

## Chapter 1.

分析工具：统计软件：SPSS最早、SAS最著名、R&R Studio、Python

## Chapter 2.

2.1 X 2.2 概率论问题本质：把局部的随机性转为整体上的确定性

2.3 黑天鹅：风险类型未知（不确定性）灰犀牛：已知（随机性）

· 随机试验：可重复、结果可知但不确定

· 基本结果：样本点  $\xrightarrow{\text{集合}} \text{样本空间 } (\Omega)$ , 随机事件是儿子集

p.s. AB同时发生，只有一个发生： $B=A^c \mid \bar{A}$

De Morgan's Laws:  $\overline{\bigcup_{i=1}^{\infty} A_i} = \bigcap_{i=1}^{\infty} \overline{A_i}$ ,  $\overline{\bigcap_{i=1}^{\infty} A_i} = \bigcup_{i=1}^{\infty} \overline{A_i}$

概率的公理化定义（简略版）

概率测度(probability measure), 简称概率, 是定义在样本空间 $\Omega$ 的子集上的实函数, 满足以下公理:

- 规范性:  $P(\Omega) = 1$ ;
- 非负性: 如果  $A \subset \Omega$ , 那么  $P(A) \geq 0$ ;
- 可加性: 对两两互不相容的事件序列  $A_1, A_2, \dots$ , 有

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

贝叶斯定理 (Bayes' Theorem)

设  $A$  和  $B$  为随机事件, 且  $P(B) > 0$ , 则:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)},$$

- 其中  $P(A|B)$  为已知  $B$  发生时  $A$  发生的条件概率, 也称作  $A$  的后验概率(posterior probability)。
- $P(A)$  是  $A$  的先验概率(或边缘概率, prior/marginal probability), 其不考虑任何  $B$  是否发生的因素。

古典概型: 有限个样本点, 每个样本点都可能出现

条件概率:  $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(AB)}{P(B)}$

$$P(AB) = P(A)P(B)$$

$$P(A) = P(A|B_1)P(B_1) + \dots + P(A|B_n)P(B_n) \quad B_1, \dots, B_n \text{ 为 } \Omega \text{ 的划分 (e.g. } B \text{ 和 } \bar{B})$$

$$P(AB) = P(A)P(B) \Leftrightarrow \text{相互独立}$$

## Chapter 3. 随机变量

- 离散型随机变量可使用概率质量函数(probability mass function, PMF)描述其
- 连续型随机变量可使用概率密度函数(probability density function, PDF)描述其
- 两类随机变量都可使用累积分布函数(cumulative distribution function, CDF)描述其概率分布。

$$\text{PMF: } p(a_i) = p_i = P(X=a_i), \quad \sum p_i = 1 \quad \text{高散型.}$$

① 伯努利分布: 结果只有 A,  $\bar{A}$ ,  $P(X=1)=p$ ,  $P(X=0)=1-p \Rightarrow X \sim B(p)$

$$\text{PMF: } p(x) = p^x (1-p)^{1-x}, \quad x=0,1$$

n重伯努利  $\Rightarrow$  二项分布  $X \sim B(n,p)$   $\text{PMF: } p(x) = \binom{n}{x} p^x (1-p)^{n-x}$

可列重伯努利

可列重伯努利试验与几何分布

- 可列重伯努利试验是指将伯努利试验 独立重复 进行无穷次的试验, 也就是  $n$  重伯努利试验中, 试验次数  $n \rightarrow \infty$  时的试验。
- 令随机变量  $X$  表示在可列重伯努利试验中事件  $A$  首次发生前  $\bar{A}$  发生的次数, 且  $P(A) = p$ , 则称  $X$  服从参数为  $p$  的 几何分布 (geometric distribution), 记为  $X \sim \text{Geo}(p)$ 。  
③
- 不难推出几何分布的概率质量函数为  
 $p(x) = p(1-p)^x, x=0,1,2,\dots$
- 几何分布是唯一具有 无记忆性(memoryless property)的离散型概率分布: 设  $X \sim \text{Geo}(p)$ , 则对任何正整数  $m$  和  $n$ , 都有

$$P(X > m + n | X > m) = P(X > n).$$

$X \in \{0,1,2,\dots\}$   $\text{PMF: } p(x) = \frac{x^x}{x!} e^{-\lambda}, \quad x=0,1,2,\dots \Rightarrow X \sim P(\lambda)$

