3.95	E-service	Yes
38	54.25	54.85	1,228.70	2,102.25	2,421.30	2,785.00	2,895.85	Yes	Yes	Yes	Yes	No	No	Yes	Yes	5.35	Total service	No
39	36	41.65	428.25	1,379.55	0	1,535.00	1,826.70	Yes	Yes	No	No	Yes	Yes	Yes	No	5.23	Plus service	No
40	19.75	0	67.8	303.65	0	200	0	No	Yes	No	No	Yes	Yes	Yes	No	3.66	Plus service	Yes
41	18.75	0	3,043.05	3,581.00	0	1,360.00	0	Yes	No	No	No	Yes	Yes	Yes	No	3.09	Plus service	No
42	53.75	0	31.25	178.2	0	490	0	No	No	No	No	Yes	Yes	Yes	No	3.4	Plus service	No
43	17.75	0	78	426.05	0	255	0	No	No	No	No	Yes	Yes	Yes	No	3.37	Plus service	No
44	0	0	180.15	0	0	0	0	No	No	No	No	No	No	No	No	3.14	Basic service	No
45	0	48.5	43.3	0	386.6	0	364.85	Yes	Yes	Yes	Yes	No	No	Yes	Yes	4.13	Total service	Yes

- region, tenure, age, marital, address, income, ed, employ, retire, gender, reside, tollfree, equip, callcard, wireless, longmon, tollmon, equipmon, cardmon, wiremon, longten, tollten, equipten, cardten, wireten, multline, voice, pager, internet, callwait, forward, confer, ebill, lninc, custcat, churn

Code

```
data = pd.read_csv('tel.csv', skipinitialspace=True, thousands=',')
print u'原始数据: \n', data.head(10)

le = LabelEncoder()
for col in data.columns:
    data[col] = le.fit_transform(data[col])

# 年龄分组
bins = [-1, 6, 12, 18, 24, 35, 50, 70]
data['age'] = pd.cut(data['age'], bins=bins, labels=np.arange(len(bins)-1))

# 取对数
columns_log = ['income', 'tollten', 'longmon', 'tollmon', 'equipmon', 'cardmon',
               'wiremon', 'longten', 'tollten', 'equipten', 'cardten', 'wireten', ]
mms = MinMaxScaler()
for col in columns_log:
    data[col] = np.log(data[col] - data[col].min() + 1)
    # data[col] = pd.cut(data[col], bins=10, labels=np.arange(10))    # 可不做
    data[col] = mms.fit_transform(data[col].values.reshape(-1, 1))

# one-hot编码
columns_one_hot = ['region', 'age', 'address', 'ed', 'reside', 'c
for col in columns_one_hot:
    data = data.join(pd.get_dummies(data[col], prefix=col))

data.drop(columns_one_hot, axis=1, inplace=True)

columns = list(data.columns)
columns.remove('churn')
x = data[columns]
y = data['churn']
print u'分组与One-Hot编码后: \n', x.head(10)

x_train, x_test, y_train, y_test = train_test_split(x, y, train_size=0.75, random_state=0)

clf = RandomForestClassifier(n_estimators=100, criterion='gini', max_depth=12, min_samples_split=5,
                             oob_score=True, class_weight={0: 1, 1: 1/y_train.mean()})
clf.fit(x_train, y_train)
```

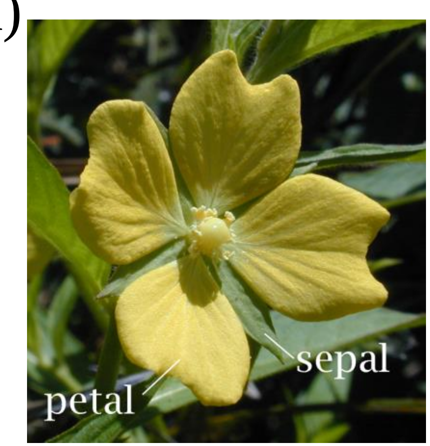
```
OOB Score:  0.742666666667
训练集准确率:  0.988
训练集查准率:  0.962264150943
训练集查全率:  0.99512195122
训练集f1 Score:  0.978417266187
训练集准确率:  0.784
训练集查准率:  0.647058823529
训练集查全率:  0.478260869565
训练集f1 Score:  0.55
```

