目录

CONTENTS

锻炼是否可以减肥

10个经常锻炼超过3个月的人的减肥数据

样本减重均值 = 2 kg

样本标准差 = 1 kg

H0：锻炼不会影响体重，不相关

H1：锻炼会影响体重

概率值实际上是p值，它就是我们假设零假设成立时观察到的结果或极端结果的概率，统计学家把这个阈值称为显著性水平()，在大多数的情况下，取=0.05

先验知识—假设检验：又称显著性检验

由样本间存在的差别对样本所代表的总体间是否存在着差别做出判断

建立假设
确定检验水准

H_0：无效假设或称原假设、零假设

H_1：一备择假设，H_0的对立面

检验水准：根据不同研究目的事先设置是否拒绝原假设的判断标准，也称显著性水准，它指无效假设H0为真，但被错误地拒绝的一个小概率值，一般取检验水准α =0.05

选定检验方法
计算统计量

T检验、卡方检验、Z检验…

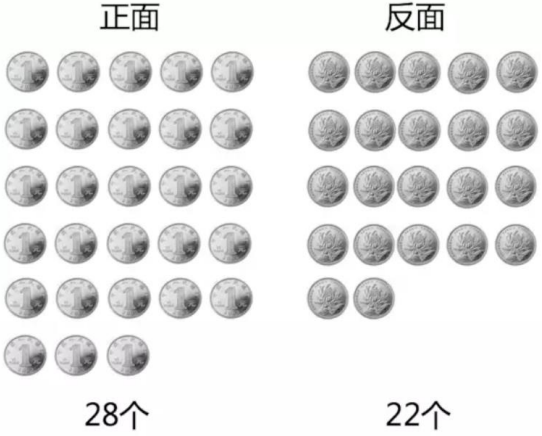
确定概率P值
做出推论

当p≤ α时，拒绝H_0，接受H_1

在H_0成立的条件下，出现等于及大于现有检验统计量值的概率P≤ α ，是小概率事件，这在一次抽样中是不大可能发生的，即现有样本信息不支持H_0因而拒绝它

检验两个变量之间有没有关系

投50次



投50次

计算表格（2行2列）	正面	反面
理论频次（E）	25	25
观察频次（O）	28	22

$$\frac{(O-E)^2}{E}+\frac{(O-E)^2}{E}$$

①求卡方值 $\frac{(28-25)^2}{25}+\frac{(22-25)^2}{25}=0.72$

②求自由度 (行数-1) * (列数-1) = (2-1) * (2-1) = 1

③ 置信度 95%

实际投出来是多少次数　　如果硬币是正常的话，投出来的理论上是多少次数

$$X^2=\sum\frac{(\text{observed}-\text{expected})^2}{\text{expected}}$$

如果硬币是正常的话，投出来的理论上是多少次数

H_0：假设硬币均衡

H_1：假设硬币不均衡　　检验水准：0.05

查表：0.72<3.841 不能推翻假设

结论：有95%的把握说这个硬币是均衡的

先验知识—P值&95%IC

P值即概率，反映某一事件发生的可能性大小　　原假设为真时，样本出现的概率

P值	碰巧的概率	对无效假设	统计意义
P>0.05	碰巧出现的可能性大于5%	不能否定无效假设	两组差别无显著意义
P<0.05	碰巧出现的可能性小于5%	可以否定无效假设	两组差别有显著意义
P <0.01	碰巧出现的可能性小于1%	可以否定无效假设	两者差别有非常显著意义

置信水平为95%的意思是多次抽样中有95%的置信区间包含未知的参数值而另外的5%则不包含真值。

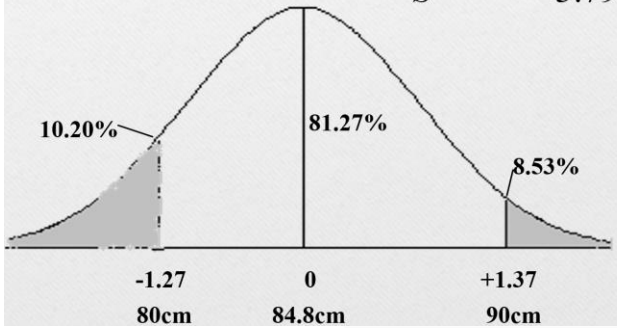
先验知识—正态分布

正态分布是用来观察数据分布的概率密度函数

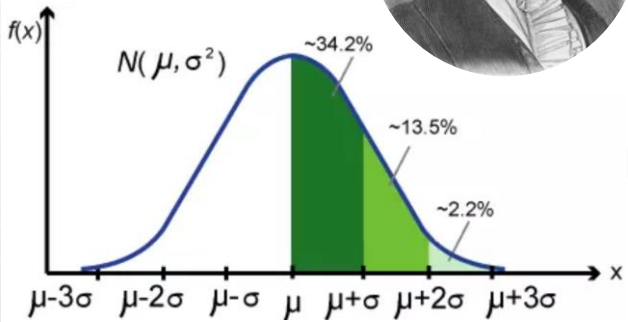
例：某地区2000年100名2岁男童的身高资料，已知均数$\overline{X}=84.8cm$，标准差S=3.79cm。

