

# 目录

精通数据科学

精通数据科学

# ONE 定量变量的线性陷阱 边际效应恒定假设

精通数据科学

TWO 变量离散化

从定量到定性

精通教师和潜

THREE 卡方检验划分法

有效的离散化方法

糖通数概制资

# 定量变量的线性陷阱

回顾模型结果

系数

大于0

Logit Regression Result	Loc	iit	Reg	ression	Resul	ts
-------------------------	-----	-----	-----	---------	-------	----

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

第五章的模型结果

糖通数抵到的秘珠度管的

预测变量:年收入是否大于50k

2.73/12	1 230 C 3	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
<pre>C(workclass, contrast_mat,</pre>	levels=l) Private	0.7371	0.226	3.267	0.001	0.295	1.179
<pre>C(workclass, contrast_mat,</pre>	levels=l) Federal-gov 💢 🦪	1.1899	0.280	4.249	0.000	0.641	1.739
<pre>C(workclass, contrast_mat,</pre>	levels=l) Local-gov	0.9964	0.254	3.916	0.000	0.498	1.495
<pre>C(workclass, contrast_mat,</pre>	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

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

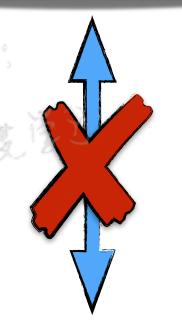
模型系数大于O

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

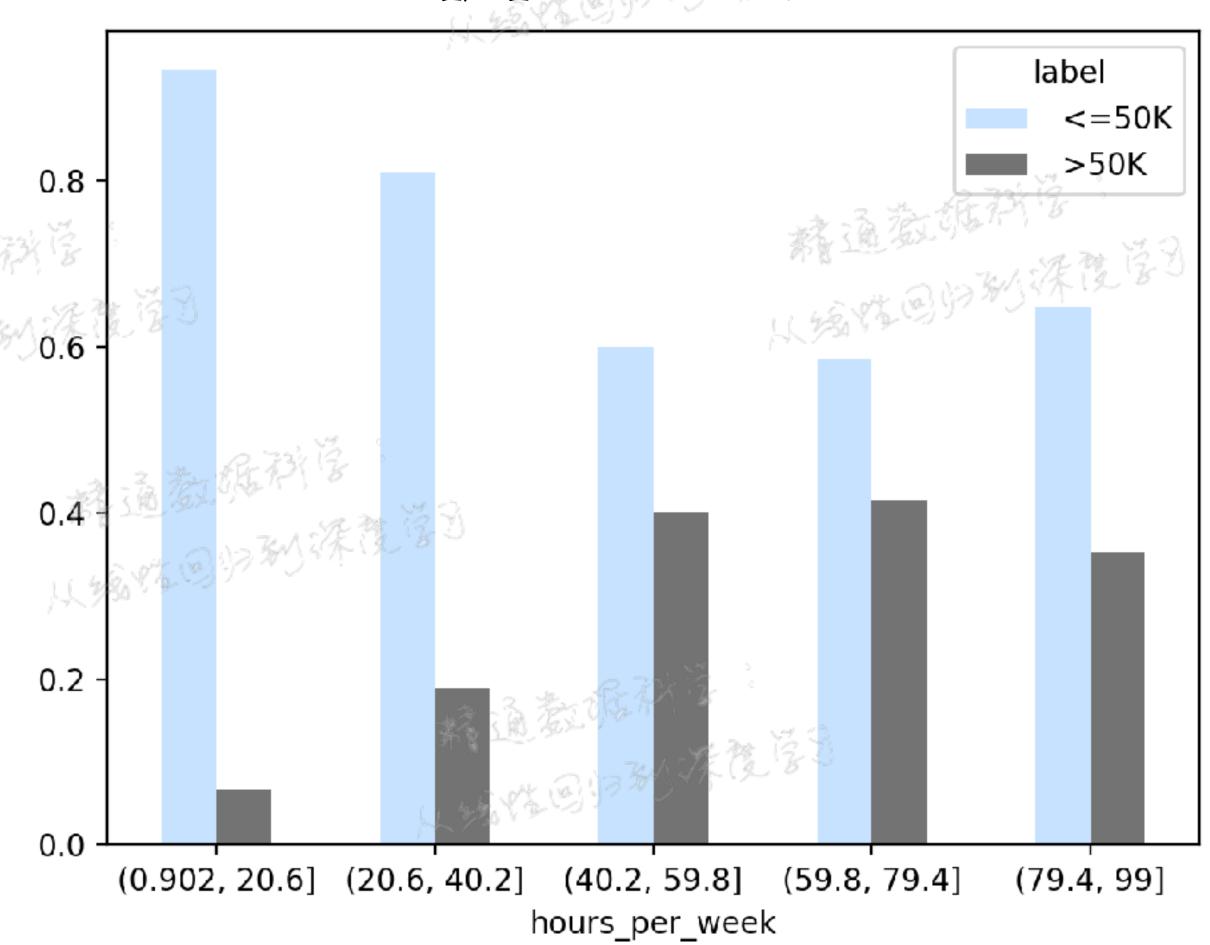
随着每星期工作时间的增加, 年收入大于50k的概率增加



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

hours\_per\_week和label的交叉报表



# 定量变量的线性陷阱

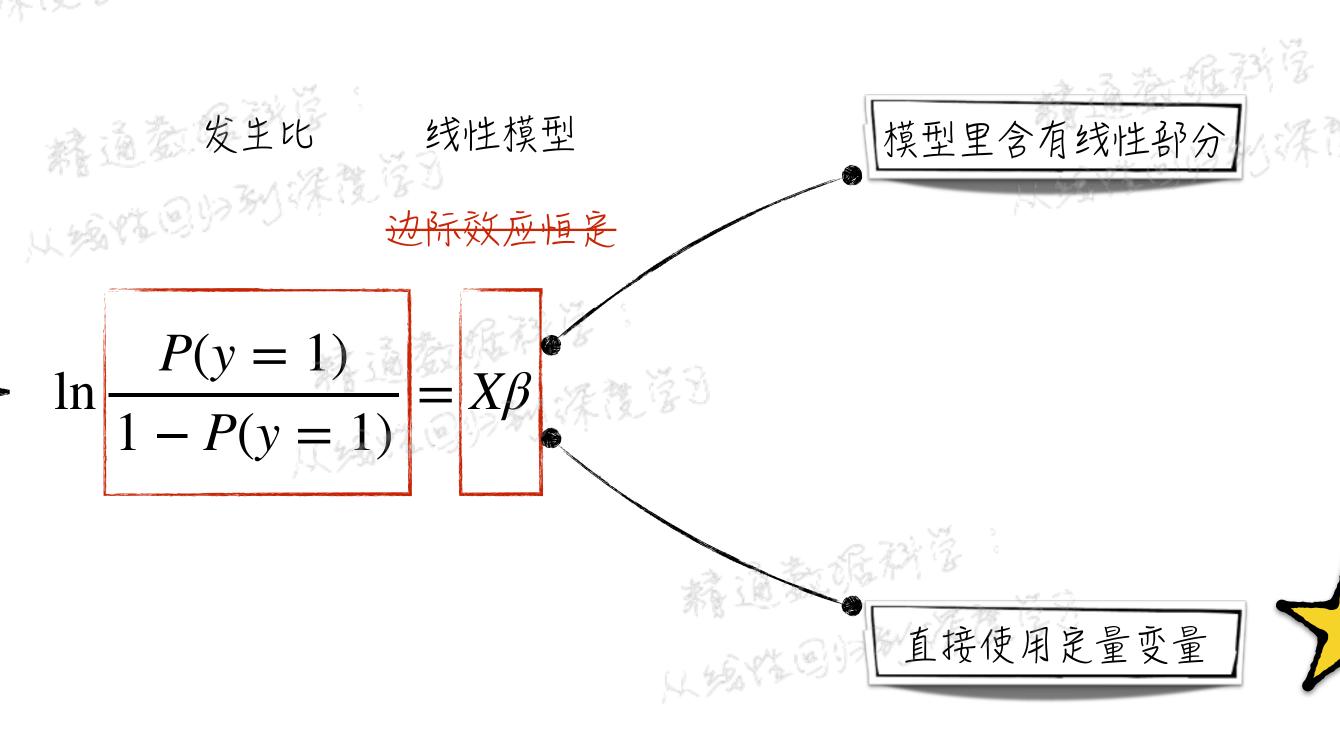