$X \sim P(\lambda)$  的方差和期望:

$$P(X=k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k=0,1,\dots$$

$$\therefore E(X) = \sum_{k=0}^{\infty} k \cdot P(X=k) = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \stackrel{\text{Taylor}}{=} \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda$$

$$D(X) = E(X^2) - (E(X))^2$$

$$E(X^2) = E[X(x-1)+x] = \sum_{k=0}^{\infty} k(k-1) \frac{\lambda^k}{k!} e^{-\lambda} + E(X) = \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda = \lambda^2 \cdot e^{-\lambda} \cdot e^{\lambda} + \lambda = \lambda^2 + \lambda$$

## 连续型

$$\text{PDF: } f(x) \geq 0, \int_{-\infty}^{\infty} f(x) dx = 1 \quad \therefore P(X=x) = 0$$

$f(x)$  反映概率集中于  $x$  附近的程度

① 正态分布与标准正态分布

如果连续型随机变量  $X$  具有概率密度函数

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

则称  $X$  服从参数为  $\mu$  和  $\sigma^2$  的 正态分布(normal distribution), 记为  $X \sim N(\mu, \sigma^2)$ . 特别地,  $N(0, 1)$  称为 标准正态分布(standard normal distribution).

固定  $p, n$  很大,  $B(n, p)$  逼近  $N(np, np(1-p))$

入很大时,  $P(\lambda) \approx N(\lambda, \lambda)$

$$X \sim N(\mu, \sigma^2) \quad Z = \frac{X-\mu}{\sigma} \Rightarrow Z \sim N(0, 1)$$

$$\Rightarrow P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P(Z \leq z)$$

② 指数分布  $f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}, \quad X \sim \text{Exp}(\lambda)$

$$\text{CDF: } F(x) = \int_{-\infty}^x f(t) dt = \begin{cases} 1-e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

"事件发生的时间间隔". 唯一无记忆性连续性分布.

③ 均匀分布 POF:  $f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}, \quad X \sim U(a,b)$

期望与方差的性质

令  $X_1, X_2, \dots, X_n$  均表示随机变量,  $c$  表示常数, 则:

- $E(cX) = cE(X), \text{Var}(cX) = c^2 \text{Var}(X)$ .
- $E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$ .
- $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) + 2E[(X_1 - E(X_1))(X_2 - E(X_2))]$ .
- 如果随机变量  $X_1, X_2, \dots, X_n$  相互独立, 则  
 $E(X_1 X_2 \dots X_n) = E(X_1)E(X_2) \dots E(X_n)$ ,  
 $\text{Var}(X_1 + X_2 + \dots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n)$ .

· 大数定律

大数定律本质上说的就是: 当我们对随机变量的值进行重复观测时, 这些值的算术平均会趋近于随机变量的期望值:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \xrightarrow{n \rightarrow \infty} \mu. \quad \text{并不严格的描述}$$

特别地, 当随机变量服从 Bernoulli 分布时, 它表示观测到的事件发生频率会收敛到事件的真实概率.

## 中心极限定理

中心极限定理本质上说的就是: 一组满足一定要求的随机变量的算术平均渐近服从正态分布:

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{n \rightarrow \infty} N(0, 1) \quad \text{or} \quad \bar{X} \xrightarrow{n \rightarrow \infty} N\left(\mu, \frac{\sigma^2}{n}\right).$$

## Chapter 4.

"简单随机抽样"  $\Rightarrow$  独立同分布 好成绩不好  $\Rightarrow$  伯克森悖论

系统抽样、分层抽样、整群抽样 R 语言  $\rightarrow$  ggplot2 包

可视化: 散点图、柱状图、直方图、箱线图、相关热图

Chapter 5. 5.1 估计量:  $\bar{X} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

$$\text{clip} = \begin{cases} x_{n+1}, & x_{n+1} \\ \frac{x_{n+1} + x_{n+2}}{2}, & 1 \leq x_{n+1} \leq x_{n+2} \\ x_{n+2}, & x_{n+2} \geq x_{n+1} \end{cases} \quad \begin{array}{c} \text{上估计} \\ \downarrow \end{array} \quad X \sim f(x; \theta) \quad \theta(X_1, \dots, X_n) \text{ 估计 } \theta$$

$\hat{\theta}_1, \hat{\theta}_2$  估计量  $\hat{\theta}_1(X_1, \dots, X_n)$  估计值 可能成立?

