

# lec03 Mathematical Preliminaries

## Mathematical Preliminaries

### Common Discrete Probability Distributions

#### Generalised Bernoulli Distribution

分类分布 (Categorical Distribution), 也称广义伯努利分布

$$P(X = 1) = p_1, P(X = 2) = p_2, \dots, P(X = k) = p_k$$

其中,  $\sum_{i=1}^k p_i = 1$

#### Multinomial Distribution

多项式分布是二项分布的扩展, 进行  $n$  次独立重复试验, 每次试验都有  $k$  种可能结果

$$\frac{n!}{\prod_{i=1}^k n_i!} \cdot \prod_{i=1}^k p_i^{n_i}$$

其中,  $n_i$  表示事件  $i$  出现的次数,  $\sum_{n_i}^k = n$ ,  $p_i$  表示事件  $i$  发生的概率

## Missing Values

ID	Height (cm)	Weight (kg)
1	175	60
2	183	80
3	183	85
4	178	65
5	203	90
6	185	78
7	155	98
8	180	75
9	173	65
10		85
11	184	80

对于这个表格, 缺少一个数据, 处理缺少数据的方法如下:

- discard
- fill in by hand
- set "missing value"
- mean
- predict

- accept

## Discard

丢弃缺失的数据对应的训练实例（在这个例子下是这一行）  
然后后面两段英语我看不懂，她写了些什么玩意

- Might get away because of the redundancy in the dataset. We might have another student with the same height as student no.10 for whom we have measured the height
- Might not get away because student no.10 was the only student who had obesity. By ignoring student no.10, we lose all the information we had about the positive class

## Fill in by hand

重新注释数据或重新测量缺少特征的实例

- 可靠

但是

- 可能无法再接触这些 subject
- 太慢
- 成本高
- 缺少值可能很多（不还是成本高和太慢吗）

## Set "missingValue"

将 "缺失" 视为该特征的一个类别，并设置一些指示该确实的常量，比如 "missingValue"。  
但是对于数值数据来说不可能，并不能解决问题

## Mean

计算整个数据集的对应特征的均值，并填上

如果缺失值的数据点是数据集中的代表性样本，这可能是一个不错的选择  
但如果这些数据点是异常值，则此方法不准确

## Predict

我们可以训练一个新的分类器来首先预测数据实例中的缺失值  
然后训练第二个分类器来使用所有（原始+预测的缺失值）数据点来预测目标类

## Accept

（这又是在说啥，还是没看懂）

只需保留缺失值的数据点不变，然后让算法（例如分类器）以适当的方式处理缺失值

分类器可能首先尝试提出一个规则来对数据进行分类，而不使用具有缺失值的特征  
如果它能以高精度做到这一点，那么就可以了，不用担心缺失值

## Noisy Data

"Noisy data" 是指分散在数据中的随机误差，可能由于记录不准确、数据损坏导致

在假设 noisy data 是正确的前提下 (没有发现有 noisy data)，根据数据集训练的分类器很可能会发生 **过拟合**

## 检测噪声

显然的噪声：

- 该数据类型与其他数据类型不一致
- 该数据值与其他数据值显著不一致

不明显的噪声：

- 打错了 (0.25打成0.52)

## 处理噪声

- 人工
- 对数据使用聚类 (clustering) 来查找位于主聚类之外的实例或特征 (outlier detection) 并将其删除
- 使用线性回归来确定函数，然后删除那些远离预测值的函数
- 忽略低于特定频率阈值的所有值，这种方法能有效检测文本中的拼写错误
- 如果可以识别并删除噪声点，我们可以应用缺失值技术来填充缺失的特征

## 冗余值

一些机器学习算法可能会受到重复数据的不利影响

# Over-fitting vs. Under-fitting

## 欠拟合解决方法

- 学习尚未收敛：再多跑
- feature 空间太小：实现更多功能
- 训练数据质量差：进行清理与重新注释
- 算法不行：换算法

## 过拟合解决方法

- Reduce the flexibility of the model：正则化 (regularisation)、删除 features
- 提前停止
- 增加数据集
- 交叉验证 (cross-validation)

# Feature Normalization

## [0,1]-scaling

$$\hat{x} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

## Gaussian Normalization

$$\hat{x} = \frac{x - \mu}{\sigma}$$

其中  $\mu$  是均值， $\sigma$  是标准差