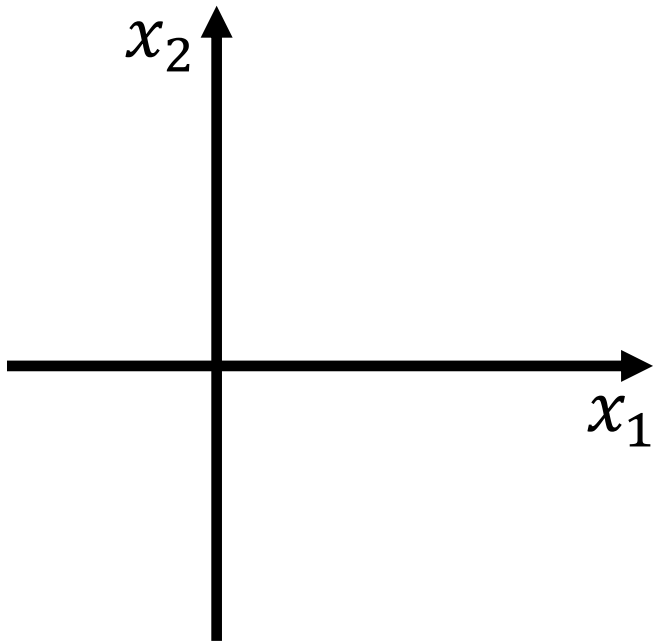
特征工程

Part 3　2025/03/18　凤维杰

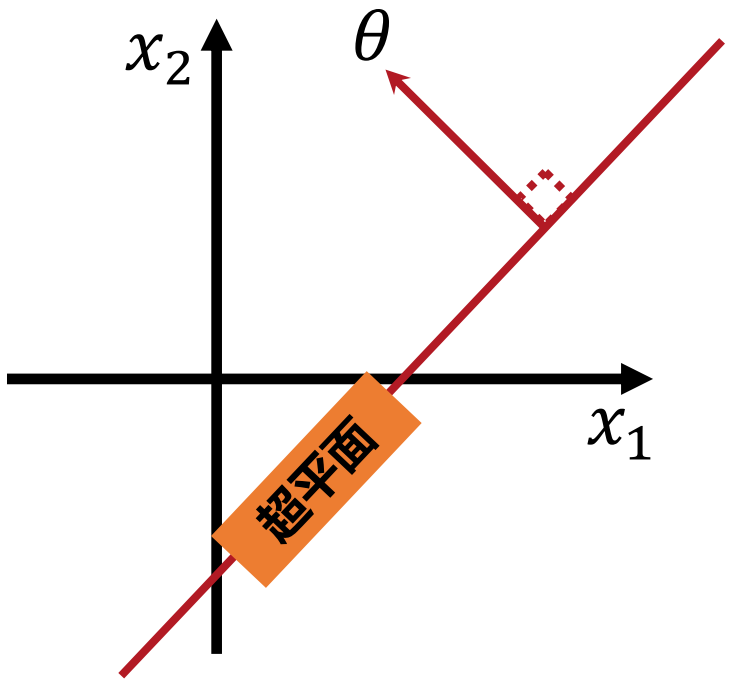
回顾：分类器

- 线性分类器$h(x;\theta,\theta_0)=\text{sign}(\theta^{\mathrm{T}}x+\theta_0)$



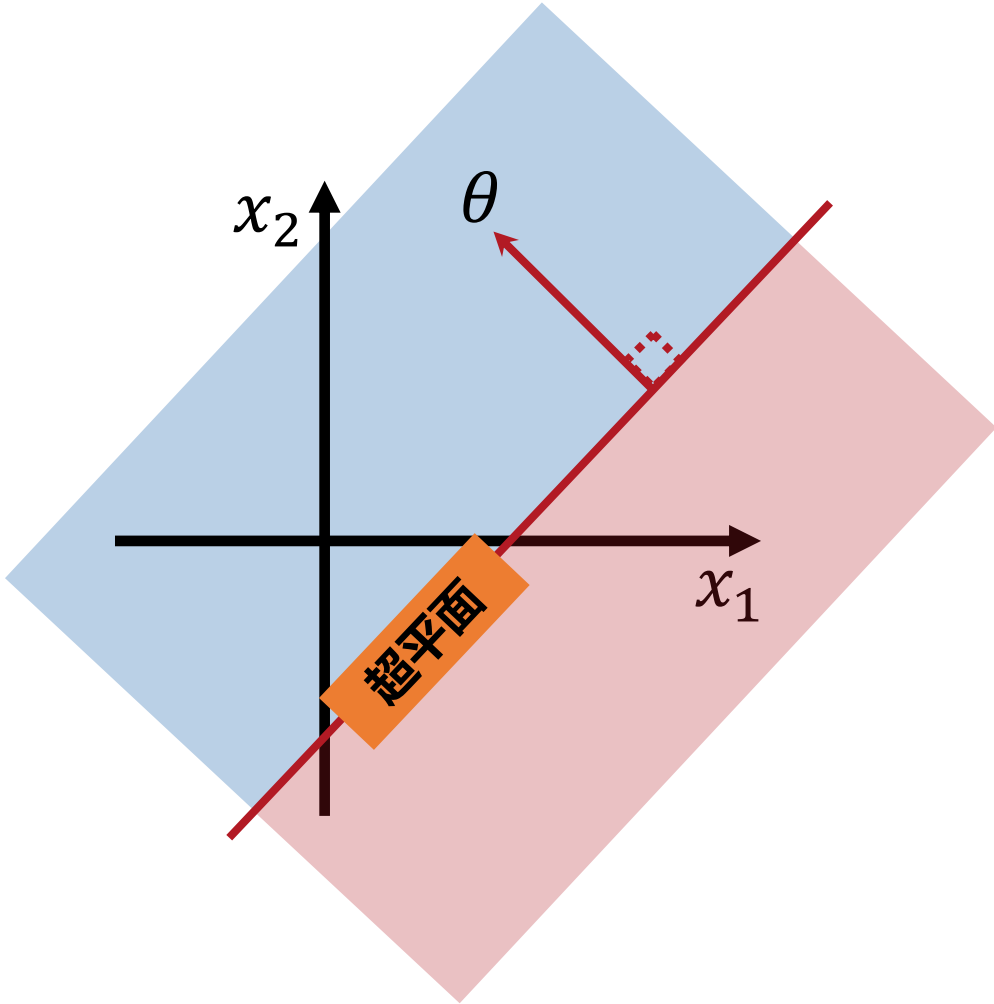
回顾：分类器

- 线性分类器$h(x;\theta,\theta_0)=\text{sign}(\theta^{\mathrm{T}}x+\theta_0)$



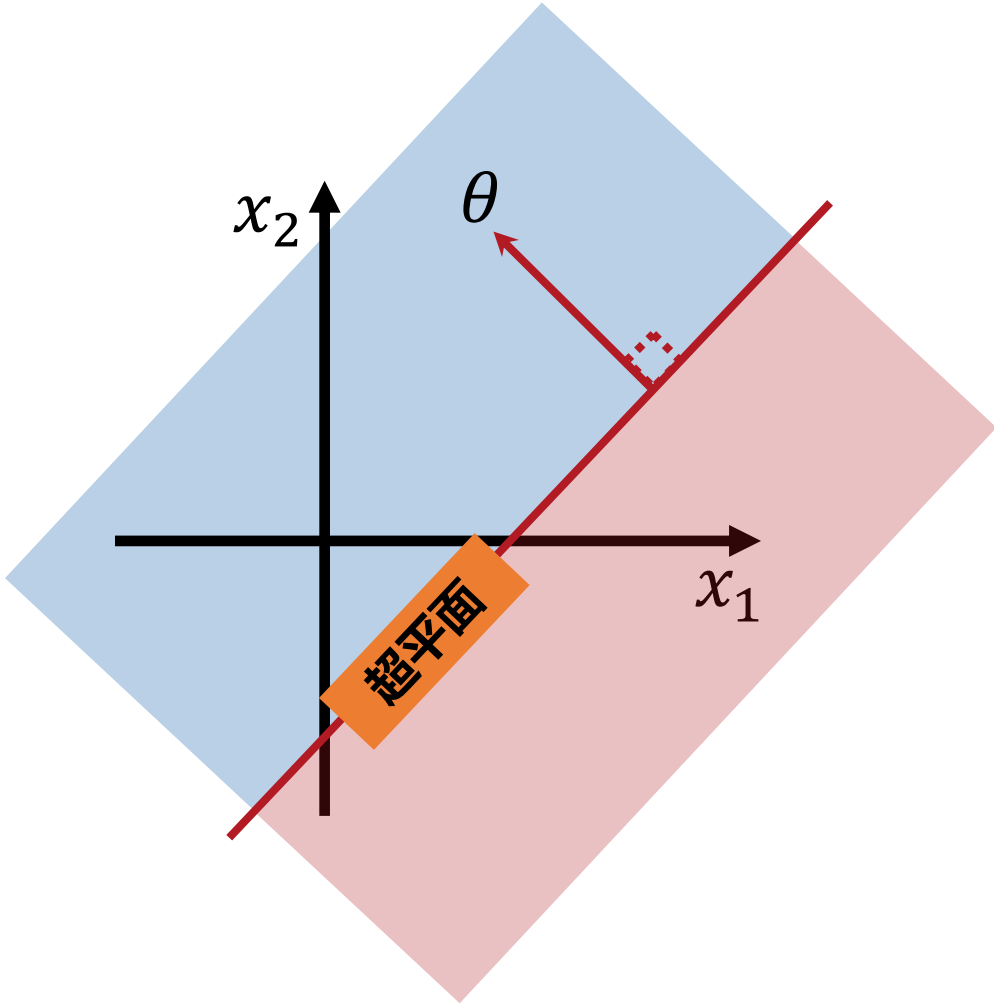
回顾：分类器

- 线性分类器$h(x;\theta,\theta_0)=\text{sign}(\theta^{\mathrm{T}}x+\theta_0)$



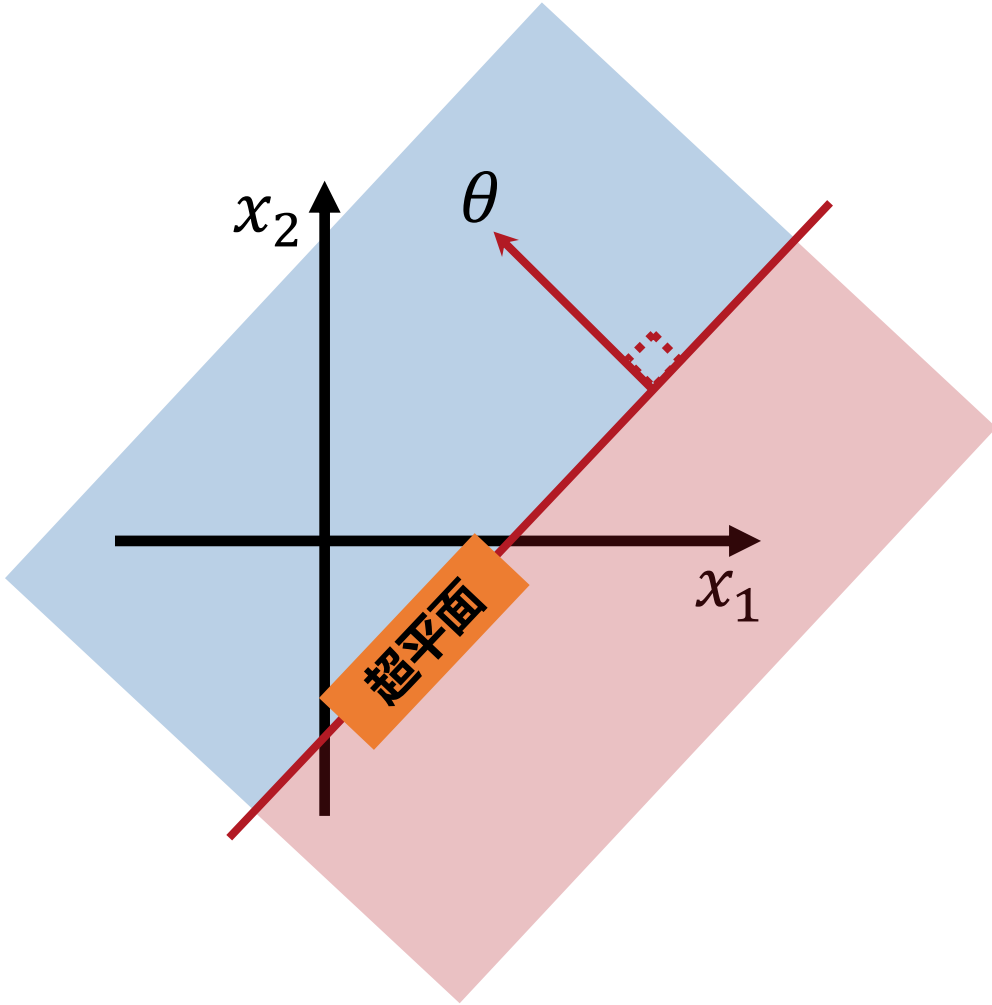
回顾：分类器

- 线性分类器$h(x;\theta,\theta_0) = \text{sign}(\theta^{\mathrm{T}}x + \theta_0)$
- 0-1损失：$L(g,a) = \begin{cases} 0, & \text{if } g = a \\ 1, & \text{else} \end{cases}$



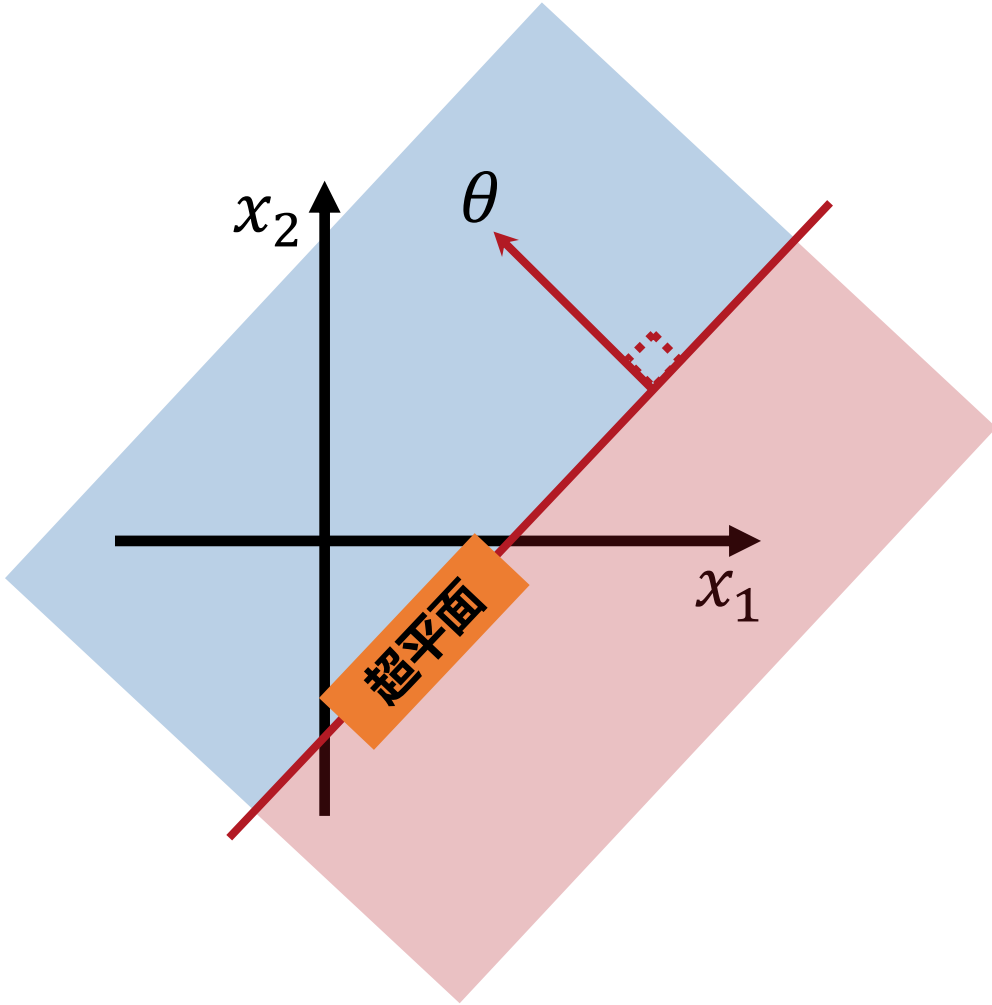
回顾：分类器

- 线性分类器$h(x;\theta,\theta_0) = \text{sign}(\theta^{\mathrm{T}}x + \theta_0)$