该地区2岁男童中，身高不足80.0cm者占该地区2岁男童的比例；

$$Z=\frac{X-\overline{X}}{S}=\frac{80.0-84.8}{3.79}=-1.27$$



查表，Φ（-1.27）=0.1020，故理论上该地区2岁男童中，身高不足80.0cm者占该地区2岁男童的10.2%。



$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}$$

先验知识—OR值

odds: 称为几率、比值、比数，是指某事件发生的可能性(概率)与不发生的可能性（概率）之比。

用p表示事件发生的概率，则：odds = p/(1-p)。

OR：比值比，为实验组的事件发生几率(odds1)/对照组的事件发生几率(odds2)。

模型为 $\ln(p/(1-p)) = \beta_0 + \beta_1 * sex$

荣誉班	Male	Female	Total
0	74	77	151
1	17	32	49
Total	91	109	

荣誉班	系数β	标准误	P
sex	0.593	0.3414294	0.083
截距	-1.47	0.2689555	0.000

男性：荣誉班级的概率=17/91，非荣誉班级的概率=74/91，

荣誉班级的几率=oddsM=(17/91)/(74/91) = 17/74 = 0.23；

女性：荣誉班级的几率oddsF = (32/109)/(77/109)=32/77 = 0.42

女性对男性的几率之比OR = oddsF/oddsM = 0.42/0.23 = 1.809

女性比男性在荣誉班的几率高80.9%

回到Logistic回归结果。截距的系数（ β0 ） -1.47是男性odds的对数ln(0.23) = -1.47。

变量sex的系数为0.593，是女性对男性的OR值的对数，ln(1.809) = 0.593

案例

数据背景：

西安交大一附院，儿科门诊，2019年1月1日~2019年4月2日，做甲流化验的数据

患者编号 医生编号 性别 出生日期 检查日期 检查结果

目的：

1、判断得甲流是否和年龄相关

2、判断哪个医生开的甲流化验单阳性率高

3、判断得甲流是否和性别相关

案例—数据

患者编号	医生编号	性别	出生日期	检查日期	年龄	分组	医嘱名称	甲流T	甲流V
pat_000797e1f906829e33e54df6317bb3fc	001234	女性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-01-11 04:31:48	阴性
pat_0010be4fdb3fa834d5b6f065af9e1c84	001235	女性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-01-19 17:57:50	阴性
pat_0057d84c919dac39025e6c5c3952fce1	001236	女性	2018	2019	1	1	甲型流感病毒抗原检测	2019-01-16 02:33:48	阳性
pat_0084fced365530c84a1300cad3f471b7	001237	男性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-03-20 13:25:54	阴性
pat_00d1403bffee5a8a8ac8296e73e52db5	001238	男性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-02-07 21:08:55	阴性
pat_00f2c15b978650f20997ad74a8302952	001239	女性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-01-31 16:21:18	阴性
pat_0118de849c1fd5e917cd722f7a4f6100	001240	男性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-01-18 19:24:14	阴性
pat_015474f6f95d3433eabda846674f9455	001241	男性	2018	2019	1	1	甲型流感病毒抗原检测	2019-01-15 12:32:18	阴性
pat_0199637d024d38c532e2e843b61a6ce6	001242	女性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-01-04 11:23:09	阴性
pat_01a2f381c05e84b6bbdb985e8d1b7961	001243	男性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-01-14 05:21:13	阴性
pat_01c9eef16f192893d8da95a9dcad687a	001244	男性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-01-20 02:59:49	阳性
pat_02b416391c57c3d191114afe917caed6	001245	男性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-03-08 19:51:59	阴性
pat_02dd407750c8886b81e5f1bb2a6e68eb	001246	男性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-01-30 11:15:17	阳性
pat_02ee63c518b1ce6b3c9887da7337cd40	001247	女性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-02-10 22:57:53	阴性
pat_034269e56a43326663fc38e8966fc458	001248	女性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-01-22 16:03:21	阴性
pat_03438c8fada6f0108675c3342cad64c3	001249	女性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-01-19 22:11:09	阴性
pat_03b59ee663912cc0360bdac7f14fd8f6	001250	女性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-03-03 07:47:45	阴性
pat_03ffce70f89b9bee1d0e32b7d03d679a	001251	男性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-03-31 01:22:13	阴性
pat_04240c5f7b9699689d9d3f0a6c0f3fd8	001252	男性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-03-08 12:20:16	阴性
pat_04770c0ae85b5b314c1be5f5f0de6645	001253	女性	2017	2018	1	1	甲乙型流感病毒抗原联合检测[复]	2018-01-21 19:20:39	阴性
pat_047cac25a69ddd3b37592bbbaded32f9	001254	男性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-01-30 11:14:53	阴性
pat_04972071c99f756be5037edf86927c6f	001255	男性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-03-24 11:49:23	阴性
pat_04ad1070d6aa07fe7f16c5ea315646a7	001256	女性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-03-09 15:55:04	阴性
pat_04ce69026ee4b5c3ba235bcdbaa83743	001257	男性	2018	2019	1	1	甲乙型流感病毒抗原联合检测[复]	2019-02-11 18:37:32	阴性

案例—判断得甲流是否和年龄相关

Age* Positive 交叉制表

			Positive		合计
			.00	1.00	
Age	1.00	计数	630	196	826
		Age 中的 %	76.3%	23.7%	100.0%
	2.00	计数	3077	1607	4684
		Age 中的 %	65.7%	34.3%	100.0%
	3.00	计数	2482	759	3241
		Age 中的 %	76.6%	23.4%	100.0%
	4.00	计数	546	215	761
		Age 中的 %	71.7%	28.3%	100.0%
合计		计数	6735	2777	9512
		Age 中的 %	70.8%	29.2%	100.0%