鸢尾花数据集

- 鸢尾花数据集或许是最有名的模式识别测试数据。
  - 早在1936年，模式识别的先驱Fisher就在论文“The use of multiple measurements in taxonomic problems”中使用了它(直至今日该论文仍然被频繁引用)。
- 该数据集包括3个鸢尾花类别，每个类别有50个样本。其中一个类别是与另外两类线性可分的，而另外两类不能线性可分。
  - 由于Fisher的最原始数据集存在两个错误(35号和38号样本)，实验中我们使用的是修正过的数据。
- 下载链接：http://archive.ics.uci.edu/ml/datasets/Iris

数据描述

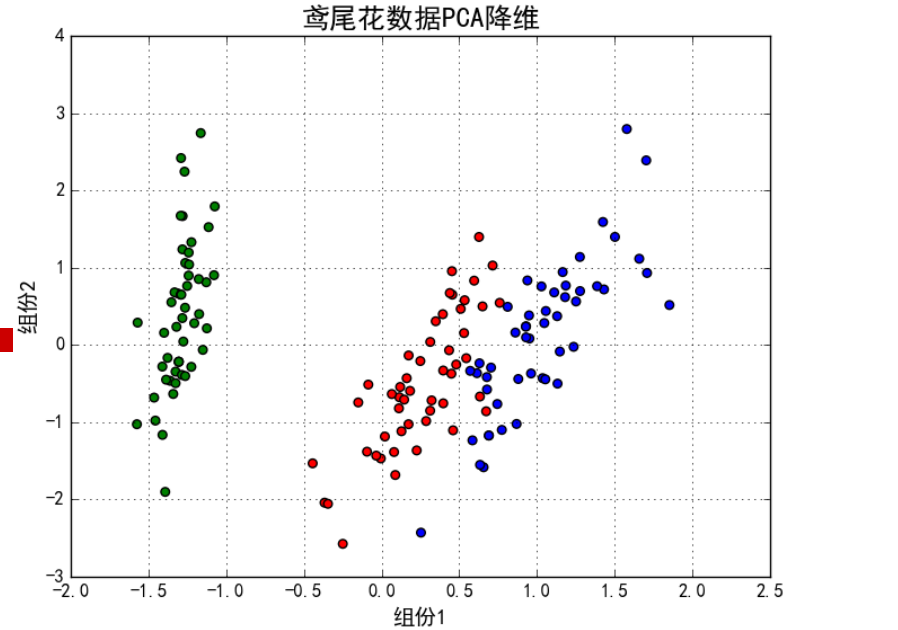
- 该数据集共150行，每行1个样本。每个样本有5个字段，分别是
  - 花萼长度 (单位cm)
  - 花萼宽度(单位：cm)
  - 花瓣长度(单位：cm)
  - 花瓣宽度(单位：cm)
  - 类别(共3类)
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica

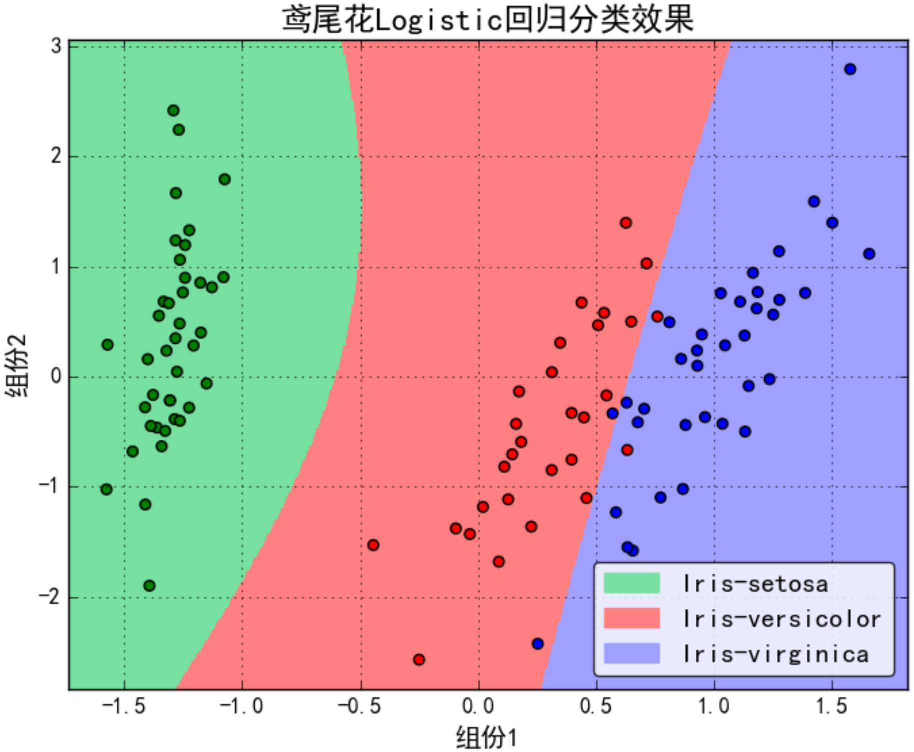


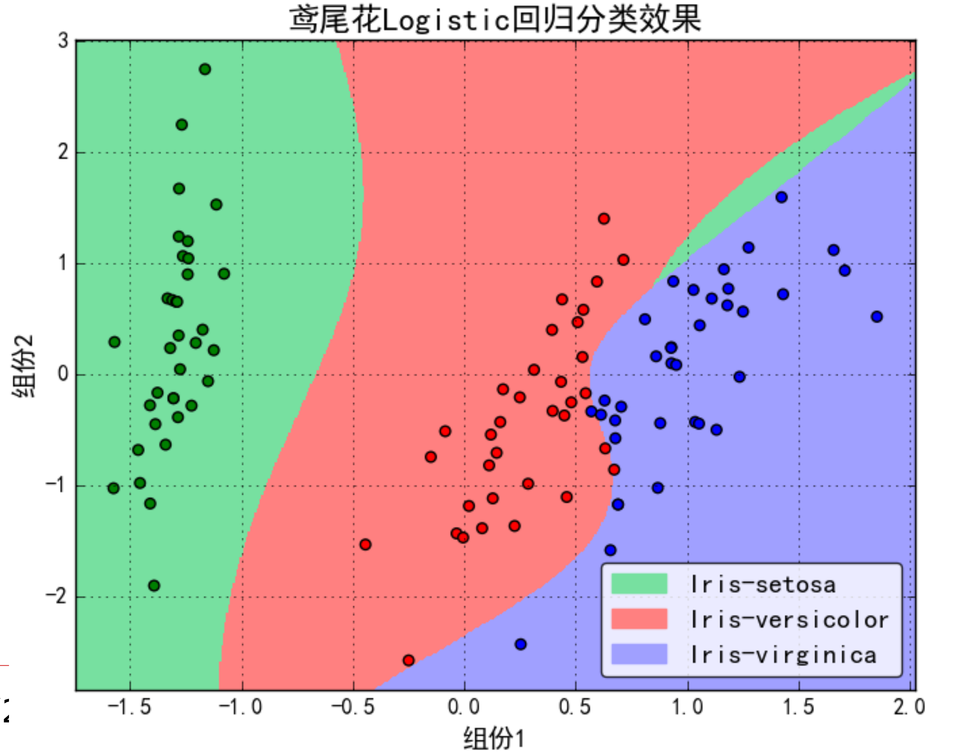

```
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
4.6,3.6,1.0,0.2,Iris-setosa
```