- 0-1损失：$L(g,a) = \begin{cases} 0, & \text{if } g = a \\ 1, & \text{else} \end{cases}$
- 训练误差：$\varepsilon_n(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(x^{(i)}), y^{(i)})$



回顾：分类器

- 线性分类器$h(x;\theta,\theta_0) = \text{sign}(\theta^{\mathrm{T}}x + \theta_0)$
- 0-1损失：$L(g,a) = \begin{cases} 0, & \text{if } g = a \\ 1, & \text{else} \end{cases}$
- 训练误差：$\varepsilon_n(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(x^{(i)}), y^{(i)})$

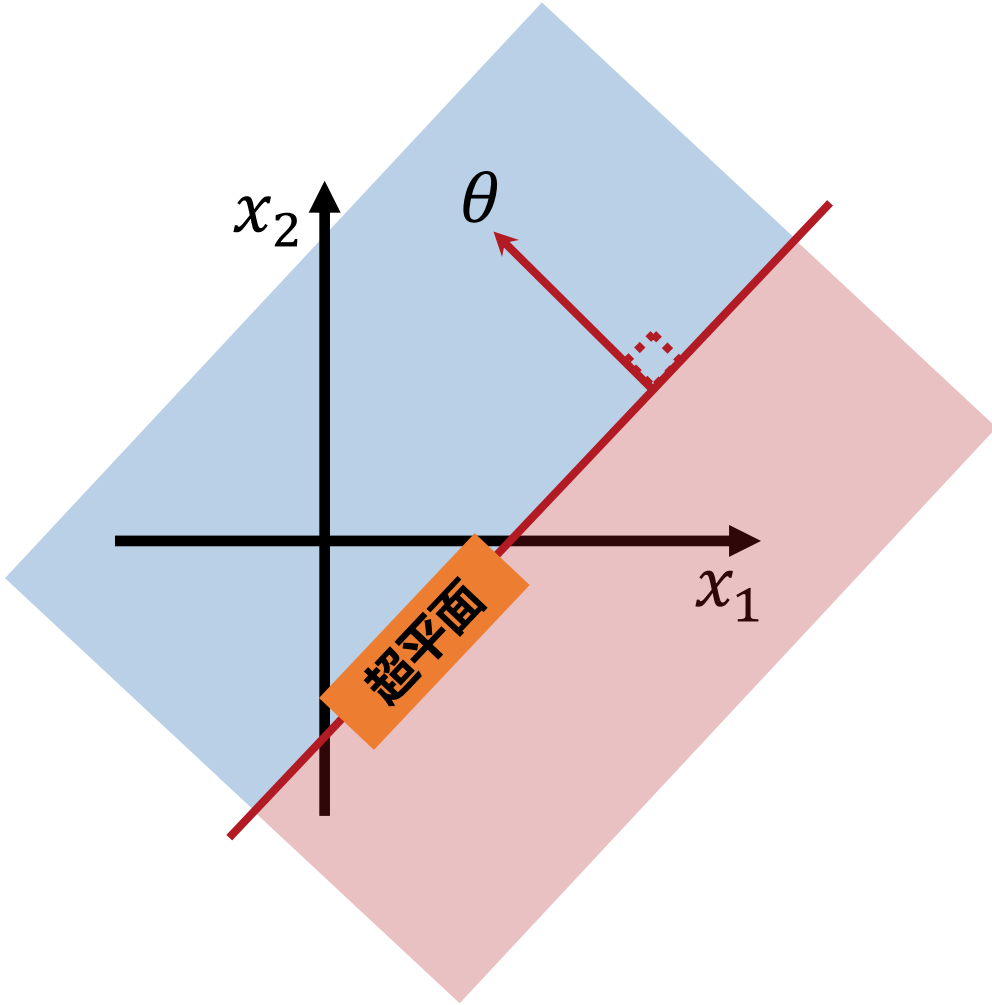


机器学习流程

1. 建立目标 & 收集数据
   - Ex：疾病预测

回顾：分类器

- 线性分类器$h(x;\theta,\theta_0) = \text{sign}(\theta^{\mathrm{T}}x + \theta_0)$
- 0-1损失：$L(g,a) = \begin{cases} 0, & \text{if } g = a \\ 1, & \text{else} \end{cases}$
- 训练误差：$\varepsilon_n(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(x^{(i)}), y^{(i)})$