1：0~1
2：1~5
3：5~10
4：10~15

相关系数

			Positive	年龄
Kendall 的 tau_b	Positive	相关系数	1.000	-.050**
		Sig.（双侧）	.	.000
		N	9512	9512
	年龄	相关系数	-.050**	1.000
		Sig.（双侧）	.000	.
		N	9512	9512
Spearman 的 rho	Positive	相关系数	1.000	-.059**
		Sig.（双侧）	.	.000
		N	9512	9512
	年龄	相关系数	-.059**	1.000
		Sig.（双侧）	.000	.
		N	9512	9512

**. 在置信度（双测）为 0.01 时，相关性是显著的。

卡方检验

	值	df	渐进 Sig. (双侧)
Pearson 卡方	123.824[a]	3	.000
似然比	124.818	3	.000
线性和线性组合	21.584	1	.000
有效案例中的 N	9512		

a. 0 单元格(0.0%) 的期望计数少于 5。最小期望计数为 222.17。

H_0：假设是不同年龄段的甲流阳性率是一样的

H_1：假设是不同年龄段的甲流阳性率是不一样的

结论：p值远小于0.01，拒绝假设H_0，接收H_1

案例—判断得甲流是否和年龄相关

方程中的变量

		B	S.E,	Wals	df	Sig.	Exp (B)	EXP(B) 的 95% C.I.	
								下限	上限
步骤 1[a]	Age			122.718	3	.000			
	Age(1)	.518	.087	35.139	1	.000	1.679	1.414	1.992
	Age(2)	-.017	.092	.035	1	.851	.983	.821	1.176
	Age(3)	.236	.115	4.215	1	.040	1.266	1.011	1.585
	常量	-1.168	.082	203.802	1	.000	.311		

a. 在步骤 1 中输入的变量: Age.

logistic

1：0~1
2：1~5
3：5~10
4：10~15

风险估计

	值	95% 置信区间	
		下限	上限
Age (1.00 / 2.00) 的几率比	1.679	1.414	1.992
用于 cohort Positive = .00	1.161	1.112	1.212
用于 cohort Positive = 1.00	.692	.608	.786
有效案例中的 N	5510		

卡方

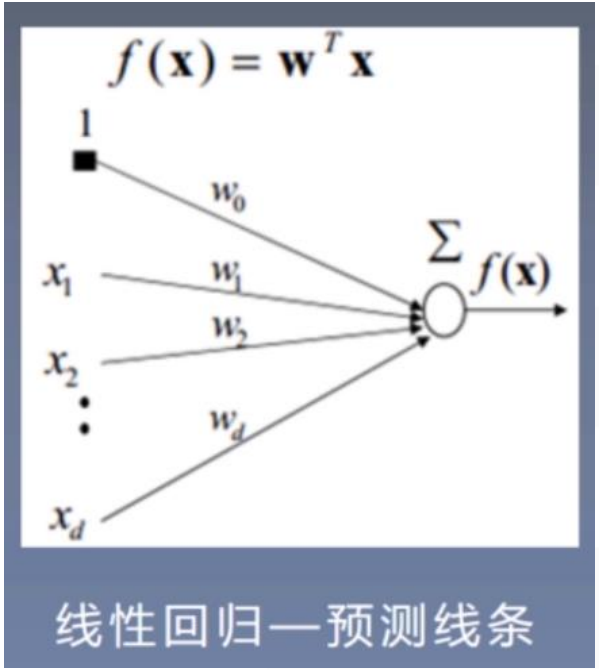
案例—判断哪个医生开的甲流化验单阳性率高

医生编号	阳性	阴性	合计	阳性率
001234	3	3	6	50.0%
001235	51	59	110	46.4%
001236	323	409	732	44.1%
001237	79	104	183	43.2%
001238	345	468	813	42.4%
001239	31	47	78	39.7%
001240	40	66	106	37.7%
001241	32	57	89	36.0%
001242	35	65	100	35.0%
001243	25	49	74	33.8%
001244	10	21	31	32.3%
001245	40	85	125	32.0%
001246	4	9	13	30.8%
001247	419	948	1367	30.7%
001248	354	835	1189	29.8%
001249	391	1014	1405	27.8%
001250	261	742	1003	26.0%
001251	17	52	69	24.6%
001252	11	35	46	23.9%