主成分分析PCA

- 多项式特征：2/3
- 管道Pipeline

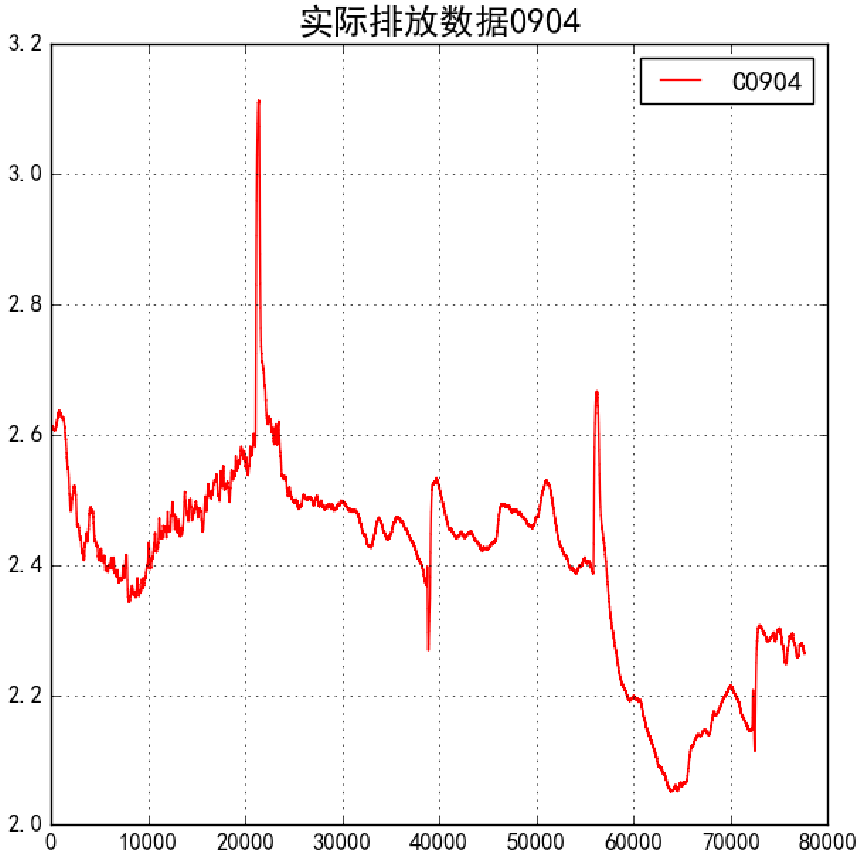


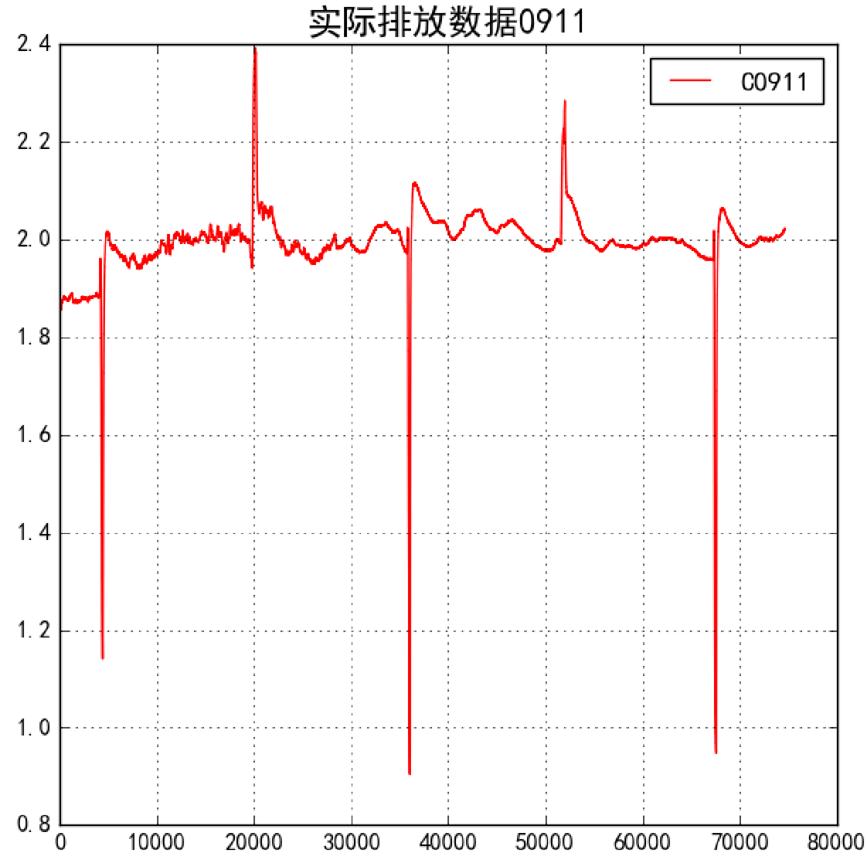


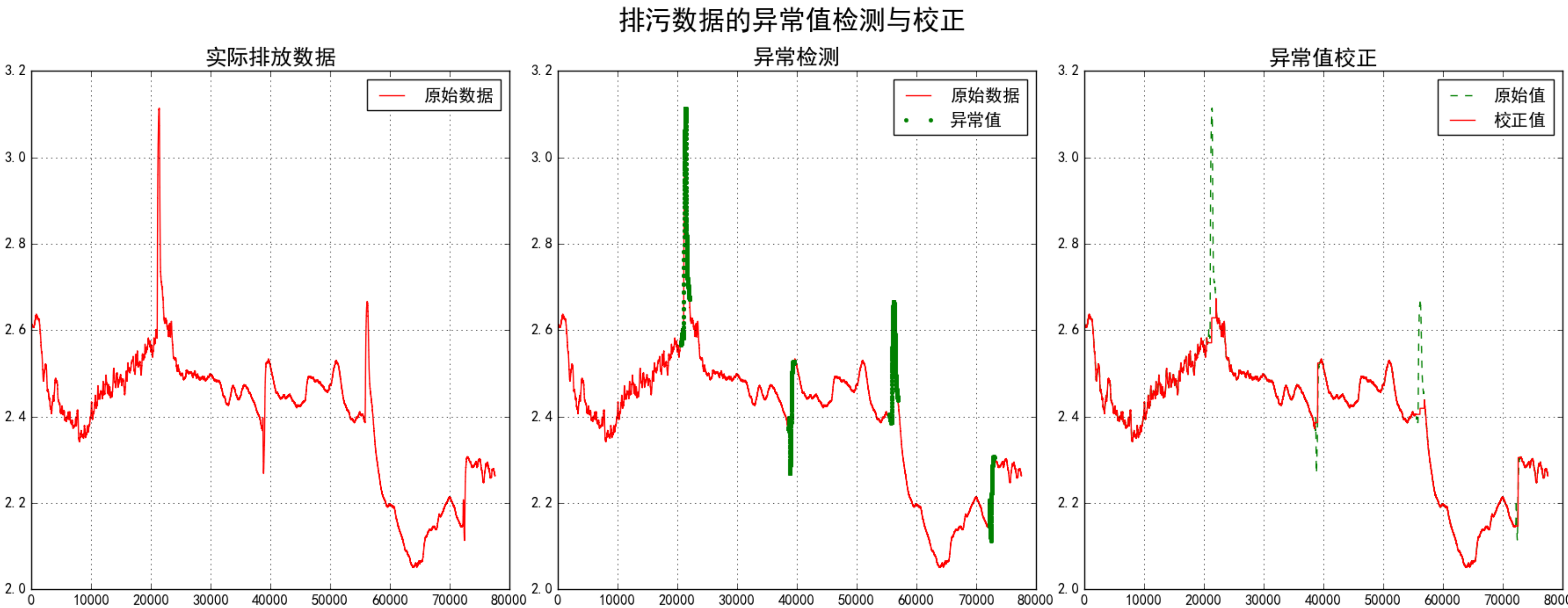


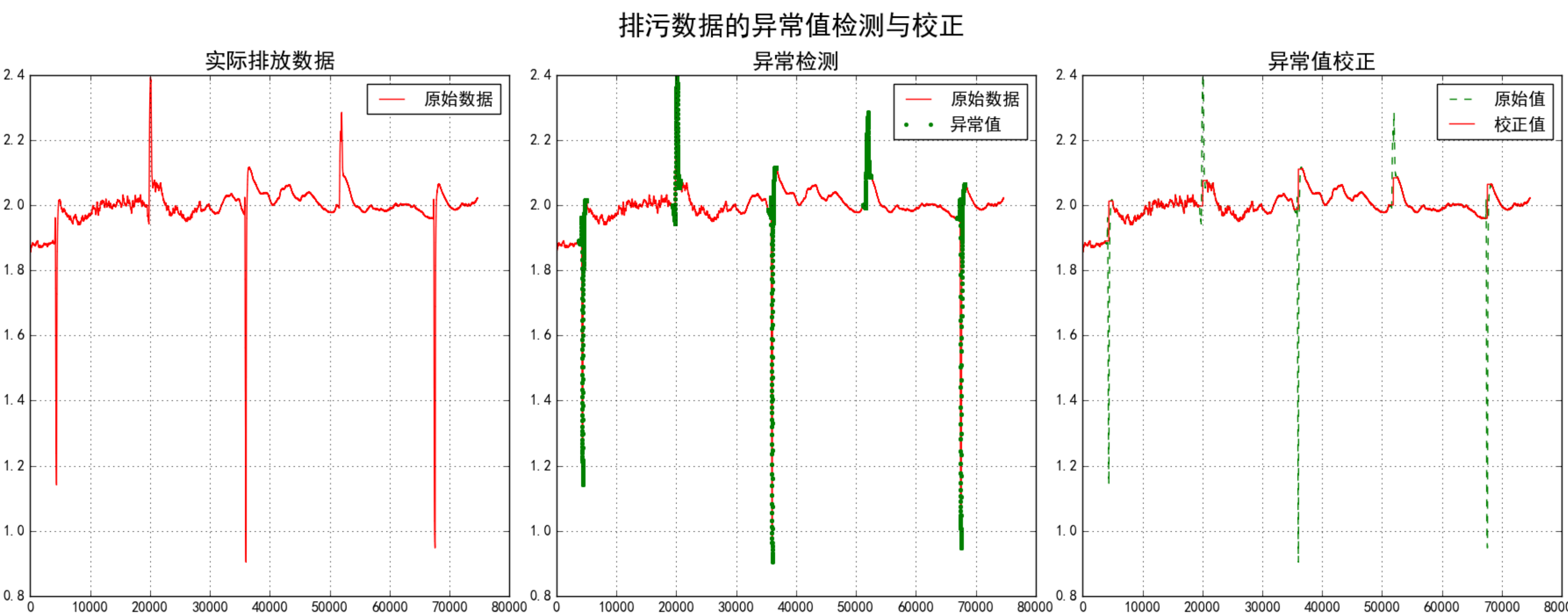
数据清洗和数据处理

如何找到下图中的异常值






排污数据的异常值检测与校正
实际排放数据
原始数据
异常检测
原始数据
异常值
异常值校正
原始值
校正值


排污数据的异常值检测与校正
实际排放数据
原始数据
异常检测
原始数据
异常值
异常值校正
原始值
校正值

车辆数据描述

- 该数据共1728个样本，每行为1个样本。每个样本有7个特征：
  - 购买价格：low/med/high/vhigh
  - 维护价格：low/med/high/vhigh
  - 车门数量：2/3/4/5more
  - 载人数目：2/4/more
  - 后备箱大小：small/med/big
  - 安全程度：low/med/high
  - 接受程度：unacc/acc/good/vgood

	A	B	C	D	E	F	G
1	buy	maintain	doors	persons	boot	safety	accept