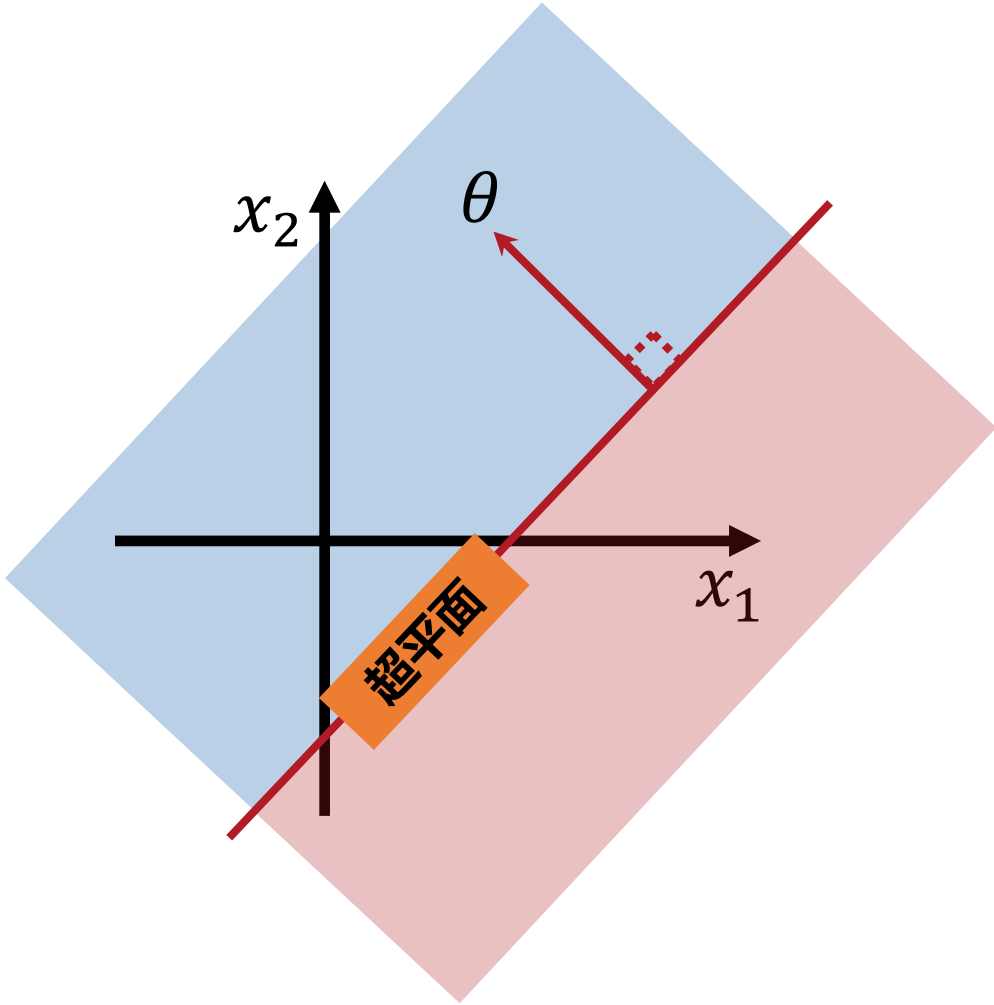


机器学习流程

1. 建立目标 & 收集数据
   - Ex：疾病预测
2. 特征工程

回顾：分类器

- 线性分类器$h(x;\theta,\theta_0) = \text{sign}(\theta^{\mathrm{T}}x + \theta_0)$
- 0-1损失：$L(g,a) = \begin{cases} 0, & \text{if } g = a \\ 1, & \text{else} \end{cases}$
- 训练误差：$\varepsilon_n(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(x^{(i)}), y^{(i)})$



机器学习流程

1. 建立目标 & 收集数据
   - Ex：疾病预测
2. 特征工程
3. 运行学习算法 & 学习分类器
   - Ex：比较学习算法、感知器
4. 分析 & 评估

目标 & 数据

- 目标 & 数据
  - E. g. 心脏病预测

目标 & 数据

- 目标 & 数据
  - E. g. 心脏病预测

	心脏病情况	心率	疼痛症状	职业	用药	年龄	收入
1	否	55	否	护士	止痛药	40+	123000
2	否	71	否	管理	降压药&止痛药	20+	134000
3	是	89	是	护士	降压药	50+	140000
4	否	67	否	医生	无	50+	130000

目标 & 数据

- 目标 & 数据
  - E. g. 心脏病预测
- 将数据编码成适合学习算法的形式

	心脏病情况	心率	疼痛症状	职业	用药	年龄	收入
1	否	55	否	护士	止痛药	40+	123000
2	否	71	否	管理	降压药&止痛药	20+	134000
3	是	89	是	护士	降压药	50+	140000
4	否	67	否	医生	无	50+	130000

目标 & 数据

- 目标 & 数据
  - E. g. 心脏病预测
- 将数据编码成适合学习算法的形式

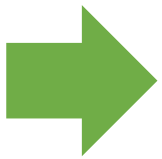
	心脏病情况	心率	疼痛症状	职业	用药	年龄	收入
1	否	55	否	护士	止痛药	40+	123000
2	否	71	否	管理	降压药&止痛药	20+	134000
3	是	89	是	护士	降压药	50+	140000
4	否	67	否	医生	无	50+	130000

数据编码

- 识别标签，并编码成实数

	心脏病情况
1	否
2	否
3	是
4	否

{‘是’, ‘否’}↔{+1, -1}



1	-1
2	-1
3	+1
4	-1

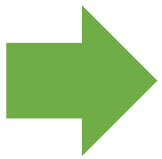
$y^{(1)} = -1$

数据编码

- 识别标签，并编码成实数

	心脏病情况
1	否
2	否
3	是
4	否

{‘是’, ‘否’}↔{+1, -1}



1	-1
2	-1
3	+1
4	-1

$y = y(1)$

- 编码形式自由(e.g. {0,1})，取决于算法
- 保存映射字典以实现标签的输出

数据编码

- 识别特征，并编码成实数
- 特征：除标签外的其余部分

	心率	疼痛症状	职业	用药	年龄	收入
1	55	否	护士	止痛药	40+	123000
2	71	否	管理	降压药&止痛药	20+	134000
3	89	是	护士	降压药	50+	140000
4	67	否	医生	无	50+	130000

数据编码

- 识别特征，并编码成实数
- 特征：除标签外的其余部分

$\left(x^{(1)}\right)^T$

	心率	疼痛症状	职业	用药	年龄	收入
1	55	否	护士	止痛药	40+	123000
2	71	否	管理	降压药&止痛药	20+	134000
3	89	是	护士	降压药	50+	140000
4	67	否	医生	无	50+	130000

数据编码

- 识别特征，并编码成实数
- 特征：除标签外的其余部分
- 原特征x；新特征： $\phi(x)$

	心率	疼痛症状	职业	用药	年龄	收入
$(x^{(1)})^T$ 1	55	否	护士	止痛药	40+	123000
2	71	否	管理	降压药&止痛药	20+	134000
3	89	是	护士	降压药	50+	140000
4	67	否	医生	无	50+	130000

数据编码

- 识别特征，并编码成实数
- 特征：除标签外的其余部分
- 原特征x；新特征： $\phi(x)$

	心率	疼痛症状	职业	用药	年龄	收入
1	55	否	护士	止痛药	40+	123000
2	71	否	管理	降压药&止痛药	20+	134000
3	89	是	护士	降压药	50+	140000
4	67	否	医生	无	50+	130000

数据编码

- 识别特征，并编码成实数
- 特征：除标签外的其余部分
- 原特征x；新特征： $\phi(x)$

	心率	疼痛症状	职业	用药	年龄	收入
1	55	否	护士	止痛药	40+	123000
2	71	否	管理	降压药&止痛药	20+	134000
3	89	是	护士	降压药	50+	140000
4	67	否	医生	无	50+	130000

数据编码

- 识别特征，并编码成实数
- 特征：除标签外的其余部分
- 原特征x；新特征： $\phi(x)$

	心率	疼痛症状	职业	用药	年龄	收入
1	55	0	护士	止痛药	40+	123000
2	71	0	管理	降压药&止痛药	20+	134000
3	89	1	护士	降压药	50+	140000
4	67	0	医生	无	50+	130000

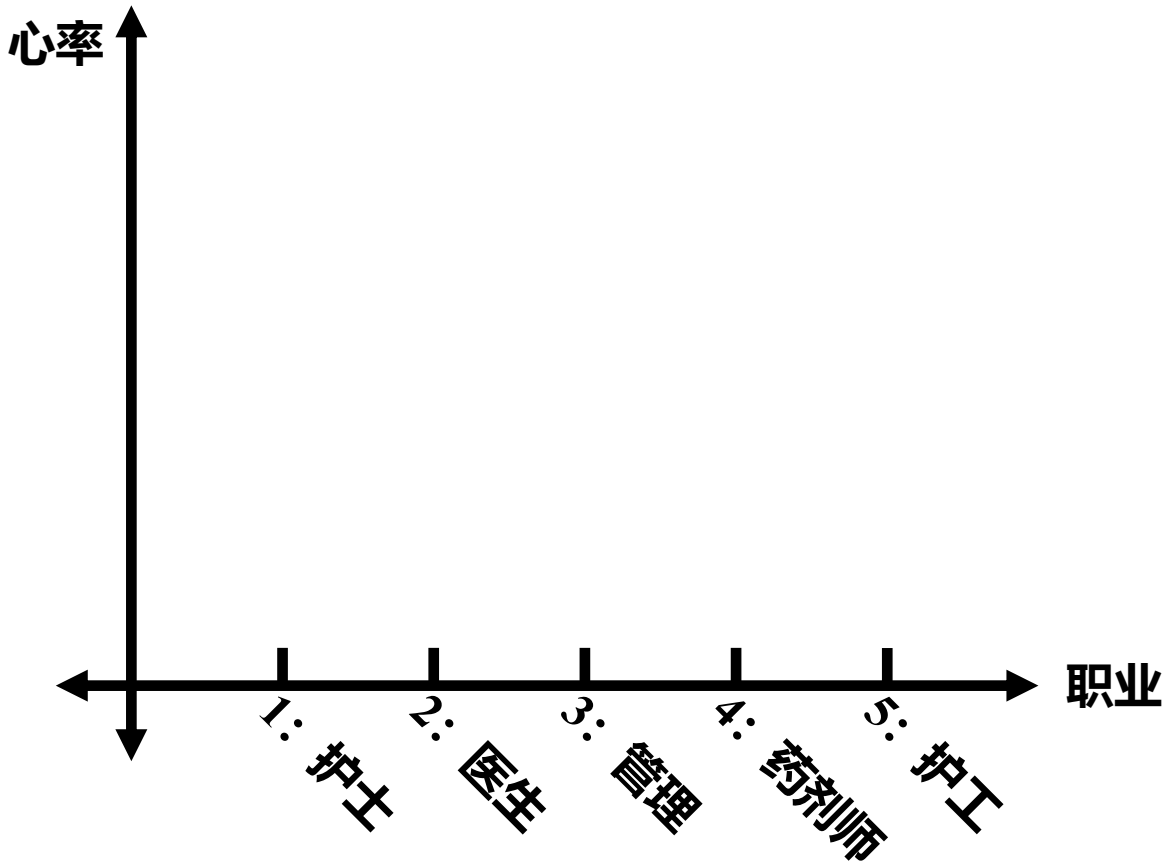
数据编码

- 识别特征，并编码成实数
- 特征：除标签外的其余部分
- 原特征x；新特征： $\phi(x)$

	心率	疼痛症状	职业	用药	年龄	收入
1	55	0	护士	止痛药	40+	123000
2	71	0	管理	降压药&止痛药	20+	134000
3	89	1	护士	降压药	50+	140000
4	67	0	医生	无	50+	130000