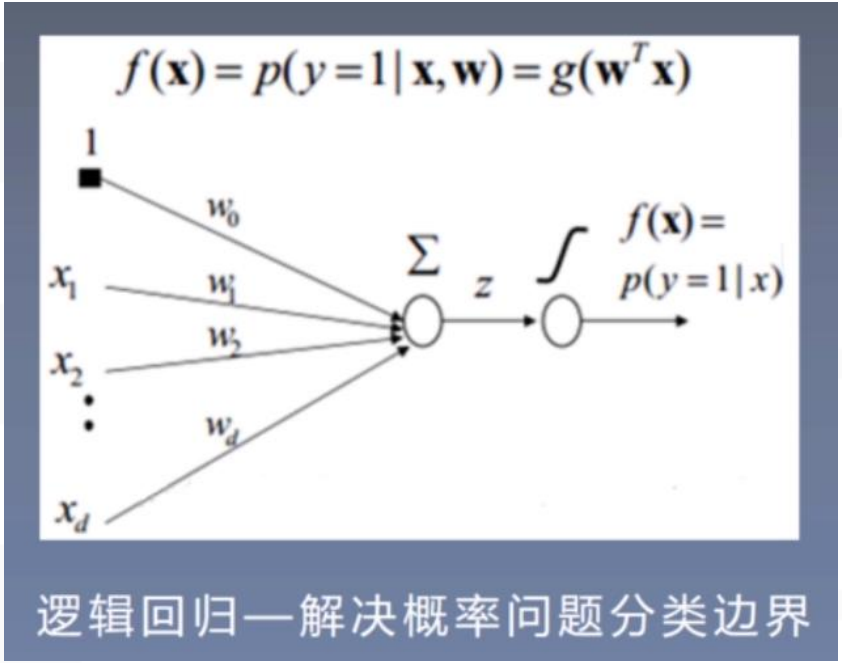
统计结果

Microsoft Excel
工作表

逻辑回归—模型



线性回归—预测线条



逻辑回归—解决概率问题分类边界

Logistic回归虽然名字叫”回归”，但却是一种分类学习方法。

使用场景大概有两个：第一用来预测，第二寻找因变量的影响因素。

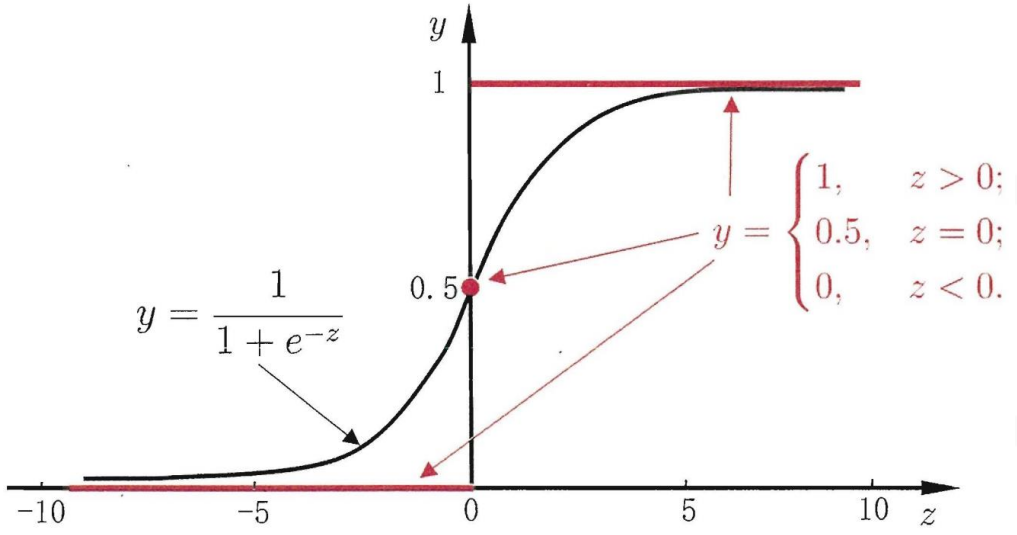
逻辑回归—模型

线性回归模型产生的预测值是实值，二分类任务需要讲实值Z转换为{0,1}

$$z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

$$y = \begin{cases} 0, & z < 0\,; \\ 0.5, & z = 0\,; \\ 1, & z > 0\,, \end{cases}$$

单位阶跃函数



对数几率函数

逻辑回归—模型

边界

$$z = w_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

- w_i为回归系数

逻辑回归方程 $h(x) = g(\mathbf{w}^{\mathrm{T}}\mathbf{x}) = \dfrac{1}{1 + e^{-\mathbf{w}^{\mathrm{T}}\mathbf{x}}}$

$h(x)$函数值的含义：分类为1的概率。

- 对于输入x分类结果为类别1和类别0的概率分别为：

$$p(y=1) = h(x)$$
$$p(y=0) = 1 - h(x)$$

逻辑回归—建模1

$$z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

$$y = \frac{1}{1 + e^{-z}} .$$

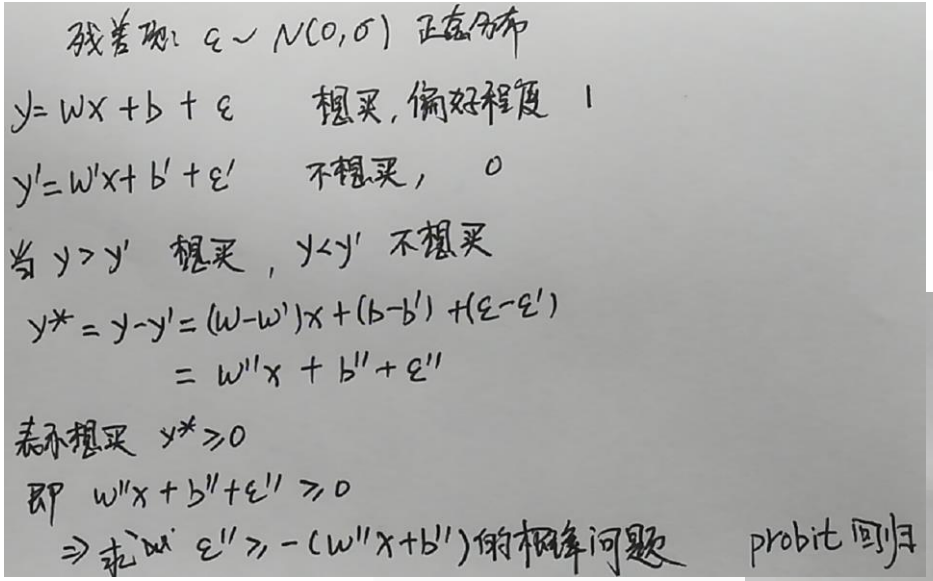
$$\Rightarrow \quad y = \frac{1}{1 + e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}} . \quad \Rightarrow \quad \ln \frac{y}{1-y} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