2	vhigh	vhigh	2	2	small	low	unacc
3	vhigh	vhigh	2	2	small	med	unacc
4	vhigh	vhigh	2	2	small	high	unacc
5	vhigh	vhigh	2	2	med	low	unacc
6	vhigh	vhigh	2	2	med	med	unacc
7	vhigh	vhigh	2	2	med	high	unacc
8	vhigh	vhigh	2	2	big	low	unacc
9	vhigh	vhigh	2	2	big	med	unacc
10	vhigh	vhigh	2	2	big	high	unacc
11	vhigh	vhigh	2	4	small	low	unacc
12	vhigh	vhigh	2	4	small	med	unacc
13	vhigh	vhigh	2	4	small	high	unacc
14	vhigh	vhigh	2	4	med	low	unacc
15	vhigh	vhigh	2	4	med	med	unacc
16	vhigh	vhigh	2	4	med	high	unacc
17	vhigh	vhigh	2	4	big	low	unacc
18	vhigh	vhigh	2	4	big	med	unacc
19	vhigh	vhigh	2	4	big	high	unacc
20	vhigh	vhigh	2	more	small	low	unacc
21	vhigh	vhigh	2	more	small	med	unacc
22	vhigh	vhigh	2	more	small	high	unacc
23	vhigh	vhigh	2	more	med	low	unacc
24	vhigh	vhigh	2	more	med	med	unacc
25	vhigh	vhigh	2	more	med	high	unacc
26	vhigh	vhigh	2	more	big	low	unacc
27	vhigh	vhigh	2	more	big	med	unacc
28	vhigh	vhigh	2	more	big	high	unacc
29	vhigh	vhigh	3	2	small	low	unacc
30	vhigh	vhigh	3	2	small	med	unacc
31	vhigh	vhigh	3	2	small	high	unacc
32	vhigh	vhigh	3	2	med	low	unacc
33	vhigh	vhigh	3	2	med	med	unacc
34	vhigh	vhigh	3	2	med	high	unacc
35	vhigh	vhigh	3	2	big	low	unacc
36	vhigh	vhigh	3	2	big	med	unacc