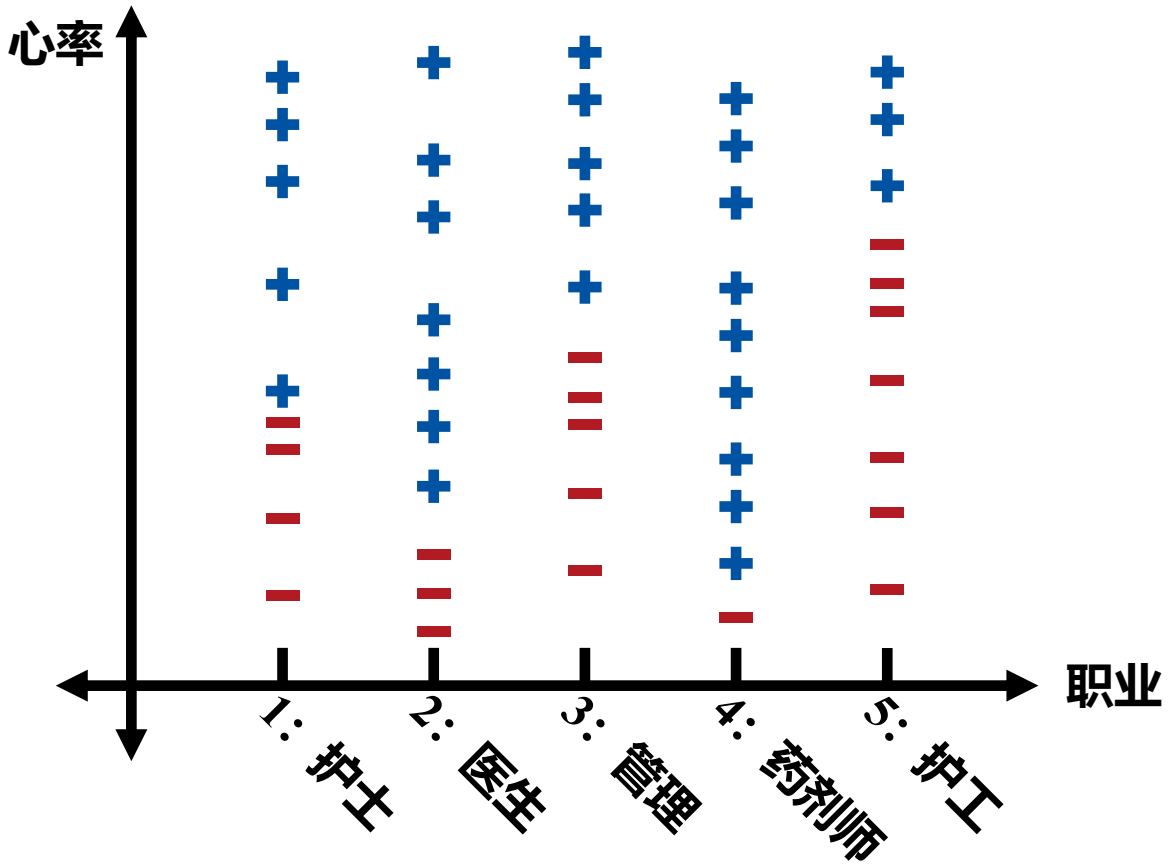
编码类别数据

- Idea 1：将每个类别转换成唯一的自然数



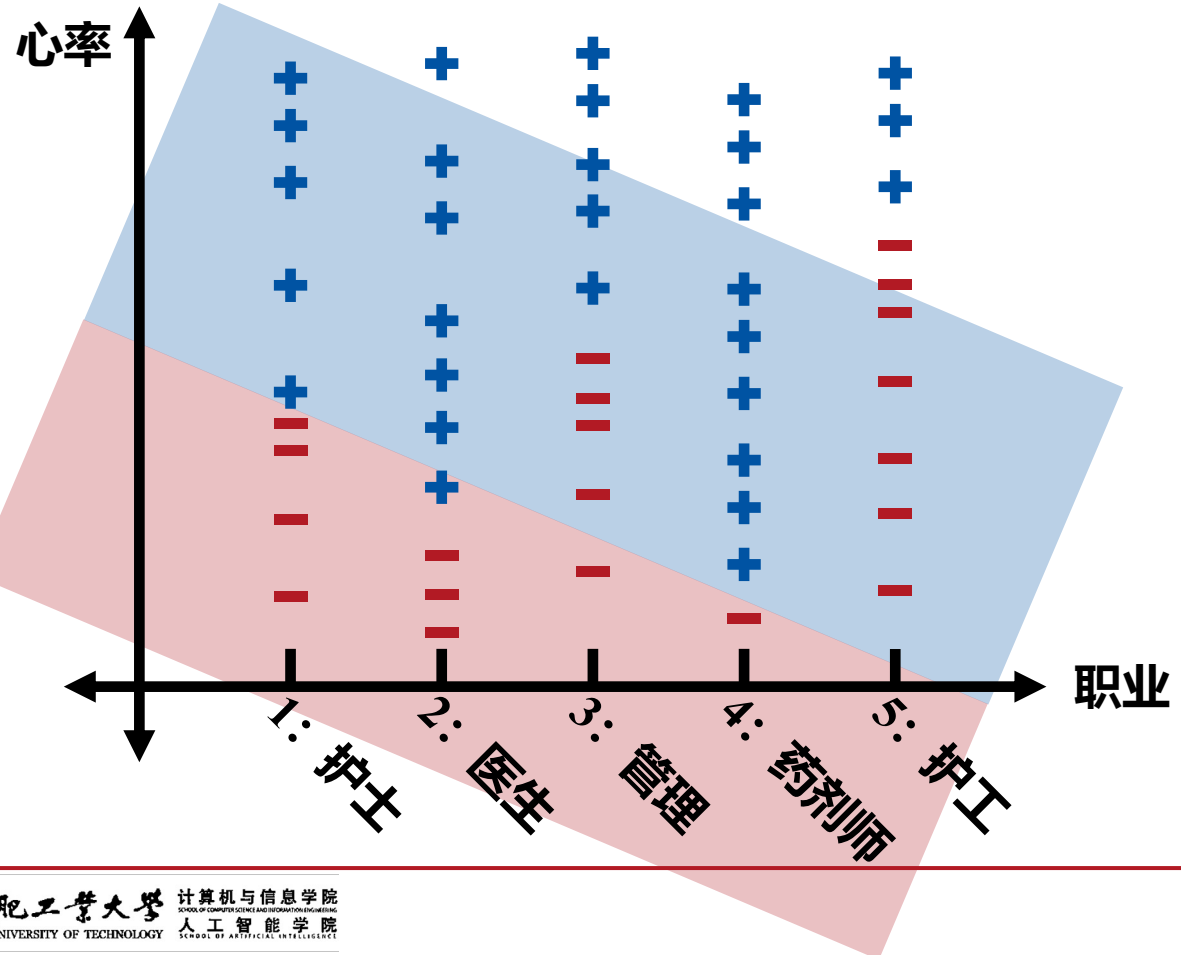
编码类别数据

- Idea 1：将每个类别转换成唯一的自然数



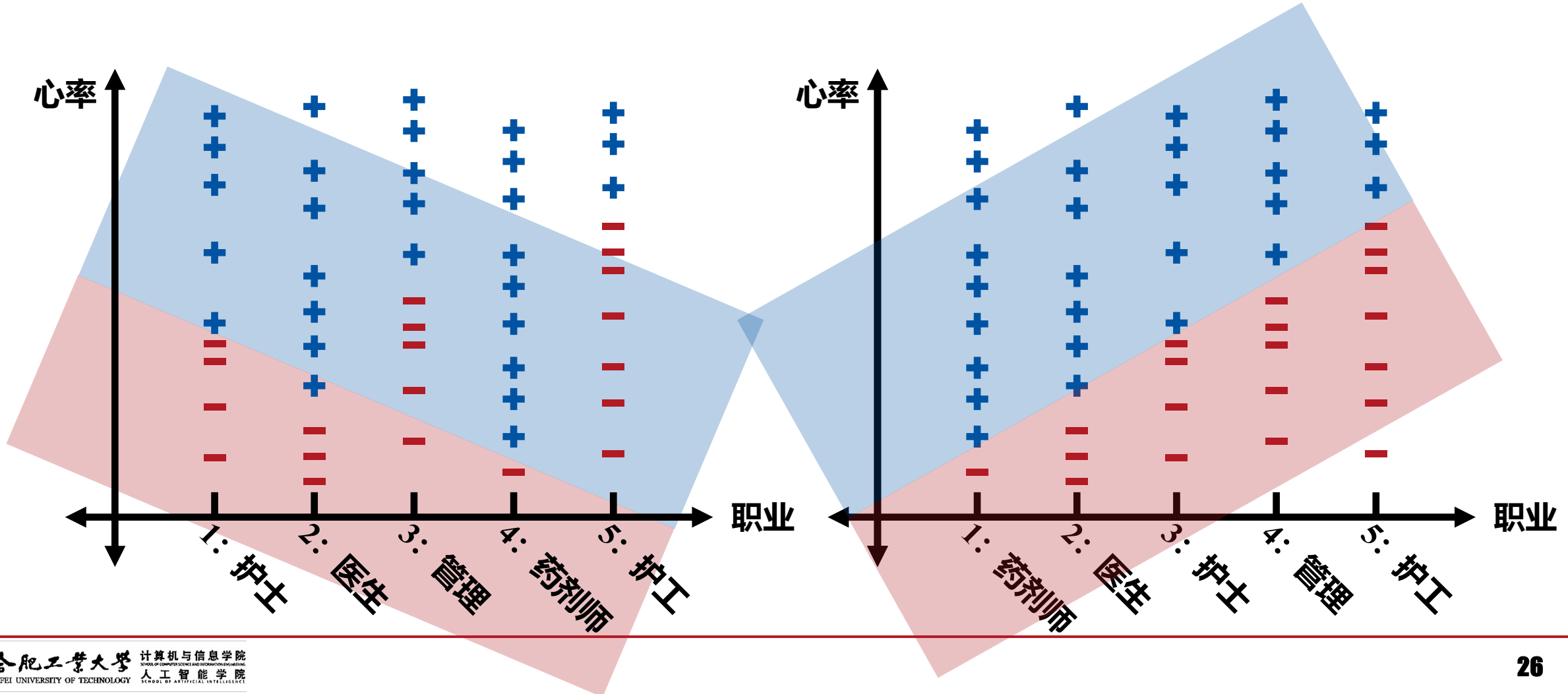
编码类别数据

- Idea 1：将每个类别转换成唯一的自然数



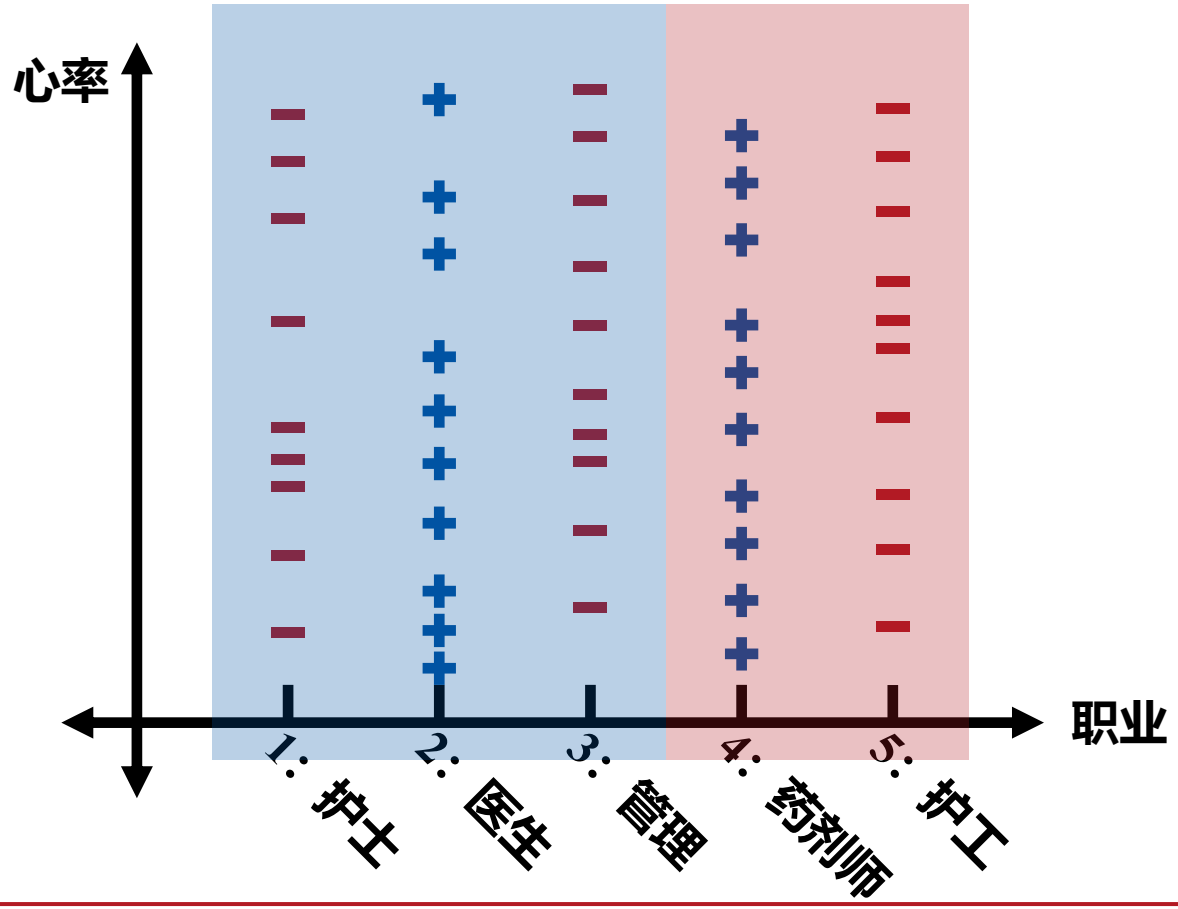
编码类别数据

- Idea 1：将每个类别转换成唯一的自然数



编码类别数据

- Idea 1：将每个类别转换成唯一的自然数



编码类别数据

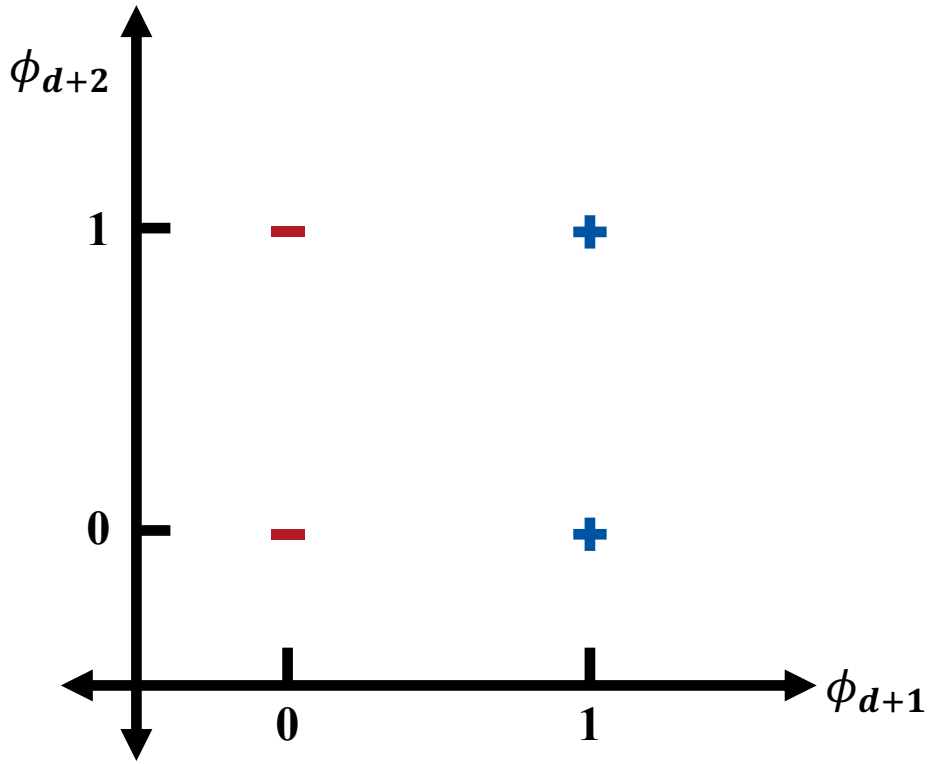
- Idea 2：将每个类别转换成唯一的二进制数