无偏估计:  $\hat{\theta}_1(\theta) = \theta$   $\hat{\theta}_2(\theta) = \theta$   $\hat{\theta}_3(\theta) = \theta$  无偏估计量

$$E(\hat{\theta}_1) = \theta \quad E(\hat{\theta}_2) = \theta \quad E(\hat{\theta}_3) = \theta$$

证明方法:

有效性:  $\hat{\theta}_1, \hat{\theta}_2$  为偏估计量  $\forall \theta \neq \theta$ ,  $\text{Var}_{\theta}(\hat{\theta}_1) \leq \text{Var}_{\theta}(\hat{\theta}_2)$  且不恒等,  $\Rightarrow \hat{\theta}_1$  比  $\hat{\theta}_2$  有效

点估计求解 { 短估计  $E(X^k)$   $k$  所需量.  $E(X-E(X))^k$   $k$  所需中心量 }

| 极大似然估计 依据: 大数定律

$\hat{\theta}_1(\theta_1, \dots, \theta_n)$  似然函数 取 max 时,  $\hat{\theta}_1(X_1, \dots, X_n)$  极大似然估计量

置信区间  $X \sim f(x; \theta)$ ,  $P_{\theta}(\hat{\theta}_1 < \theta < \hat{\theta}_2) \geq 1-\alpha$ ,  $\forall \theta \in \Theta$ ,  $0 < \alpha < 1$

$(\hat{\theta}_1, \hat{\theta}_2)$  置信水平为  $1-\alpha$  的置信区间  $\hat{\theta}_1$  下限  $\hat{\theta}_2$  上限

## 5.3 假设检验 $H_0, H_1$ , RR: 拒绝域

I类错误:  $H_0 \vee$  拒绝  $H_0$

II类错误:  $H_1 \vee$  没拒绝  $H_0$

不可能同时减小  $\Rightarrow$  I不太高, II尽量低

第I类错误(Type I error): 当  $H_0$  是真相时, 我们拒绝了  $H_0$ . 假阳

第II类错误(Type II error): 当  $H_1$  是真相时, 我们没能拒绝  $H_0$ . 假阴

显著性检验

设  $X = (X_1, X_2, \dots, X_n)$  为来自总体  $X \sim f(x; \theta)$  的一个简单随机样本, 对于一个检验问题  $H_0: \theta \in \Theta_0 \leftrightarrow H_1: \theta \in \Theta_1$  及任意  $0 < \alpha < 1$ , 若一个检验的拒绝域  $RR$  满足

$$P_\theta(X \in RR) \leq \alpha, \forall \theta \in \Theta_0.$$

则称该检验为显著性水平(significance level)为  $\alpha$  的显著性检验(significance test / test of significance), 简称水平为  $\alpha$  的检验(test with level  $\alpha$ ).

P-value: 假定  $H_0 = \text{true}$ , 观测数据或更极端的数据能观测到的概率

$P < \alpha$  (如 0.95) 拒绝  $H_0$ .

## Chapter 6.

### 6.1 回归回归

回归模型(regression model)简单来说就是用于分析一个因变量(响应变量, response variable)的期望和一些自变量(解释变量, explanatory variables)之间的关系:

$$E(Y|X) = f(X).$$

以一元线性回归模型(simple linear regression model)为例:

$$Y = E(Y|X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon.$$

参数  $\beta_0$  (截距, intercept)与  $\beta_1$  (斜率, slope)是需要估计的参数, 统称为回归系数(regression coefficients),  $\varepsilon$  是随机误差.

随机误差表示的是无法通过  $X$  解释的  $Y$  的取值波动.

通常假设随机误差服从正态分布, 即  $\varepsilon \sim N(0, \sigma^2)$ .

需要通过观察数据  $\{(x_i, y_i), i=1, 2, \dots, n\}$  估计参数  $\beta_0$  与  $\beta_1$ .

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

$$\text{方差 } S^2: \text{无偏估计: } S^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2$$

6.2 逻辑回归 先将  $f(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$  映射到 [0, 1]

$$\text{logistic 函数: } \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \Rightarrow p = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}.$$

估计  $\beta_0, \beta_1, \dots, \beta_p$ : 极大似然法.  $L(\beta_0, \beta_1) = \prod p^{y_i} (1-p)^{1-y_i}$

$$\Rightarrow L(\cdot) = \ln L(\cdot)$$

R:  $\text{glm}()$   $\Rightarrow$  回归系数 梯度 =  $\frac{\partial}{\partial \beta_j}$

$$\frac{\partial \ln L}{\partial \beta_j}$$

评价:  $\hat{y}_i = 1 \quad \hat{y}_i = 0$

$y_i = 1$	$\hat{y}_i = 1$	① Accuracy = $\frac{TP + TN}{n} \times 100\%$ . 准确率
TP	FP	② Precision = $\frac{TP}{TP + FP} \times 100\%$ . 精确率
$y_i = 0$	$\hat{y}_i = 0$	性, sensitivity): 表示在实际为正的
FP	TN	③ Recall = $\frac{TP}{TP + FN} \times 100\%$ . 召回率/敏感性
ROC 曲线	④ * Perfect Random	实际为负的记录中, 判断为负的比
		⑤ Specificity = $\frac{TN}{TN + FP} \times 100\%$ . 鉴别率

下面讲(AUC) Random: AUC=0.5 AUC  $\uparrow$  Better.