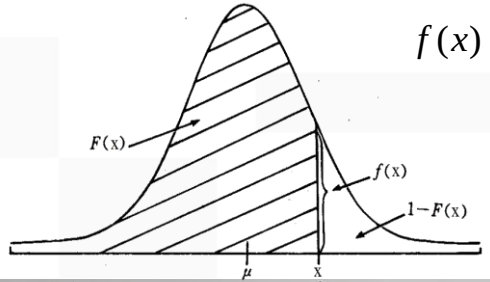
$$\ln \frac{p(y=1 \mid \boldsymbol{x})}{p(y=0 \mid \boldsymbol{x})} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \quad \Rightarrow$$

$$p(y=1 \mid \boldsymbol{x}) = \frac{e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}}{1 + e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}} ,$$

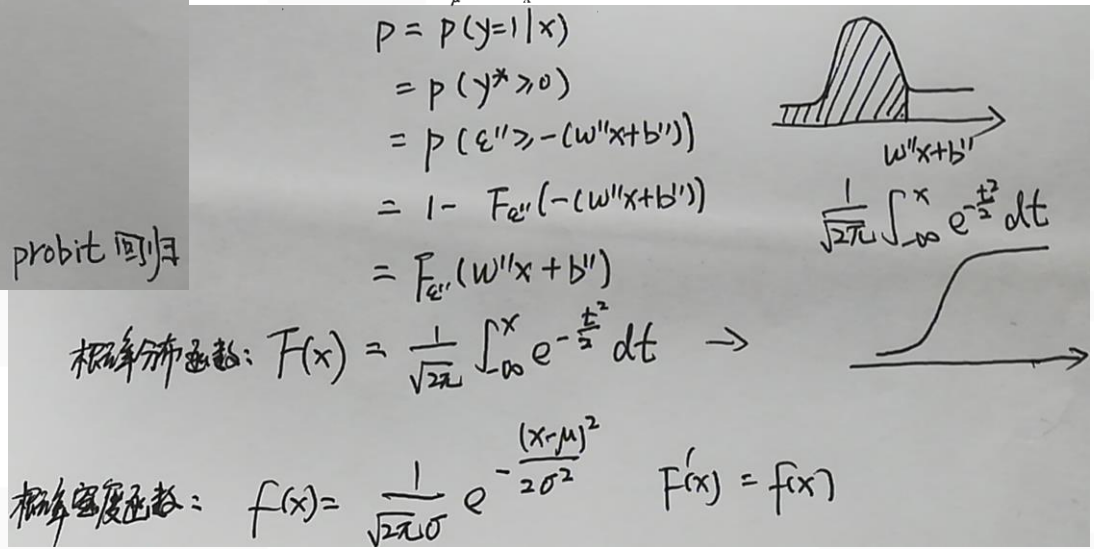
$$p(y=0 \mid \boldsymbol{x}) = \frac{1}{1 + e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}} .$$

逻辑回归—建模2





$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-(x-\mu)^2/(2\sigma^2)}$$

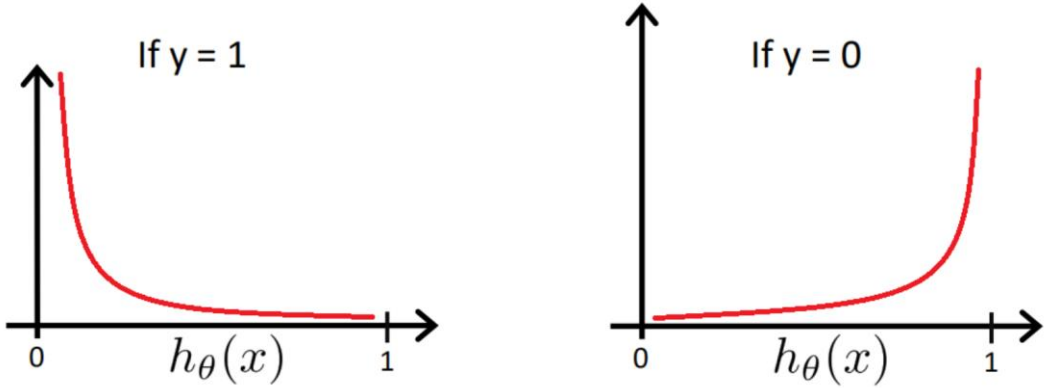


正态分布的累计分布函数

近似等于sigmoid函数

逻辑回归—损失函数

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



当y = 1，$h_\theta(x) = 1$时，cost=0　当y = 0，$h_\theta(x) = 1$时，cost=∞

当y = 1，$h_\theta(x) = 0$时，cost=∞　当y = 0，$h_\theta(x) = 0$时，cost=0

$$Cost(h_\theta(x), y) = -y log(h_\theta(x)) - (1 - y) log(1 - h_\theta(x))$$

逻辑回归—损失函数推导1

$$P(y=1 \mid x;\theta) = h_\theta(x)$$
$$P(y=0 \mid x;\theta) = 1 - h_\theta(x)$$

$$p(y \mid x;\theta) = (h_\theta(x))^y (1-h_\theta(x))^{1-y}$$

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

$$\begin{aligned} L(\theta) &= p(\vec{y} \mid X;\theta) \\ &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta) \\ &= \prod_{i=1}^{m} (h_\theta(x^{(i)}))^{y^{(i)}} (1-h_\theta(x^{(i)}))^{1-y^{(i)}} \end{aligned}$$