	ϕ_d	ϕ_{d+1}	ϕ_{d+2}
护士	**0**	**0**	**0**
管理	**0**	**0**	**1**
医生	**0**	**1**	**0**
药剂师	**0**	**1**	**1**
护工	**1**	**0**	**0**

编码类别数据

- Idea 2：将每个类别转换成唯一的二进制数

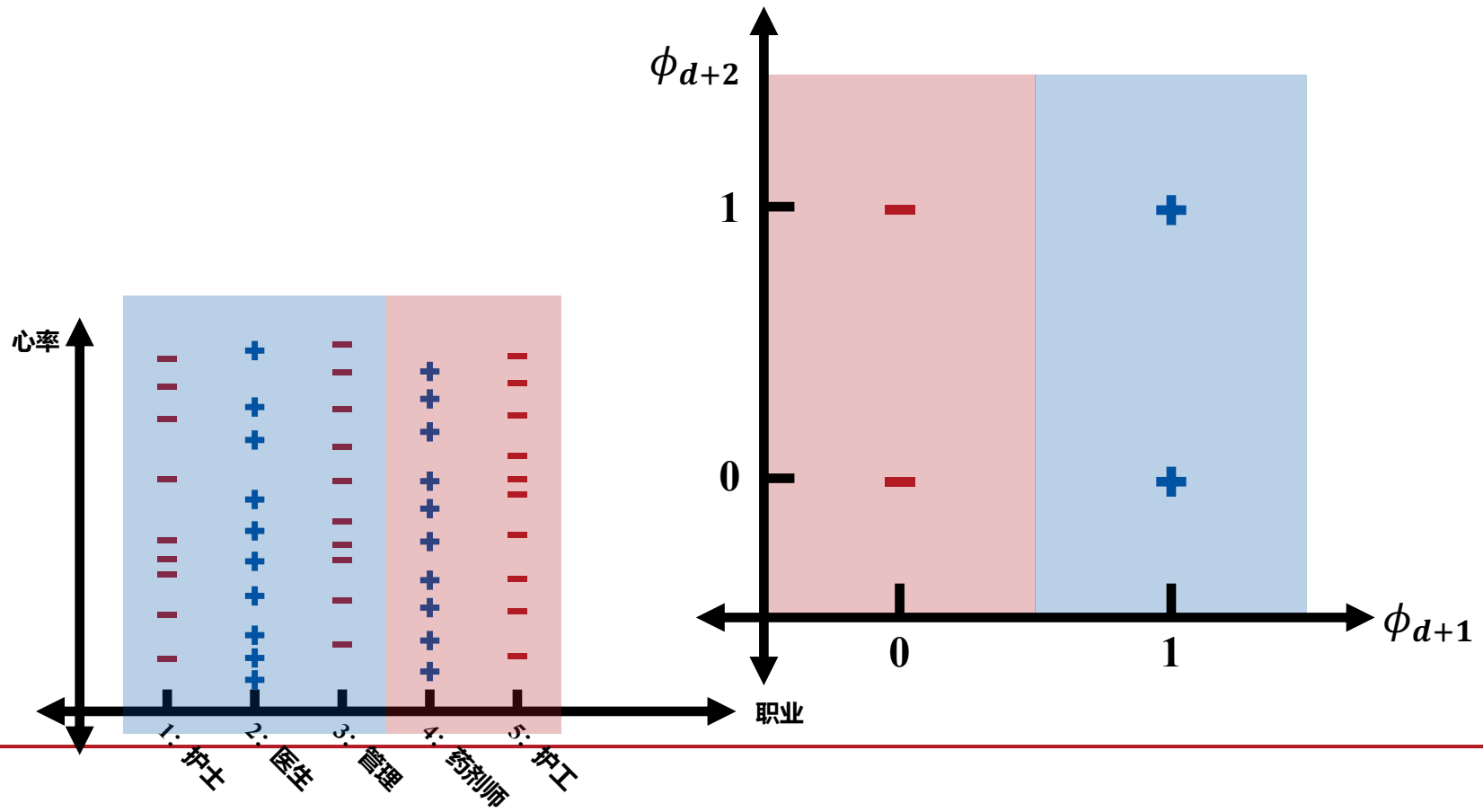
	ϕ_d	ϕ_{d+1}	ϕ_{d+2}
护士	0	0	0
管理	0	0	1
医生	0	1	0
药剂师	0	1	1
护工	1	0	0



编码类别数据

- Idea 2：将每个类别转换成唯一的二进制数

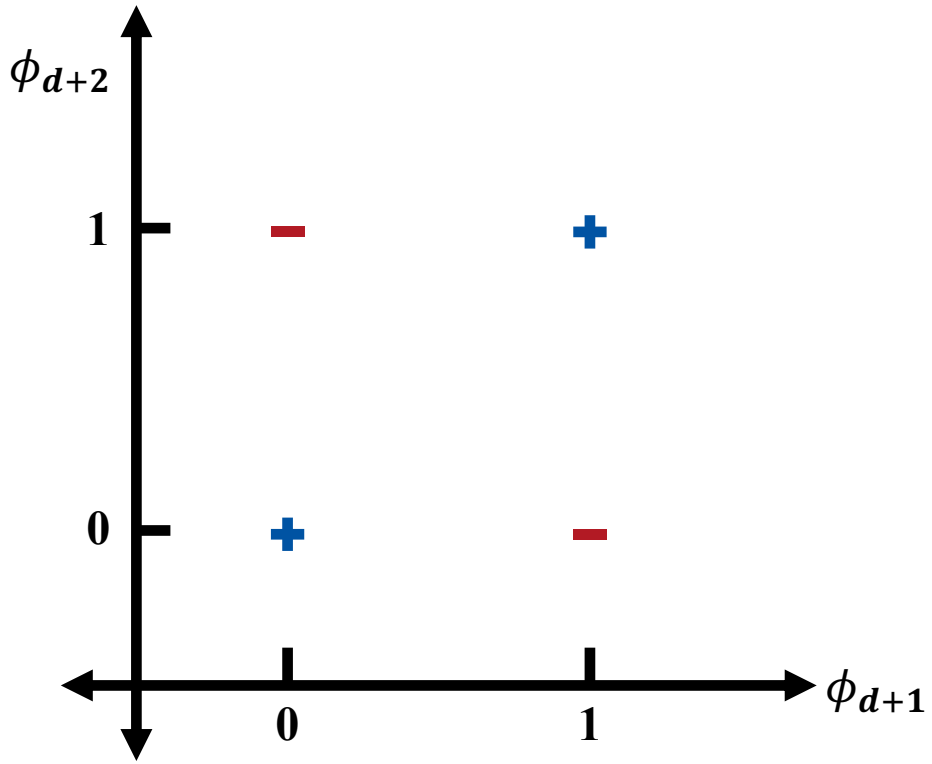
	ϕ_d	ϕ_{d+1}	ϕ_{d+2}
护士	0	0	0
管理	0	0	1
医生	0	1	0
药剂师	0	1	1
护工	1	0	0



