如何更有效地使用定量变量

从定量到定性

小胖

目录

定量变量的线性陷阱

回顾模型结果

Logit Regression Results

Dep. Variable:	label_code	No. Observations:	6512
Model:	Logit	Df Residuals:	6500
Method:	MLE	Df Model:	11
Date:	Fri, 31 May 2019	Pseudo R-squ.:	0.2732
Time:	12:36:09	Log-Likelihood:	-2611.6
converged:	True	LL-Null:	-3593.4
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.8418	0.304	-25.757	0.000	-8.438	-7.245
C(workclass, contrast_mat, levels=l) State-gov	0.7614	0.278	2.743	0.006	0.217	1.306
C(workclass, contrast_mat, levels=l) Self-emp-not-inc	0.7133	0.254	2.810	0.005	0.216	1.211
C(workclass, contrast_mat, levels=l) Private	0.7371	0.226	3.267	0.001	0.295	1.179
C(workclass, contrast_mat, levels=l) Federal-gov	1.1899	0.280	4.249	0.000	0.641	1.739
C(workclass, contrast_mat, levels=l) Local-gov	0.9964	0.254	3.916	0.000	0.498	1.495
C(workclass, contrast_mat, levels=l) Self-emp-inc	1.6298	0.280	5.820	0.000	1.081	2.179
C(sex)[T. Male]	1.2566	0.090	13.959	0.000	1.080	1.433
education_num	0.3361	0.016	21.347	0.000	0.305	0.367
capital_gain	0.0003	2.13e-05	14.196	0.000	0.000	0.000
capital_loss	0.0009	7.55e-05	12.085	0.000	0.001	0.001
hours_per_week	0.0265	0.003	8.841	0.000	0.021	0.032