似然函数

$$l(\theta) = log(L(\theta)) = \sum_{i-1}^{m} log((h_\theta(x^{(i)}))^{y^{(i)}}) + log((1-h_\theta(x^{(i)}))^{1-y^{(i)}})$$

对数似然函数

$$l(\theta) = log(L(\theta)) = \sum_{i=1}^{m} y^{(i)} log(h_\theta(x^{(i)})) + (1-y^{(i)}) log(1-h_\theta(x^{(i)}))$$

逻辑回归—损失函数推导1

样本x得到输出y的生成概率

$$p(y \mid x;\theta) = (h_\theta(x))^y \left(1 - h_\theta(x)\right)^{1-y}$$

假定样本与样本之间相互独立，那么整个样本集生成的概率即似然函数为：

$$\begin{aligned} L(\theta) &= p(\vec{y} \mid X;\theta) \\ &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta) \\ &= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}} \end{aligned}$$

其中，m为样本的总数，$y^{(i)}$表示第i个样本的类别，$x^{(i)}$表示第i个样本

逻辑回归—梯度下降

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

Want $\min_\theta J(\theta)$:

Repeat $\{$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

$\}$

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} \left(y^{(i)} \log\left(h_\theta(x^{(i)})\right) + (1 - y^{(i)}) \log\left(1 - h_\theta(x^{(i)})\right)\right)$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m} (y^{(i)} \cdot \frac{1}{h_\theta(x^{(i)})} - (1 - y^{(i)}) \cdot \frac{1}{1 - h_\theta(x^{(i)})}) \cdot \frac{\partial h_\theta(x^{(i)})}{\partial \theta_j}$$

$$= -\frac{1}{m}\sum_{i=1}^{m} (\frac{y^{(i)}}{h_\theta(x^{(i)})} - \frac{1 - y^{(i)}}{1 - h_\theta(x^{(i)})}) \cdot h_\theta(x^{(i)})(1 - h_\theta(x^{(i)})) \frac{\partial \theta^T x}{\partial \theta_j}$$

$$= -\frac{1}{m}\sum_{i=1}^{m} (y^{(i)} \cdot \left(1 - h_\theta(x^{(i)})\right) - (1 - y^{(i)}) \cdot h_\theta(x^{(i)})) \cdot x_j$$

$$= -\frac{1}{m}\sum_{i=1}^{m} (y^{(i)} - y^{(i)} \cdot h_\theta(x^{(i)}) - h_\theta(x^{(i)}) + y^{(i)} \cdot h_\theta(x^{(i)})) \cdot x_j$$

$$= -\frac{1}{m}\sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) \cdot x_j$$

$$= \frac{1}{m}\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j$$

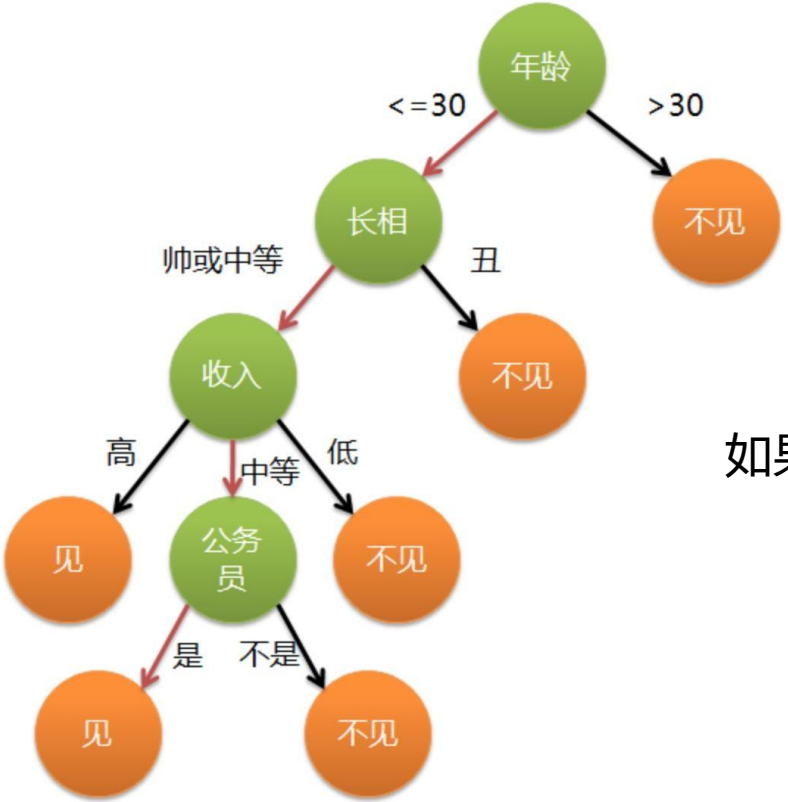
$$g(z) = \frac{1}{1 + e^{-z}}$$

$$g'(x) = \left(\frac{1}{1 + e^{-x}}\right) = \frac{e^{-x}}{(1 + e^{-x})^2}$$

$$= \frac{1}{1 + e^{-x}} \cdot \frac{e^{-x}}{1 + e^{-x}} = \frac{1}{1 + e^{-x}} \cdot \left(1 - \frac{1}{1 + e^{-x}}\right)$$

$$= g(x) \cdot (1 - g(x))$$

决策树



比较适合分析离散数据

如果是连续数据要先转成离散数据再做分析

决策树—信息量

一个具体事件的信息量应该是随着其发生概率而递减的，且不能为负。

信息奠基人香农（Shannon）认为“信息是用来消除随机不确定性的东西”