编码类别数据

- Idea 2：将每个类别转换成唯一的二进制数

	ϕ_d	ϕ_{d+1}	ϕ_{d+2}
护士	0	0	0
管理	0	0	1
医生	0	1	0
药剂师	0	1	1
护工	1	0	0



编码类别数据

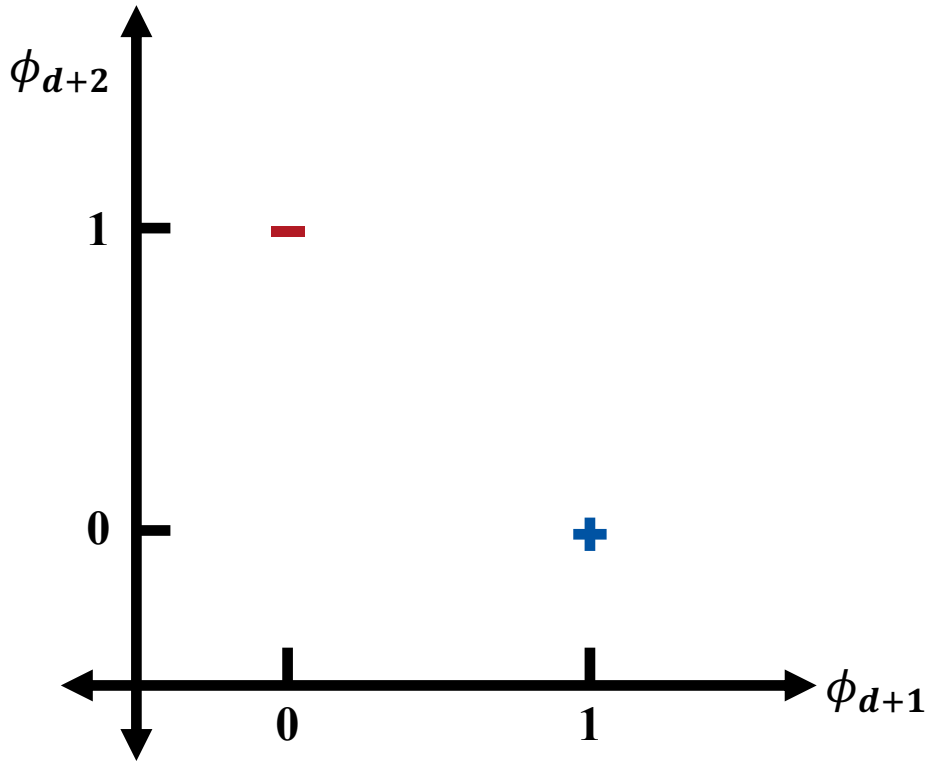
- Idea 3：将每个类别转换成唯一的0-1向量

	ϕ_d	ϕ_{d+1}	ϕ_{d+2}	ϕ_{d+3}	ϕ_{d+4}
护士	**1**	**0**	**0**	**0**	**0**
管理	**0**	**1**	**0**	**0**	**0**
医生	**0**	**0**	**1**	**0**	**0**
药剂师	**0**	**0**	**0**	**1**	**0**
护工	**0**	**0**	**0**	**0**	**1**

编码类别数据

- Idea 3：将每个类别转换成唯一的0-1向量

	ϕ_d	ϕ_{d+1}	ϕ_{d+2}	ϕ_{d+3}	ϕ_{d+4}
护士	1	0	0	0	0
管理	0	1	0	0	0
医生	0	0	1	0	0
药剂师	0	0	0	1	0
护工	0	0	0	0	1

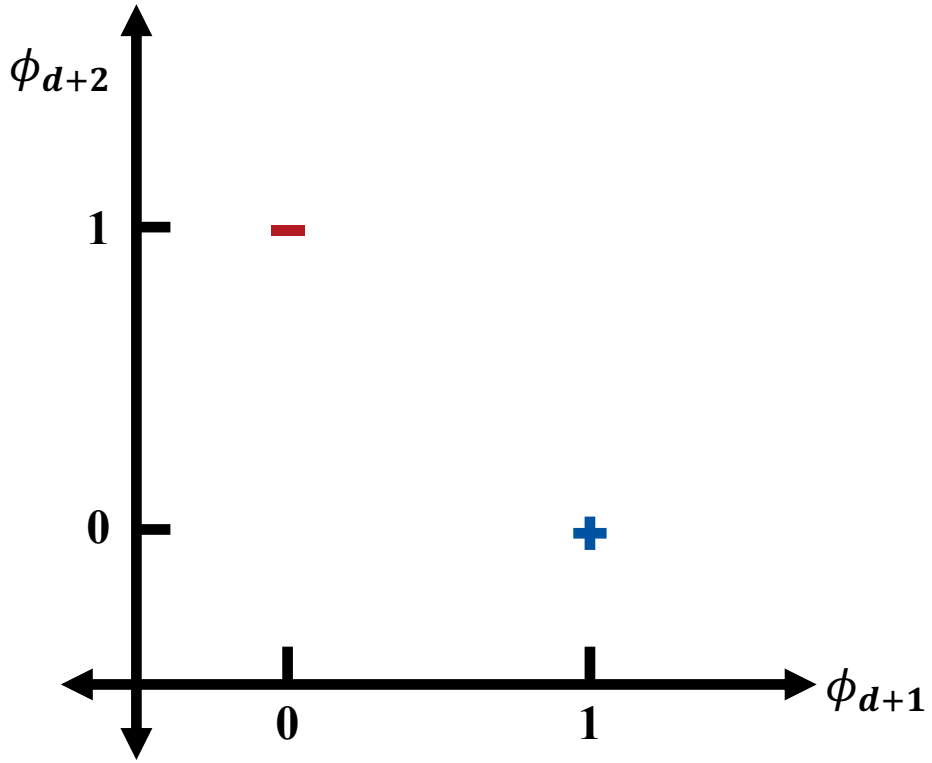


编码类别数据

- Idea 3：将每个类别转换成唯一的0-1向量

	ϕ_d	ϕ_{d+1}	ϕ_{d+2}	ϕ_{d+3}	ϕ_{d+4}
护士	1	0	0	0	0
管理	0	1	0	0	0
医生	0	0	1	0	0
药剂师	0	0	0	1	0
护工	0	0	0	0	1

- 独热编码 one-hot encoding



数据编码

- 识别特征，并编码成实数

	心率	疼痛症状	职业	用药	年龄	收入
1	55	0	护士	止痛药	40+	123000
2	71	0	管理	降压药&止痛药	20+	134000
3	89	1	护士	降压药	50+	140000
4	67	0	医生	无	50+	130000

数据编码

- 识别特征，并编码成实数

	心率	疼痛症状	j1,j2,j3,j4,j5	用药	年龄	收入
1	55	0	1, 0, 0, 0, 0	止痛药	40+	123000
2	71	0	0, 1, 0, 0, 0	降压药&止痛药	20+	134000
3	89	1	0, 0, 1, 0, 0	降压药	50+	140000
4	67	0	0, 0, 0, 1, 0	无	50+	130000

数据编码

- 识别特征，并编码成实数

	心率	疼痛症状	j1,j2,j3,j4,j5	用药	年龄	收入
1	55	0	1, 0, 0, 0, 0	止痛药	40+	123000
2	71	0	0, 1, 0, 0, 0	降压药&止痛药	20+	134000
3	89	1	0, 0, 1, 0, 0	降压药	50+	140000
4	67	0	0, 0, 0, 1, 0	无	50+	130000

编码类别数据

- 是否继续使用one-hot编码？

止痛药

止痛药 & 降压药

降压药

无

编码类别数据

- 是否继续使用one-hot编码？

	ϕ_d	ϕ_{d+1}	ϕ_{d+2}	ϕ_{d+3}
止痛药	**1**	**0**	**0**	**0**
止痛药 & 降压药	**0**	**1**	**0**	**0**
降压药	**0**	**0**	**1**	**0**
无	**0**	**0**	**0**	**1**

编码类别数据

- 是否继续使用one-hot编码？

	ϕ_d	ϕ_{d+1}	ϕ_{d+2}	ϕ_{d+3}
止痛药	**1**	**0**	**0**	**0**
止痛药 & 降压药	**0**	**1**	**0**	**0**
降压药	**0**	**0**	**1**	**0**
无	**0**	**0**	**0**	**1**

- Idea：分解编码

	ϕ_d	ϕ_{d+1}
止痛药	**1**	**0**
止痛药 & 降压药	**1**	**1**
降压药	**0**	**1**
无	**0**	**0**

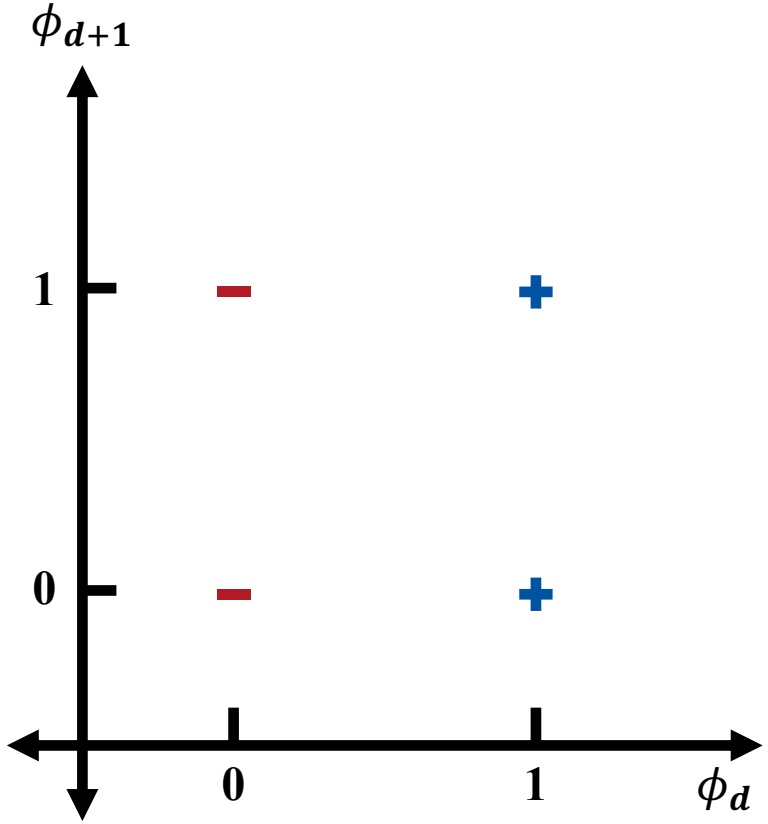
编码类别数据

- 是否继续使用one-hot编码？

	ϕ_d	ϕ_{d+1}	ϕ_{d+2}	ϕ_{d+3}
止痛药	**1**	**0**	**0**	**0**
止痛药 & 降压药	**0**	**1**	**0**	**0**
降压药	**0**	**0**	**1**	**0**
无	**0**	**0**	**0**	**1**