隐含的边际效应恒定假设

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



精通数据和资

解决问题的 突破口

# 目录

精通数据科学 ONE 定量变量的线性陷阱

边际效应恒定假设

精通数据积管。 从给此回的秘证不能管别

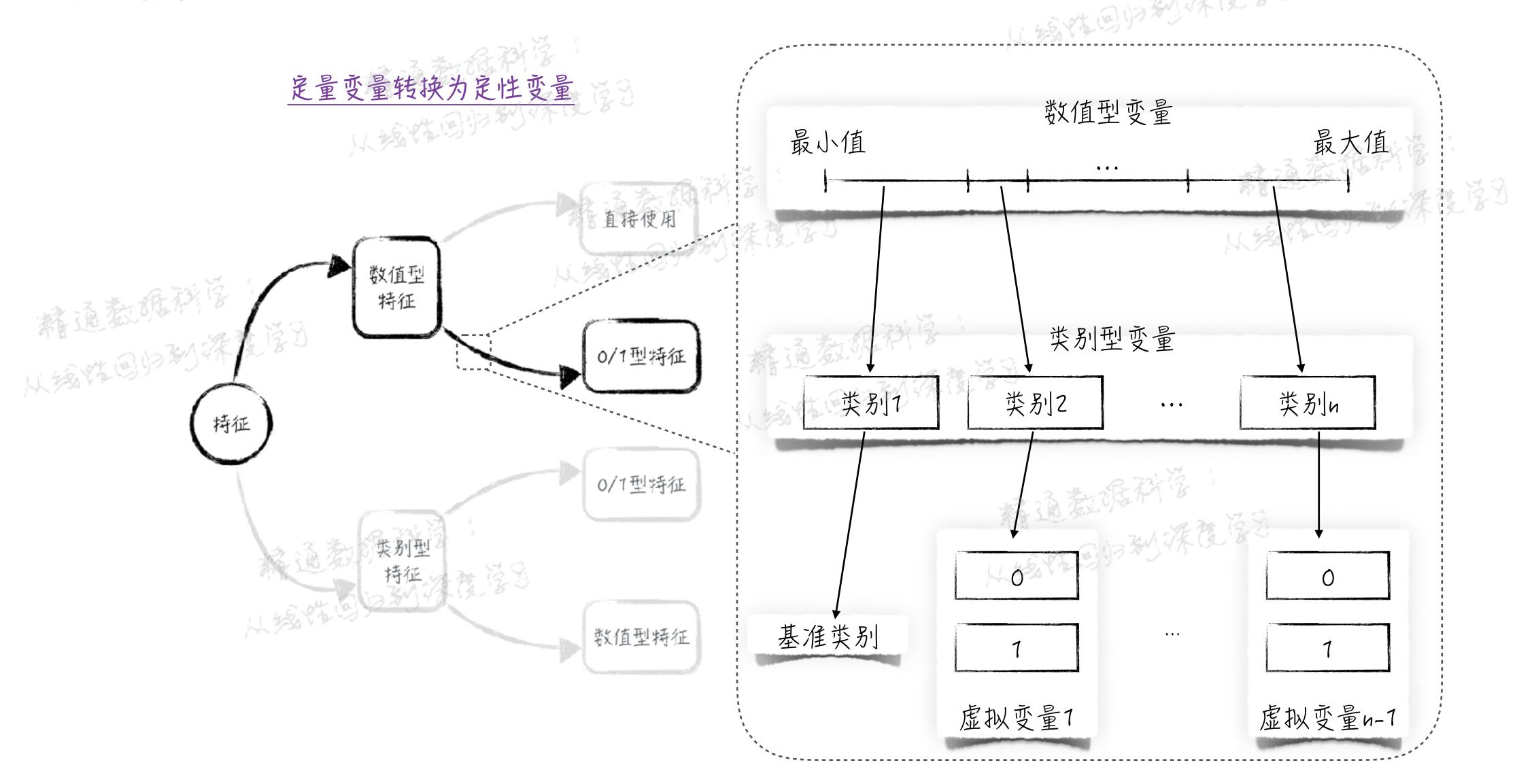
精通数据科学

精通数据科学。从线性回归初深度管

精通数据科学 从编唱的多数深度漫 THREE 卡方检验划务法

# 变量离散化

从定量到定性



# 变量离散化

从定量到定性

# 如何选择划 分区间?

划分太粗:

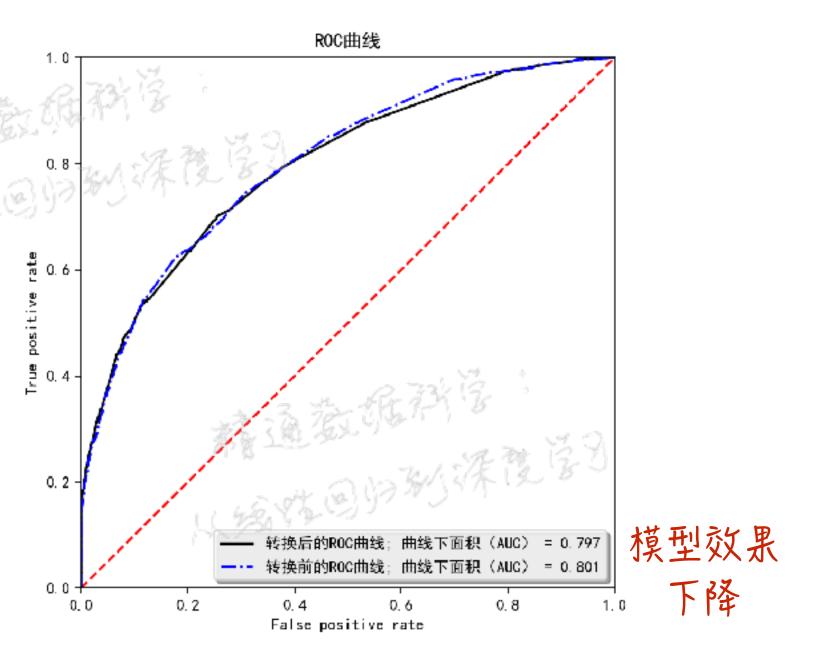
· 模型效果 欠佳 划分太细:

过拟合风 险上升

### 每星期工作时间等分为5份

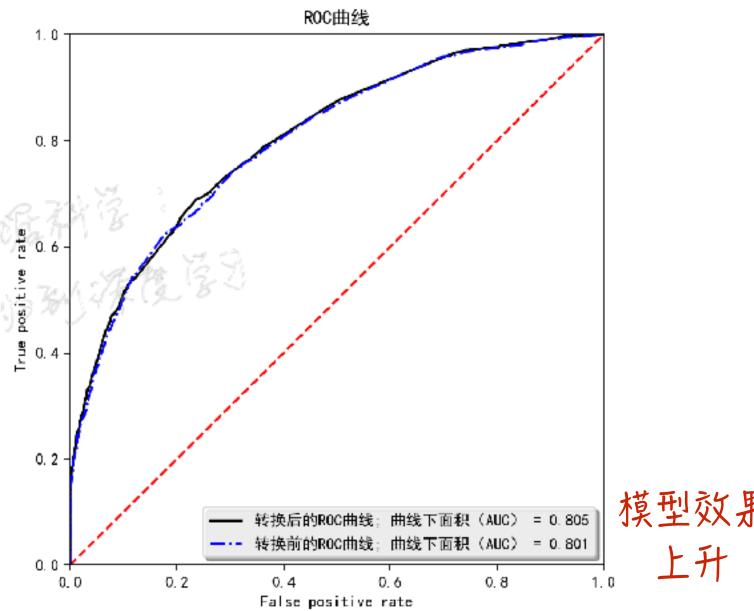
Intercept C(hours_per_week_group)[T.20-40] C(hours_per_week_group)[T.40-60] C(hours_per_week_group)[T.60-80] C(hours_per_week_group)[T.80-100] education_num capital_gain capital_loss  -6.1776 1.1916 1.9972 1.9353 0.3126 0.3126	- 4 - 6-73/3	coef
	C(hours_per_week_group)[T.20-40] C(hours_per_week_group)[T.40-60] C(hours_per_week_group)[T.60-80] C(hours_per_week_group)[T.80-100] education_num capital_gain	1.1916 1.9972 1.9353 1.7762 0.3126 0.0003

变量系 数与数 据相符



# 每星期工作时间等分为10份

1-1-7-30	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



# 目录

精通数据科学 THE 定量变量的线性陷阱

边际效应恒定假设

精通数据科学

精通数据科学

糖通数源和资

THREE 卡方检验划分法

精通数据科学。

从给此回的秘证不能管别

# 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

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

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

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)

<b>万个</b>		变量A的第	<b></b>	高数据科学	0
精通数据的	Д1	A2	АЗ	总计	130
变量 <b>B1</b>	7	7	3	5	
B的 类别 B2	° 1	1	10	12	
· · · · · · · · · · · · · · · · · · ·	22	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

精通数据科学

从编楼回归到深度管型

从始级回的秘珠度管 划分定量变量的原则是:

划分后的定性变量与被预测量少 越相关越好



精通数据科学

划分后的定性变量与y之间的卡 方统计量越大越好

精通数据科学。

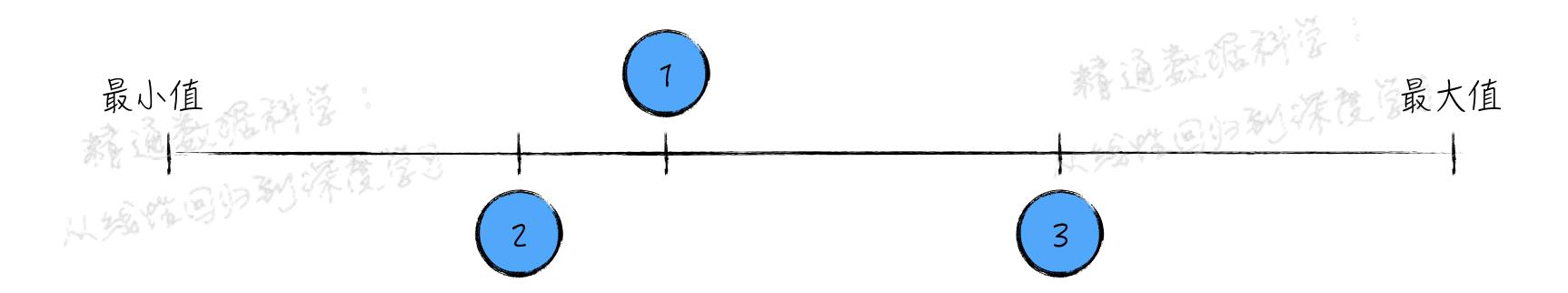
从给你回归和深度管理



精通数据科学。

从绝性回归和汽车度管引

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

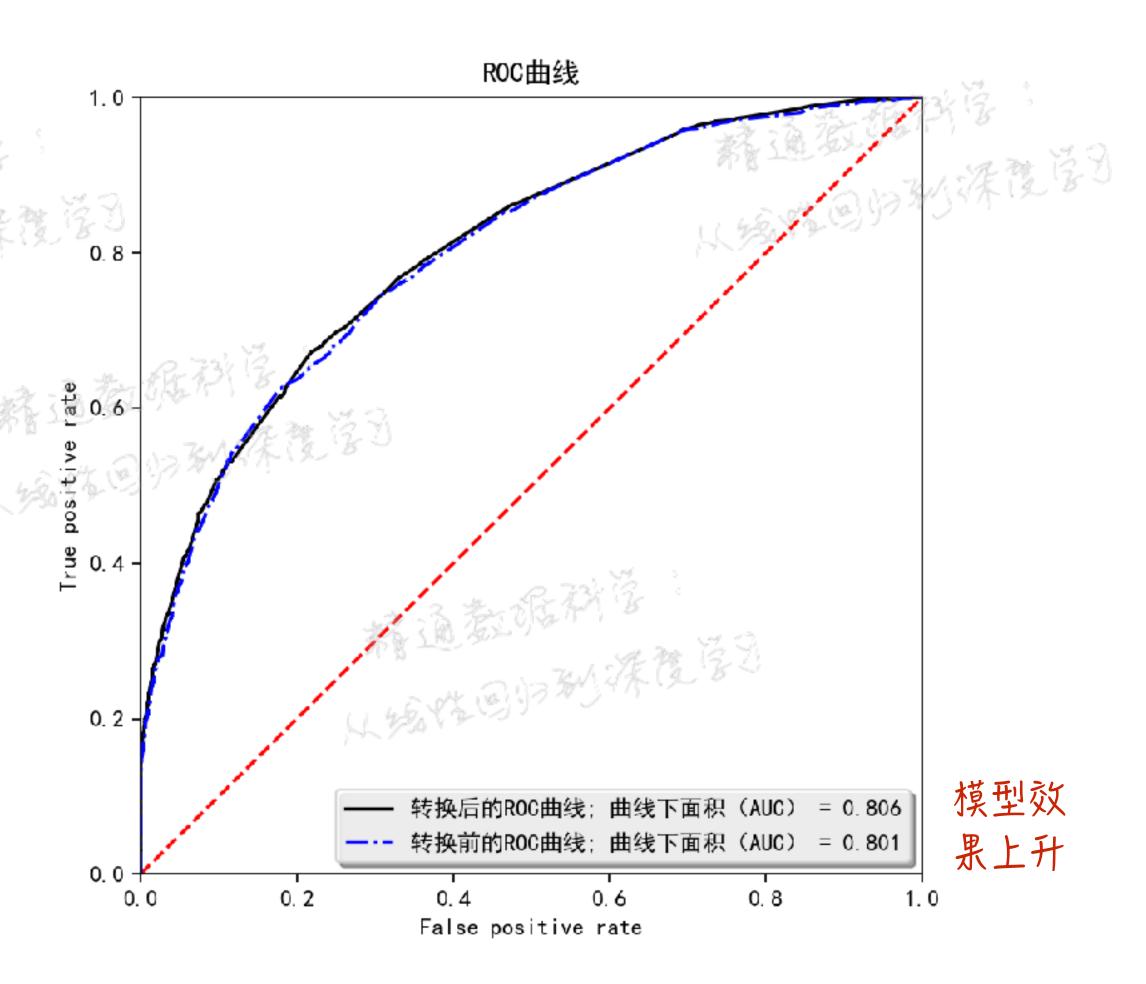


# 卡方检验划分法

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

	1947
京·伊克·科·	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
	C(hours_per_week_group)[T.34-37] C(hours_per_week_group)[T.37-41] C(hours_per_week_group)[T.41-49] C(hours_per_week_group)[T.49-99] education_num capital_gain



精通数据科学。 从验验证到的秘证不改资

# THANKSOUS

務通数据科学 从给您回归和深度管

村通教师和强。

精通数据科学。 从绝路的多处深度管

精通数据科学

精通数派科学 从给你回的秘况