衡量信息量大小就看这个信息消除不确定性的程度。

信息量的大小和事件发生的概率成反比

$$h(x) = -log_2 p(x)$$

- 信息量和事件发生的概率有关，当事件发生的概率越低，传递的信息量越大；
- 信息量应当是非负的，必然发生的信息量为0；
- 两个事件的信息量可以相加，并且两个独立事件的联合信息量应该是他们各自信息量的和；

D代表当前样本集合，p_k代表当前样本中第k类样本所占的比例，|y|代表类别总数

$$\mathrm{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$$

决策树—信息熵

事件A：小明考试及格，对应的概率$P(x_A)=0.1$，信息量为$I(x_A)=-\log(0.1)=3.3219$

事件B：小王考试及格，对应的概率$P(x_B)=0.999$，信息量为$I(x_B)=-log(0.999)=0.0014$

$$H_A(x)=-[p(x_A)\log(p(x_A))+(1-p(x_A))\log(1-p(x_A))]=0.4690$$

$$H_B(x)=-[p(x_B)\log(p(x_B))+(1-p(x_B))\log(1-p(x_B))]=0.0114$$

$$\mathrm{Ent}(D)=-\sum_{k=1}^{|\mathcal{Y}|}p_k\log_2 p_k$$

熵其实是信息量的期望值，它是一个随机变量的确定性的度量

熵越大，变量的取值越不确定，反之就越确定

决策树—信息熵

信息熵Ent(D) 是用来度量样本集合纯度的最常用指标，信息熵越大，不确定性越大

假如有一个普通骰子A，仍出1-6的概率都是1/6

有一个骰子B，扔出6的概率是50%，扔出1-5的概率都是10%

有一个骰子C，扔出6的概率是100%。

$$\mathrm{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$$

骰子A：$-\left(\frac{1}{6} \times \log_2 \frac{1}{6}\right) \times 6 \approx 2.585$

骰子B：$-\left(\frac{1}{10} \times \log_2 \frac{1}{10}\right) \times 5 - \frac{1}{2} \times \log_2 \frac{1}{2} \approx 2.161$

骰子C：$-(1 \times \log_2 1) = 0$

决策树—ID3算法：信息增益

一句话：前后熵的值变化，谁下降的快就把谁分叉。

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

$$\mathrm{Gain}(D,a)=\mathrm{Ent}(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v)\ .$$

$$\mathrm{Ent}(D)=-\sum_{k=1}^{|\mathcal{Y}|}p_k\log_2 p_k\ .$$

当前信息熵：

$$\mathrm{Ent}(D)=-\sum_{k=1}^{2}p_k\log_2 p_k=-\left(\frac{8}{17}\log_2\frac{8}{17}+\frac{9}{17}\log_2\frac{9}{17}\right)=0.998\ .$$

决策树—ID3算法

按照好坏分类：
好瓜：8/17
坏瓜：9/17

$$\mathrm{Ent}(D) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{8}{17}\log_2\frac{8}{17} + \frac{9}{17}\log_2\frac{9}{17}\right) = 0.998 .$$ 划分前

按照色泽分类好坏瓜：
青绿：{1,4,6,10,13,17}——3/6
乌黑：{2,3,7,8,9,15}——4/6
浅白：{5,11,12,14,16}——1/5

$$\mathrm{Ent}(D^1) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = 1.000 ,$$

$$\mathrm{Ent}(D^2) = -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 0.918 ,$$

$$\mathrm{Ent}(D^3) = -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right) = 0.722 ,$$

用色泽分类好坏后的熵的期望值
H(D|色泽)

$$\mathrm{Gain}(D, \text{色泽}) = \mathrm{Ent}(D) - \sum_{v=1}^{3} \frac{|D^v|}{|D|}\mathrm{Ent}(D^v)$$ 划分后

$$= 0.998 - \left(\frac{6}{17}\times 1.000 + \frac{6}{17}\times 0.918 + \frac{5}{17}\times 0.722\right)$$