- Idea：分解编码

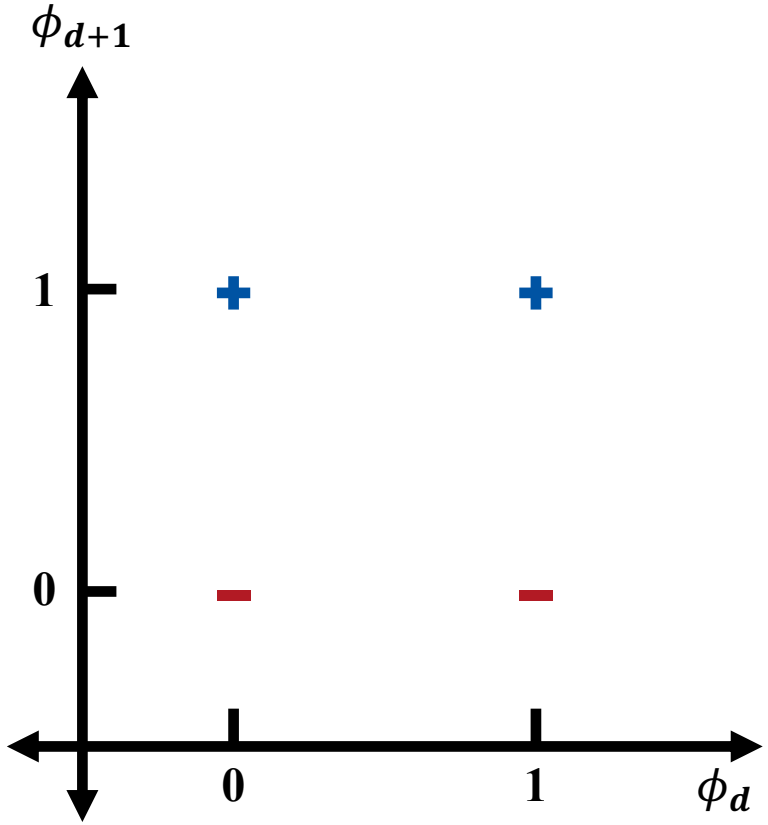
	ϕ_d	ϕ_{d+1}
止痛药	**1**	**0**
止痛药 & 降压药	**1**	**1**
降压药	**0**	**1**
无	**0**	**0**



编码类别数据

- 是否继续使用one-hot编码？

	ϕ_d	ϕ_{d+1}	ϕ_{d+2}	ϕ_{d+3}
止痛药	1	0	0	0
止痛药 & 降压药	0	1	0	0
降压药	0	0	1	0
无	0	0	0	1



- Idea：分解编码

	ϕ_d	ϕ_{d+1}
止痛药	1	0
止痛药 & 降压药	1	1
降压药	0	1
无	0	0

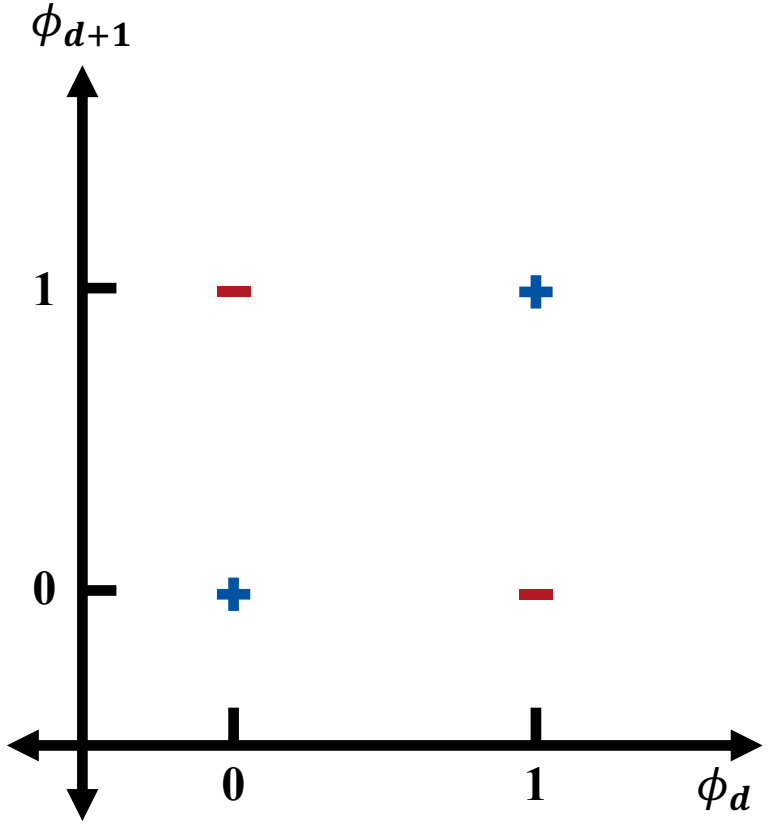
编码类别数据

- 是否继续使用one-hot编码？

	ϕ_d	ϕ_{d+1}	ϕ_{d+2}	ϕ_{d+3}
止痛药	**1**	**0**	**0**	**0**
止痛药 & 降压药	**0**	**1**	**0**	**0**
降压药	**0**	**0**	**1**	**0**
无	**0**	**0**	**0**	**1**

- Idea：分解编码

	ϕ_d	ϕ_{d+1}
止痛药	**1**	**0**
止痛药 & 降压药	**1**	**1**
降压药	**0**	**1**
无	**0**	**0**



数据编码

- 识别特征，并编码成实数

	心率	疼痛症状	j1,j2,j3,j4,j5	用药	年龄	收入
1	55	0	1, 0, 0, 0, 0	止痛药	40+	123000
2	71	0	0, 1, 0, 0, 0	降压药&止痛药	20+	134000
3	89	1	0, 0, 1, 0, 0	降压药	50+	140000
4	67	0	0, 0, 0, 1, 0	无	50+	130000

数据编码

- 识别特征，并编码成实数

	心率	疼痛症状	j1,j2,j3,j4,j5	m1,m2	年龄	收入
1	55	0	1, 0, 0, 0, 0	1, 0	40+	123000
2	71	0	0, 1, 0, 0, 0	1, 1	20+	134000
3	89	1	0, 0, 1, 0, 0	0, 1	50+	140000
4	67	0	0, 0, 0, 1, 0	0, 0	50+	130000

数据编码

- 识别特征，并编码成实数

	心率	疼痛症状	j1,j2,j3,j4,j5	m1,m2	年龄	收入
1	55	0	1, 0, 0, 0, 0	1, 0	40+	123000
2	71	0	0, 1, 0, 0, 0	1, 1	20+	134000
3	89	1	0, 0, 1, 0, 0	0, 1	50+	140000
4	67	0	0, 0, 0, 1, 0	0, 0	50+	130000

数据编码

- 识别特征，并编码成实数
- 使用代表性数字（均值、中位数）表示区间
- 缺点：忽略了数据中的细节

	心率	疼痛症状	j1,j2,j3,j4,j5	m1,m2	年龄	收入
1	55	0	1, 0, 0, 0, 0	1, 0	45	123000
2	71	0	0, 1, 0, 0, 0	1, 1	25	134000
3	89	1	0, 0, 1, 0, 0	0, 1	55	140000
4	67	0	0, 0, 0, 1, 0	0, 0	55	130000

数据编码

- 识别特征，并编码成实数

	心率	疼痛症状	j1,j2,j3,j4,j5	m1,m2	d	收入
1	55	0	1, 0, 0, 0, 0	1, 0	4	123000
2	71	0	0, 1, 0, 0, 0	1, 1	2	134000
3	89	1	0, 0, 1, 0, 0	0, 1	5	140000
4	67	0	0, 0, 0, 1, 0	0, 0	5	130000

数据编码

- 识别特征，并编码成实数

	心率	疼痛症状	j1,j2,j3,j4,j5	m1,m2	d	收入
1	55	0	1, 0, 0, 0, 0	1, 0	4	123000
2	71	0	0, 1, 0, 0, 0	1, 1	2	134000
3	89	1	0, 0, 1, 0, 0	0, 1	5	140000
4	67	0	0, 0, 0, 1, 0	0, 0	5	130000

数据编码

- 数值数据：按数据值的顺序排列，数据值的差异在某些任务中可能有意义
- 类别数据：无需考虑顺序
- 序列数据：按数据值的顺序排列，但数据值差异没有意义

数据编码

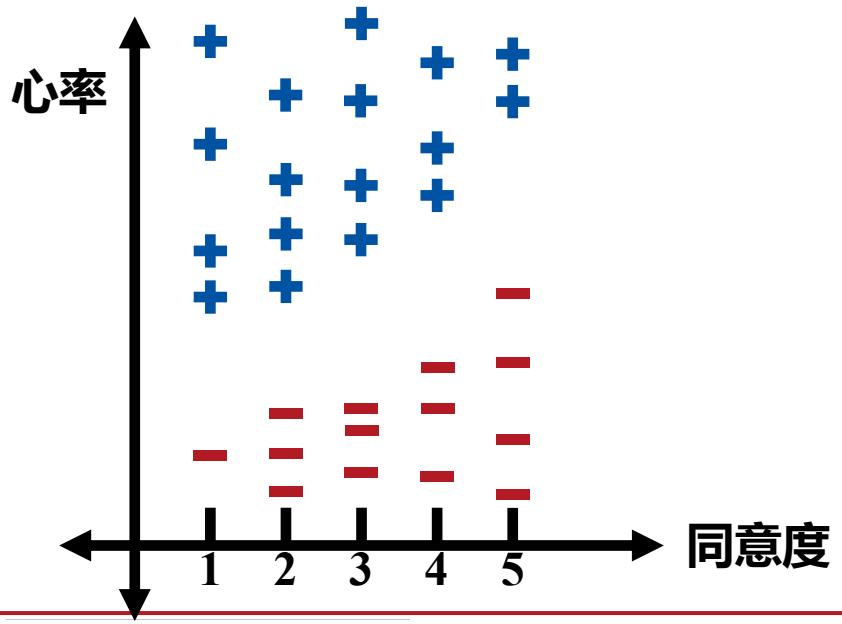
- 数值数据：按数据值的顺序排列，数据值的差异在某些任务中可能有意义
- 类别数据：无需考虑顺序
- 序列数据：按数据值的顺序排列，但数据值差异没有意义
  - E. g. 李克特量表：

强烈否定	否定	中立	同意	强烈同意
1	2	3	4	5

数据编码

- 数值数据：按数据值的顺序排列，数据值的差异在某些任务中可能有意义
- 类别数据：无需考虑顺序
- 序列数据：按数据值的顺序排列，但数据值差异没有意义
  - E. g. 李克特量表：