第五章的模型结果

预测变量：年收入是否大于50k

系数大于0

$$\ln \frac{P(y=1)}{1-P(y=1)} = X\beta$$

模型系数大于0

随着每星期工作时间的增加，年收入大于50k的概率增加

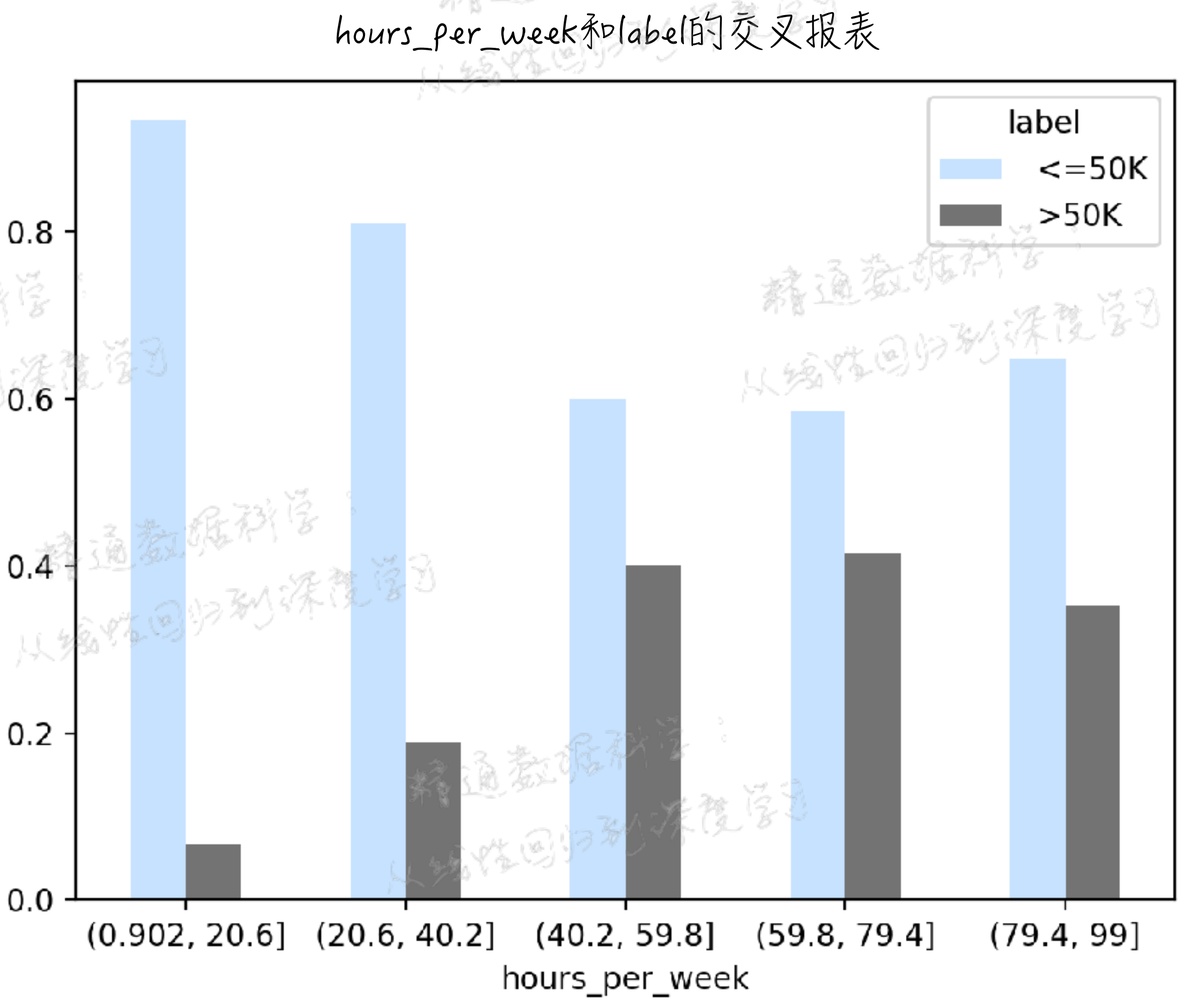
定量变量的线性陷阱

模型结果与事实的差异

随着每星期工作时间的增加，
年收入大于50k的概率增加

模型结果与
事实不符

根据数据，每周工作时间超过
80小时之后，年收入超过50k
的比例反而下降



定量变量的线性陷阱

隐含的边际效应恒定假设

随着每星期工作时间的增加，
年收入大于50k的概率增加

模型结果与
事实不符

根据数据，每周工作时间超过
80小时之后，年收入超过50k
的比例反而下降

发生比　　线性模型

~~边际效应恒定~~

$$\ln\frac{P(y=1)}{1-P(y=1)}=X\beta$$

模型里含有线性部分

直接使用定量变量