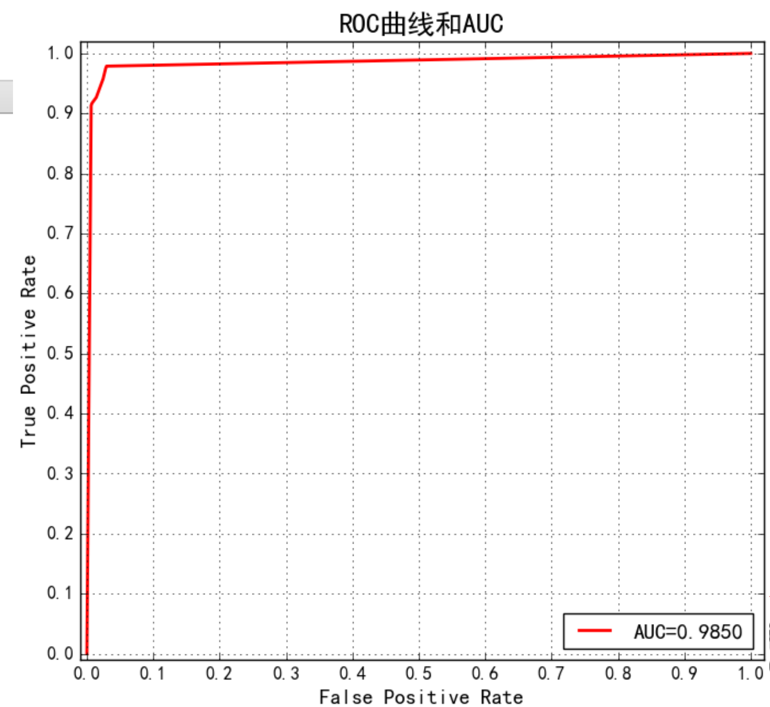
决策树和随机森林分类

```python
x = data.loc[:, columns[:-1]]
y = data['accept']
x, x_test, y, y_test = train_test_split(x, y, train_size=0.7)
if random_forest:
    clf = RandomForestClassifier(n_estimators=100, criterion='gini', max_depth=12, min_samples_split=5
else:
    clf = DecisionTreeClassifier(criterion='gini', max_depth=12, min_samples_split=5, max_features=5)
if cross_validation:
    model = GridSearchCV(clf, param_grid={'max_depth': np.arange(10,20),
                                          'min_samples_split': np.arange(5, 20),
                                          'max_features': np.arange(1, 7)
                                          }, cv=3)
    model.fit(x, y)
    print model.best_params_
```

3.pca 1.RF

1720	1	1	3	2	2	2	0
1721	1	1	3	2	2	0	1
1722	1	1	3	2	1	1	2
1723	1	1	3	2	1	2	1
1724	1	1	3	2	1	0	3
1725	1	1	3	2	0	1	2
1726	1	1	3	2	0	2	1
1727	1	1	3	2	0	0	3

```
[1728 rows x 7 columns]
训练集精确度： 0.988420181969
测试集精确度： 0.967244701349
Micro AUC： 0.991795397255
Micro AUC(System)： 0.991795397255
Macro AUC： 0.987885354719
```



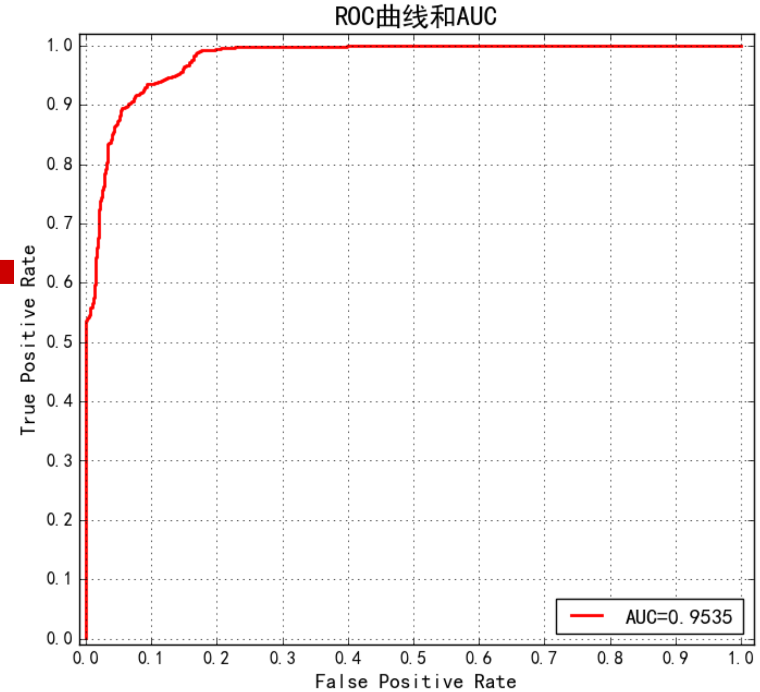
One-hot编码

	buy	maintain	doors	persons	boot	safety	accept
0	vhigh	vhigh	2	2	small	low	unacc
1	vhigh	vhigh	2	2	small	med	unacc
2	vhigh	vhigh	2	2	small	high	unacc
3	vhigh	vhigh	2	2	med	low	unacc
4	vhigh	vhigh	2	2	med	med	unacc
5	vhigh	vhigh	2	2	med	high	unacc
6	vhigh	vhigh	2	2	big	low	unacc
7	vhigh	vhigh	2	2	big	med	unacc
8	vhigh	vhigh	2	2	big	high	unacc
9	vhigh	vhigh	2	4	small	low	unacc