### 6.3 k近邻法

优点是方法逻辑简单, 易于实现, 计算效率高, 结果有很好的可解释性.

缺点是方法假设了各个特征之间的条件独立性, 这个假设在实际中往往不成立, 因此在特征之间相关性较强时, 分类效果不好.

$$\text{条件概率/似然} \quad \text{先验概率}$$

$$\Rightarrow P(\text{类别} | \text{特征}) = \frac{P(\text{特征} | \text{类别})P(\text{类别})}{P(\text{特征})}$$

后验概率

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

### 6.4 决策树

优点是方法逻辑简单, 易于实现, 计算效率高, 结果有很好的可解释性.

缺点是方法假设了各个特征之间的条件独立性, 这个假设在实际中往往不成立, 因此在特征之间相关性较强时, 分类效果不好.

将下面将这个思路用数学公式表达出来.

二维平面上的点  $(x_1, x_2)$  到直线  $w_1 x_1 + w_2 x_2 + b = 0$  的距离公式为

$$\frac{|w_1 x_1 + w_2 x_2 + b|}{\sqrt{w_1^2 + w_2^2}}.$$

推广到  $p$  维空间, 点  $x_0 = (x_{01}, x_{02}, \dots, x_{0p})$  到直线  $w^\top x + b = 0$  的距离为

$$\frac{|w^\top x_0 + b|}{\sqrt{w_1^2 + w_2^2 + \dots + w_p^2}} = \frac{|w^\top x_0 + b|}{\|w\|}.$$

最大间隔的数学表达:

对于训练数据  $\{(y_i, x_i), i=1, 2, \dots, n\}$ , 当  $w^\top x_i + b > 0$  时, 令  $y_i = 1$ , 当  $w^\top x_i + b < 0$  时, 令  $y_i = -1$ .

寻找最大间隔等价于求解如下优化问题:

$$\max_{w, b} \min_{i=1, \dots, n} \frac{y_i(w^\top x_i + b)}{\|w\|}.$$

记  $\gamma = \min_{i=1, \dots, n} y_i(w^\top x_i + b)$ , 则 SVM 的求解可表示为如下约束优化问题:

$$\max_{w, b} \frac{\gamma}{\|w\|}$$

$$\text{s.t. } y_i(w^\top x_i + b) \geq \gamma, i = 1, \dots, n.$$

由于  $\gamma$  的大小不影响此优化问题的求解, 不妨令  $\gamma = 1$  (可以理解为坐标系的缩放), 最终 SVM 的求解表示为如下约束优化问题:

$$\min_{w, b} \frac{1}{2} \|w\|^2$$

$$\text{s.t. } y_i(w^\top x_i + b) \geq 1, i = 1, \dots, n.$$

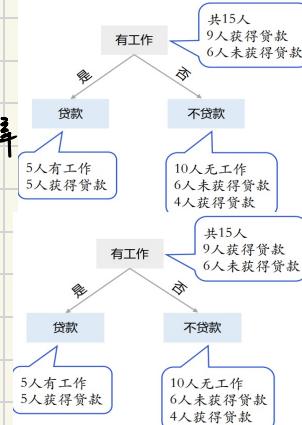
## Chapter 7. 7.1 树木和森林

$$\text{信息熵 } H(Y) = -\sum_{i=1}^k p_i \log_2 p_i$$

$$\text{信息增益 } g(Y, X) = H(Y) - H(Y|X) \quad g \uparrow \Rightarrow \text{优异}$$

$$\text{---地 } g_R(Y, X) = \frac{g(Y, X)}{H(X)}$$

$$\text{基尼不纯度: } \text{Gini}(Y) = \sum_{i=1}^k p_i(1-p_i) = 1 - \sum_{i=1}^k p_i^2 \quad \text{降低} \rightarrow \text{优异}$$