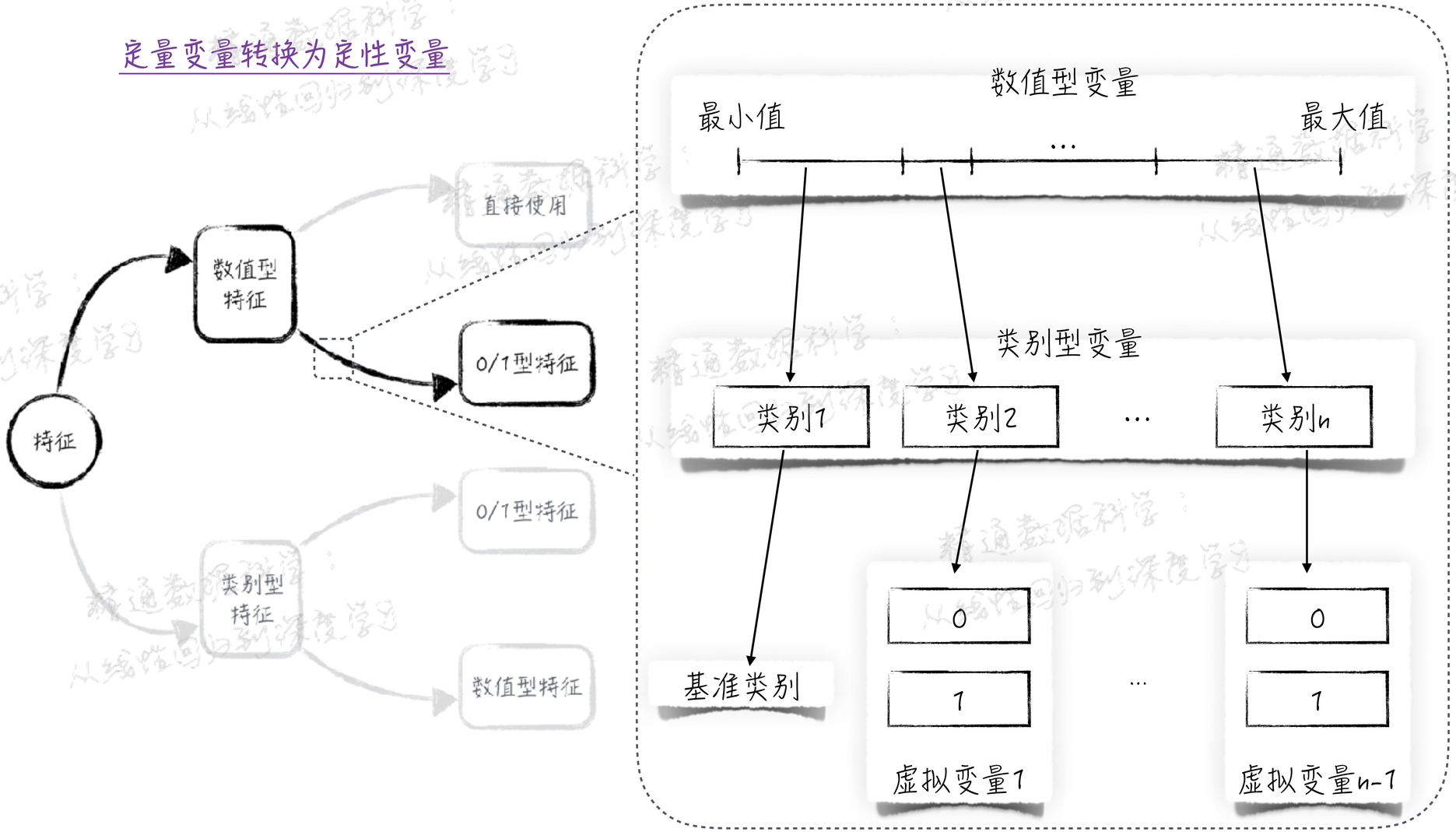
解决问题的
突破口

目录

变量离散化

从定量到定性

定量变量转换为定性变量



变量离散化

从定量到定性

如何选择划分区间？

划分太粗：

- 模型效果欠佳

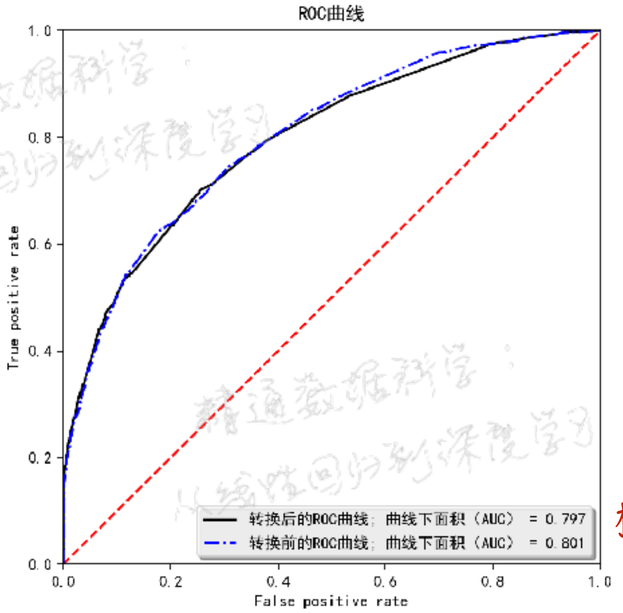
划分太细：

- 过拟合风险上升

每星期工作时间等分为5份

	coef
Intercept	-6.1776
C(hours_per_week_group)[T.20-40]	1.1916
C(hours_per_week_group)[T.40-60]	1.9972
C(hours_per_week_group)[T.60-80]	1.9353
C(hours_per_week_group)[T.80-100]	1.7762
education_num	0.3126
capital_gain	0.0003
capital_loss	0.0008