	buy_high	buy_low	buy_med	buy_vhigh	maintain_high	maintain_low	maintain_med	maintain_vhigh	doors_2	doors_3
0	0	0	0	1	0	0	0	1	1	0
1	0	0	0	1	0	0	0	1	1	0
2	0	0	0	1	0	0	0	1	1	0
3	0	0	0	1	0	0	0	1	1	0
4	0	0	0	1	0	0	0	1	1	0
5	0	0	0	1	0	0	0	1	1	0
6	0	0	0	1	0	0	0	1	1	0
7	0	0	0	1	0	0	0	1	1	0
8	0	0	0	1	0	0	0	1	1	0
9	0	0	0	1	0	0	0	1	1	0

doors_4	doors_5more	persons_2	persons_4	persons_more	boot_big	boot_med	boot_small	safety_high	safety_low	safety_med
0	0	1	0	0	0	0	1	0	1	0
0	0	1	0	0	0	0	1	0	0	1
0	0	1	0	0	0	0	1	1	0	0
0	0	1	0	0	0	1	0	0	1	0
0	0	1	0	0	0	1	0	0	0	1
0	0	1	0	0	0	1	0	1	0	0
0	0	1	0	0	1	0	0	0	1	0
0	0	1	0	0	1	0	0	0	0	1
0	0	1	0	0	1	0	0	1	0	0
0	0	0	1	0	0	0	1	0	1	0

Logistic回归




```python
# one-hot编码
x = pd.DataFrame()
for col in columns[:-1]:
    t = pd.get_dummies(data[col])
    t = t.rename(columns=lambda x: col+'_'+str(x))
    x = pd.concat((x, t), axis=1)
print x.head(10)
# print x.columns
y = pd.Categorical(data['accept']).codes

x, x_test, y, y_test = train_test_split(x, y, train_size=0.7)
clf = LogisticRegressionCV(Cs=np.logspace(-3, 4, 8), cv=5)
clf.fit(x, y)
print clf.C_
y_hat = clf.predict(x)
print '训练集精确度：', metrics.accuracy_score(y, y_hat)
y_test_hat = clf.predict(x_test)
print '测试集精确度：', metrics.accuracy_score(y_test, y_test_hat)
n_class = len(data['accept'].unique())
y_test_one_hot = label_binarize(y_test, classes=np.arange(n_class))
y_test_one_hot_hat = clf.predict_proba(x_test)
fpr, tpr, _ = metrics.roc_curve(y_test_one_hot.ravel(), y_test_one_hot_hat.ravel())
print 'Micro AUC:\t', metrics.auc(fpr, tpr)
print 'Micro AUC(System):\t', metrics.roc_auc_score(y_test_one_hot, y_test_one_hot_hat, average='micro')
auc = metrics.roc_auc_score(y_test_one_hot, y_test_one_hot_hat, average='macro')
print 'Macro AUC:\t', auc

mpl.rcParams['font.sans-serif'] = u'SimHei'
mpl.rcParams['axes.unicode_minus'] = False
plt.figure(figsize=(8, 7), dpi=80, facecolor='w')
plt.plot(fpr, tpr, 'r-', lw=2, label='AUC=%.4f' % auc)
plt.legend(loc='lower right')
plt.xlim((-0.01, 1.02))
plt.ylim((-0.01, 1.02))
plt.xticks(np.arange(0, 1.1, 0.1))
plt.yticks(np.arange(0, 1.1, 0.1))
plt.xlabel('False Positive Rate', fontsize=14)
plt.ylabel('True Positive Rate', fontsize=14)
plt.grid(b=True, ls=':')
plt.title(u'ROC曲线和AUC', fontsize=18)
```

训练集精确度： 0.9206
测试集精确度： 0.8651
Micro AUC: 0.9776
Macro AUC: 0.9535

作业

- 除准确率(accuracy)外，还有哪些评价分类模型性能的指标？为什么有这些指标？
- 什么是混淆矩阵？TPR、FPR是什么含义？
  - Precision
  - Recall
  - F1-measure
  - AUC
  - AIC/BIC

我们在这里

- http://wenda.ChinaHadoop.cn
  - 视频/课程/社区
- 微博
  - @ChinaHadoop
  - @邹博_机器学习
- 微信公众号
  - 小象学院
  - 大数据分析挖掘

感谢大家！

恳请大家批评指正！