$$H(Y) = -\frac{9}{15} \log_2 \left(\frac{9}{15}\right) - \frac{6}{15} \log_2 \left(\frac{6}{15}\right) = 0.971.$$

$$H(Y|\text{有工作}) = 0,$$

$$H(Y|\text{无工作}) = -\frac{6}{10} \log_2 \left(\frac{6}{10}\right) - \frac{4}{10} \log_2 \left(\frac{4}{10}\right) = 0.971.$$

$$H(Y|\text{工作}) = \frac{5}{15} \times H(Y|\text{有工作}) + \frac{10}{15} \times H(Y|\text{无工作}) = 0.647.$$

$$g(Y|\text{工作}) = H(Y) - H(Y|\text{工作}) = 0.971 - 0.647 = 0.324.$$

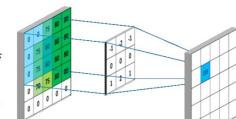
$$\text{Gini}(Y|\text{有工作}) = 1 - \left(\frac{9}{15}\right)^2 - \left(\frac{6}{15}\right)^2 = 0.48.$$

$$\text{Gini}(Y|\text{无工作}) = 1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2 = 0.48.$$

$$\text{Gini}(Y|\text{工作}) = \frac{5}{15} \times \text{Gini}(Y|\text{有工作}) + \frac{10}{15} \times \text{Gini}(Y|\text{无工作}) = 0.32.$$

$$\text{Gini}(Y; \text{工作}) = \text{Gini}(Y) - \text{Gini}(Y|\text{工作}) = 0.48 - 0.32 = 0.16.$$

$$\text{Gini}(Y; \text{房产}) = \text{Gini}(Y) - \text{Gini}(Y|\text{房产}) = 0.48 - 0.267 = 0.213.$$



6. (10 分) 假设一个卷积神经网络的输入图片是一个  $227 \times 227$  像素矩阵, 经过一个  $11 \times 11$  卷积核矩阵的卷积操作(卷积步长为 4, 无边距扩展), 得到的特征矩阵的维度是多少? 请给出计算细节.  $(227-11)/4+1 = 55$ .  $55 \times 55$

## 卷积步长、边距扩展

卷积步长表示卷积核滑动时横向及纵向移动时的单元格数量.

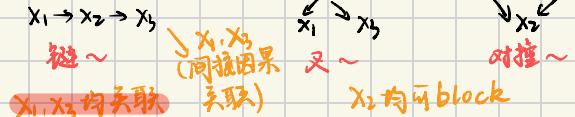
边距扩展是在指定卷积核尺寸与卷积步长后, 希望经过卷积后的特征图具有指定大小而在输入矩阵周围补 0 的操作.

## 循环神经网络 (RNN)

不能保持长期信息间依赖关系

### Chapter 8 因果推断

2 通过, 3 节点  $\Rightarrow$  缘合



在回归模型中, 把混淆因子作为协变量加入模型即可达到切断后门路径的效果. 如:

$$E(Y|T=t, X=x) = \beta_0 + \beta_1 t + \beta_2 x,$$

$\beta_1$  就是  $T$  对  $Y$  的因果效应.

但在未考虑变量间的因果关系时, 盲目地在回归模型中加入更多变量是不可靠的, 因为有些变量不是混淆因子, 而是对照因子, 控制对照因子反而会打开后门路径.

如右图, 若要研究  $X_3$  对  $Y$  的因果效应, 则要关闭  $X_3 \leftarrow X_1 \rightarrow Y$  与  $X_3 \leftarrow X_2 \rightarrow Y$  两条后门路径, 即应拟合回归模型  $Y \sim X_3 + X_1 + X_2$  而不是  $Y \sim X_3$ .

若要研究  $X_1$  对  $Y$  的因果效应, 有三条路径,  $X_1 \rightarrow Y$  (直接因果效应),  $X_1 \rightarrow X_3 \rightarrow Y$  (间接因果效应),  $X_1 \rightarrow X_3 \leftarrow X_2 \rightarrow Y$ .

由于  $X_3$  为对照因子, 第三条路径本是关闭的, 而如果在回归模型中加入  $X_3$ , 反而会打开这条路径. 因此应拟合回归模型  $Y \sim X_1$  而不是  $Y \sim X_1 + X_3$  以得到  $X_1$  对  $Y$  的总体因果效应 (=直接因果效应+间接因果效应).

而如果拟合  $Y \sim X_1 + X_3 + X_2$  会关闭第二、三条路径, 得到  $X_1$  对  $Y$  的直接因果效应.