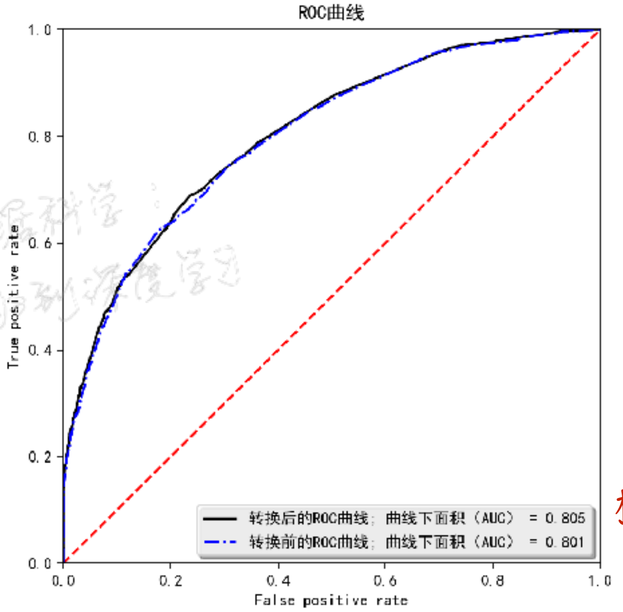
变量系数与数据相符



模型效果下降

每星期工作时间等分为10份

	coef
Intercept	-5.8862
C(hours_per_week_group)[T.10-20]	-0.4022
C(hours_per_week_group)[T.20-30]	-0.1443
C(hours_per_week_group)[T.30-40]	1.0024
C(hours_per_week_group)[T.40-50]	1.6946
C(hours_per_week_group)[T.50-60]	1.7782
C(hours_per_week_group)[T.60-70]	1.6794
C(hours_per_week_group)[T.70-80]	1.6088
C(hours_per_week_group)[T.80-90]	1.7718
C(hours_per_week_group)[T.90-100]	1.2554
education_num	0.3116
capital_gain	0.0003
capital_loss	0.0008



模型效果上升

目录

卡方检验划分法

卡方检验

卡方检验（Chi-square test）用于度量两个类别型变量之间的相关性

- 卡方统计量越大，两个变量之间的相关性也就越大

$$T_{i,j} = \frac{(\text{实际值} - \text{预测值})^2}{\text{预测值}}$$

$$T = \sum T_{i,j}$$

在这个例子中，T服从自由度为(3 -1) X (2 - 1) = 2的卡方分布

Contingency Table(列联表)

变量A的类别

变量B的类别	A1	A2	A3	总计
B1	1	1	3	5
B2	1	1	10	12
总计	2	2	13	17

类别A3,B2包含10个数据

预测值：

$$17 \times P(A = A3) \times P(B = B2)$$

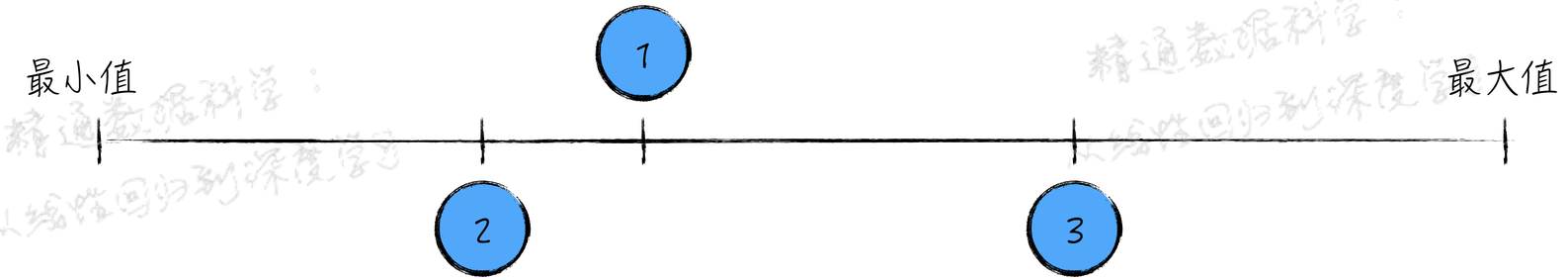
$$= 17 \times \frac{13}{17} \times \frac{12}{17} = 9.17$$