$$= 0.109 .$$ 信息增益

分类的期望熵

- 设系统*S*的熵为Entropy(*S*)，*S*具有属性A，*Values*(A)表示属性A所以可能的取值，S_v是*S*中属性A的值为*v*的集合，则

用A分类S后熵的期望值=$\sum_{v\in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$

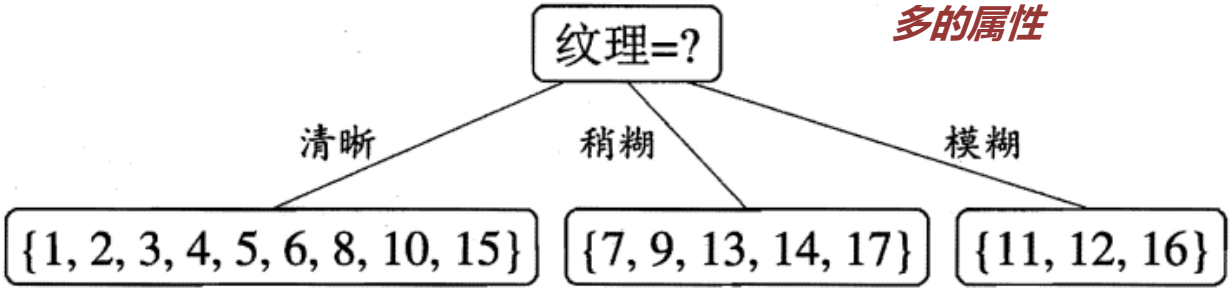
决策树—ID3算法

类似的，我们可计算出其他属性的信息增益：

$$\text{Gain}(D, \text{根蒂}) = 0.143; \quad \text{Gain}(D, \text{敲声}) = 0.141;$$

$$\text{Gain}(D, \text{纹理}) = 0.381; \quad \text{Gain}(D, \text{脐部}) = 0.289;$$

$$\text{Gain}(D, \text{触感}) = 0.006.$$



D3算法根据“最大信息熵增益”原则选择划分当前数据集的最好特征，信息熵是信息的度量方式，不确定度越大或者说越混乱，熵就越大。在建立决策树的过程中，根据特征属性划分数据，使得原本“混乱”的数据的熵(混乱度)减少，按照不同特征划分数据熵减少的程度会不一样。

缺点：只能处理离散型属性，并且对倾向于***选择取值较多的属性***

信息增益反映的给定一个条件以后不确定性减少的程度,必然是分得越细的数据集确定性更高,也就是条件熵越小,信息增益越大

决策树—C4.5算法：信息增益率

除以一个惩罚项，以解决偏向取值较多的属性的问题

$$\text{Gain\_ratio}(D,a)=\frac{\text{Gain}(D,a)}{\text{IV}(a)},$$

$$\text{IV}(a)=-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\log_2\frac{|D^v|}{|D|}$$

信息熵

以信息增益率为准则来选择划分属性的决策树

虽然解决偏向取值较多的属性的问题，单会造成倾向于选择某属性值的分类少的问题

决策树—CART算法

以基尼指数为准则来选择划分属性的决策树

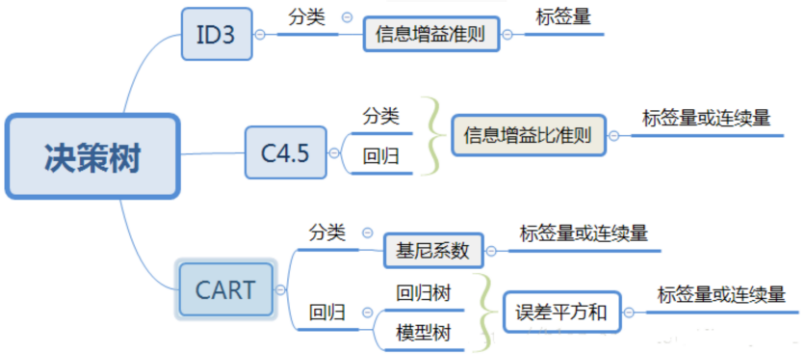
基尼值：$\mathrm{Gini}(D)=\sum_{k=1}^{|\mathcal{Y}|}\sum_{k'\neq k}p_k p_{k'}=\sum_{k=1}^{|\mathcal{Y}|}p_k\sum_{k'\neq k}p_{k'}=\sum_{k=1}^{|\mathcal{Y}|}p_k(1-p_k)=1-\sum_{k=1}^{|\mathcal{Y}|}p_k^2$

基尼指数：$\mathrm{Gini\ index}(D,a)=\sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Gini}\,(D^v)$

基尼值和基尼指数越小，样本集合纯度越高。

基尼值：从数据集D中随机抽取两个样本，其中类别标记不一致的概率，Gini越小，则数据集D的纯度越高

决策树—基尼指数和熵



$$\ln(1+x) = x - \frac{x^2}{2} + \frac{1}{3}x^3 + \cdots \quad |x| \le 1 \quad \text{泰勒展开}$$

$$p_i \log p_i \approx p_i \ln p_i$$

$$= p_i \ln(1 + p_i - 1)$$

$$= p_i(p_i - 1) + \cdots$$

$$\approx p_i^2 - p_i$$

$$-\sum p_i \log p_i \approx \sum p_i - \sum p_i^2 = 1 - \sum p_i^2 = \text{Gini}$$

逻辑回归损失&交叉熵

y为1的概率 $P(y=1 \mid x;\theta) = h_\theta(x)$

1-y为0的概率 $P(y=0 \mid x;\theta) = 1 - h_\theta(x)$

$$p(y \mid x;\theta) = (h_\theta(x))^y \, (1 - h_\theta(x))^{1-y}$$

$$\begin{aligned} L(\theta) &= p(\vec{y} \mid X;\theta) \\ &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta) \\ &= \prod_{i=1}^{m} (h_\theta(x^{(i)}))^{y^{(i)}} \, (1 - h_\theta(x^{(i)}))^{1-y^{(i)}} \end{aligned}$$

$$\mathrm{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k \, .$$

$$l(\theta) = log(L(\theta)) = \sum_{i=1}^{m} y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))$$

该损失函数的推导，没有用到逻辑回归这个条件，所以该损失函数可以应用于其他模型来表示分类问题，比如深度学习模型、SVM等