卡方检验划分法

求解算法

划分定量变量的原则是：

划分后的定性变量与被预测量y越相关越好 → 划分后的定性变量与y之间的卡方统计量越大越好 → 使用贪心算法来解决这个问题，直至卡方统计量小于某一个阈值

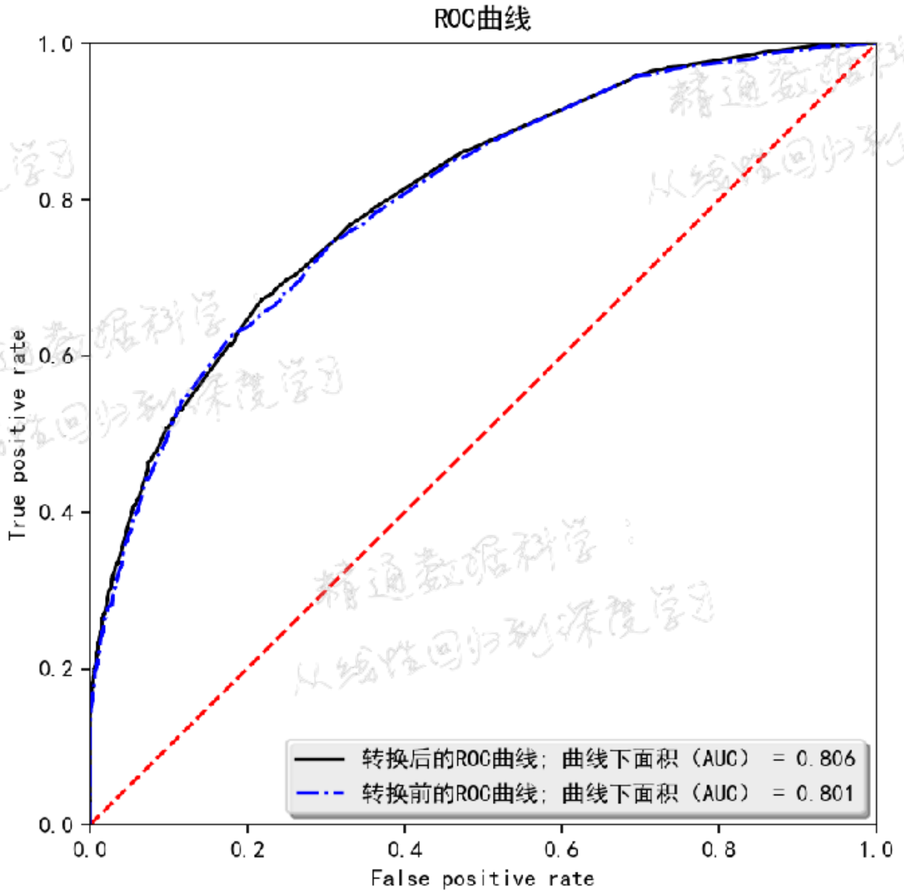


卡方检验划分法

模型结果

基于卡方检验，将每星期工作时间“最优”地划分为5段

	coef
Intercept	-6.1007
C(hours_per_week_group)[T.34-37]	0.7750
C(hours_per_week_group)[T.37-41]	1.2699
C(hours_per_week_group)[T.41-49]	1.7780
C(hours_per_week_group)[T.49-99]	1.9936
education_num	0.3118
capital_gain	0.0003
capital_loss	0.0008



模型效果上升

THANK YOU