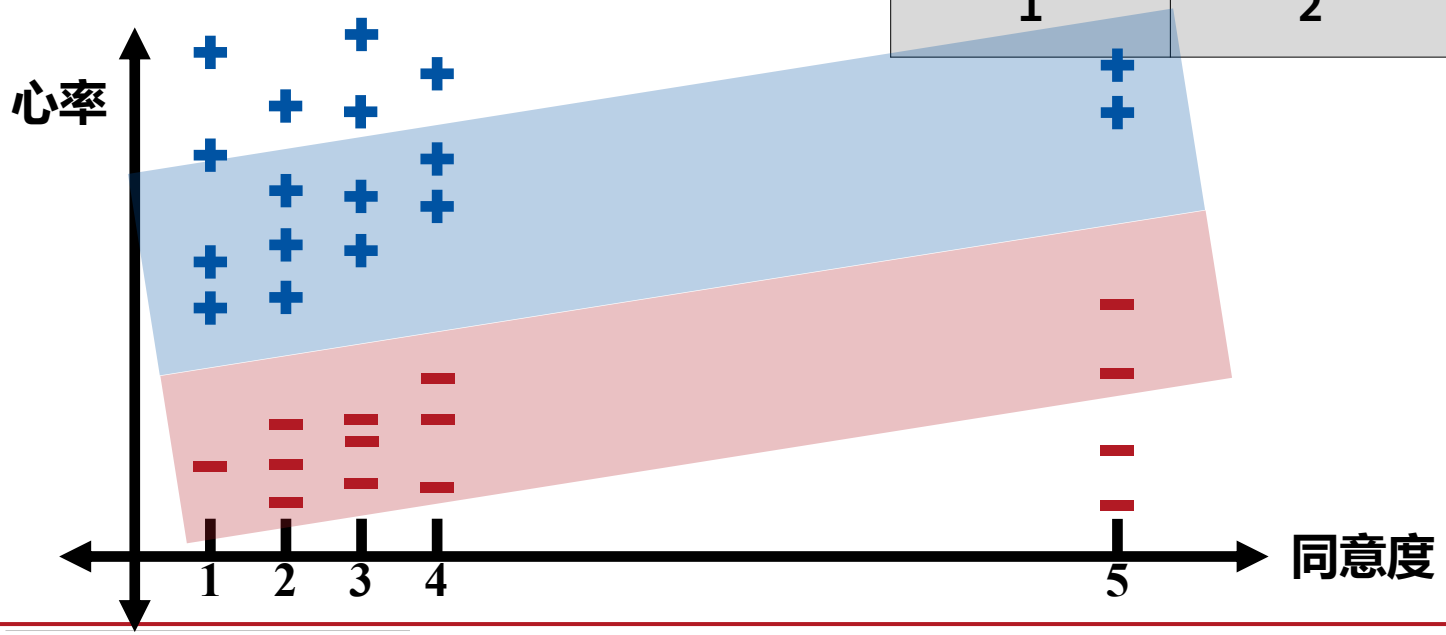
强烈否定	否定	中立	同意	强烈同意
1	2	3	4	5



数据编码

- 数值数据：按数据值的顺序排列，数据值的差异在某些任务中可能有意义
- 类别数据：无需考虑顺序
- 序列数据：按数据值的顺序排列，但数据值差异没有意义
  - E. g. 李克特量表：

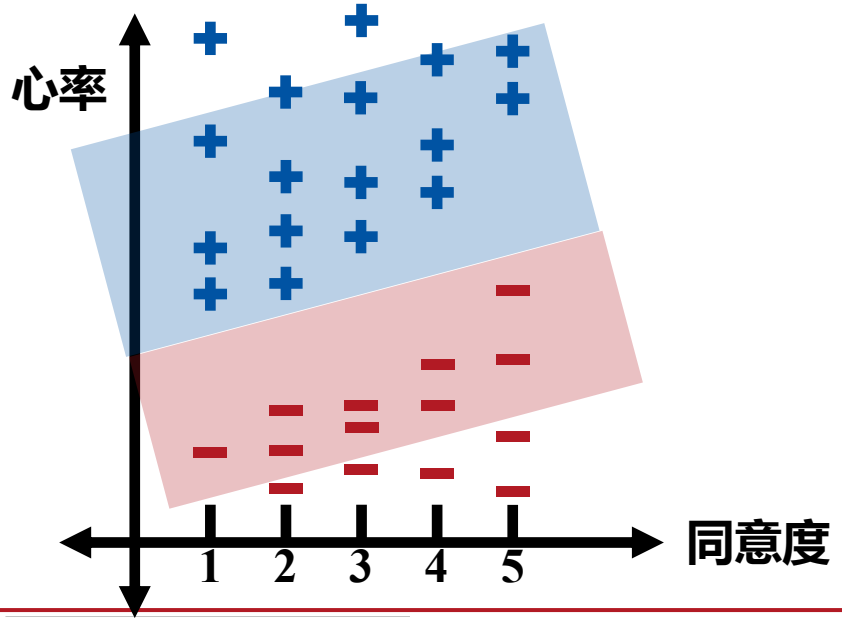
强烈否定	否定	中立	同意	强烈同意
1	2	3	4	5

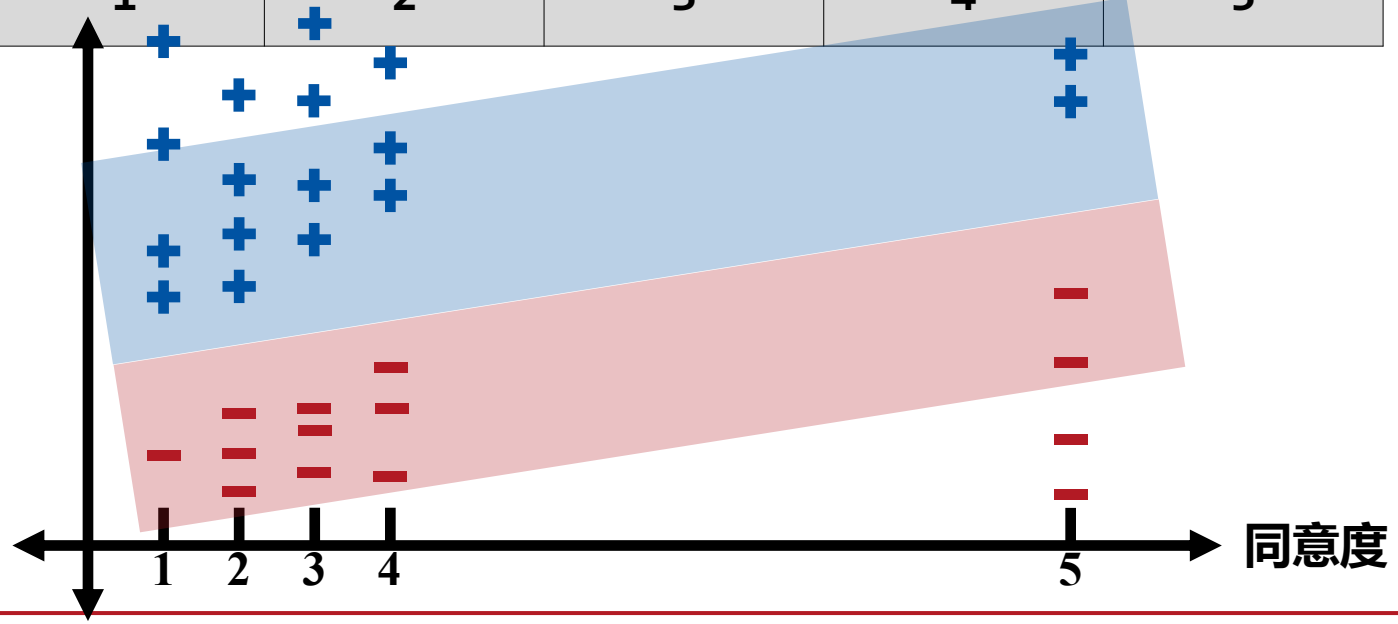


数据编码

- 数值数据：按数据值的顺序排列，数据值的差异在某些任务中可能有意义
- 类别数据：无需考虑顺序
- 序列数据：按数据值的顺序排列，但数据值差异没有意义
  - E. g. 李克特量表：

强烈否定	否定	中立	同意	强烈同意
1	2	3	4	5

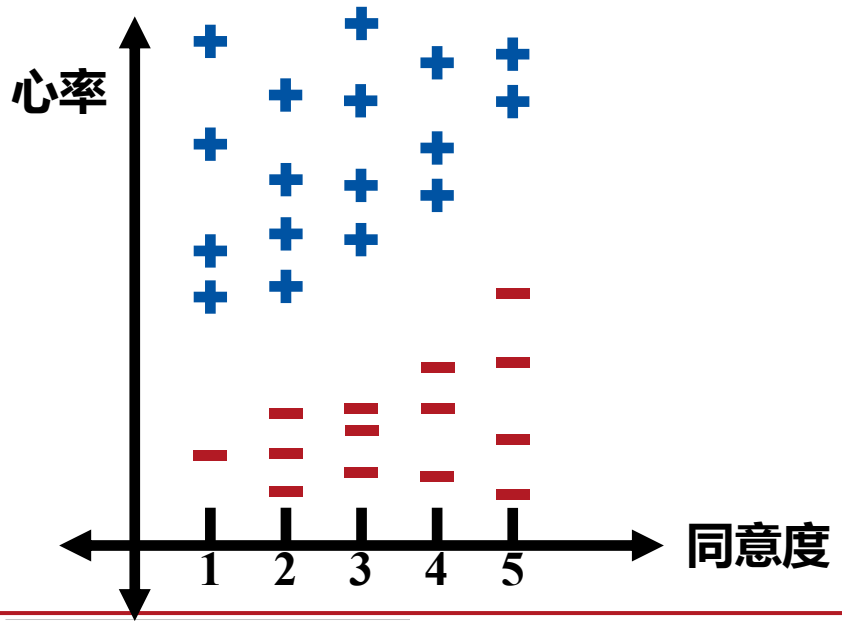




数据编码

- 数值数据：按数据值的顺序排列，数据值的差异在某些任务中可能有意义
- 类别数据：无需考虑顺序
- 序列数据：按数据值的顺序排列，但数据值差异没有意义
  - E. g. 李克特量表：

强烈否定	否定	中立	同意	强烈同意
1	2	3	4	5

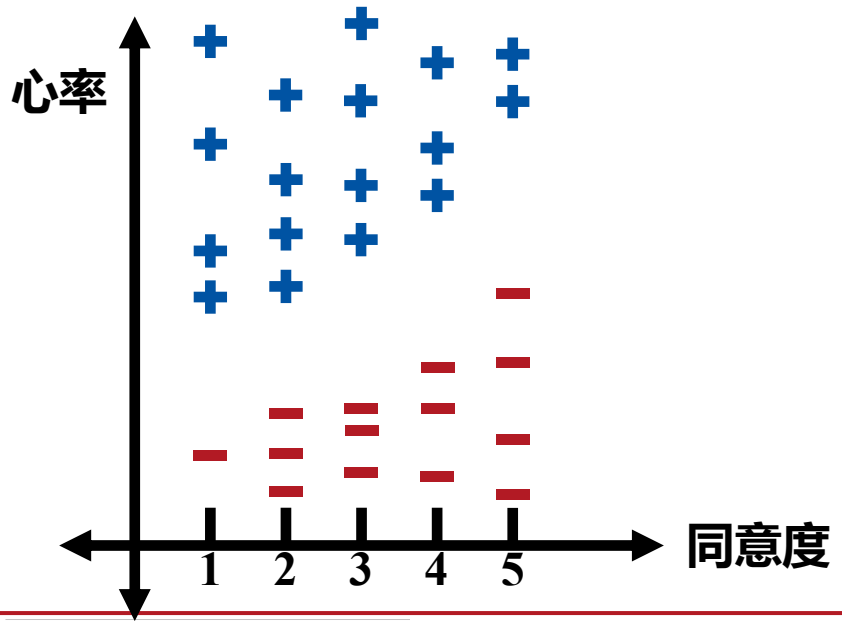


- Idea：一元码（温度计编码）

数据编码

- 数值数据：按数据值的顺序排列，数据值的差异在某些任务中可能有意义
- 类别数据：无需考虑顺序
- 序列数据：按数据值的顺序排列，但数据值差异没有意义
  - E. g. 李克特量表：

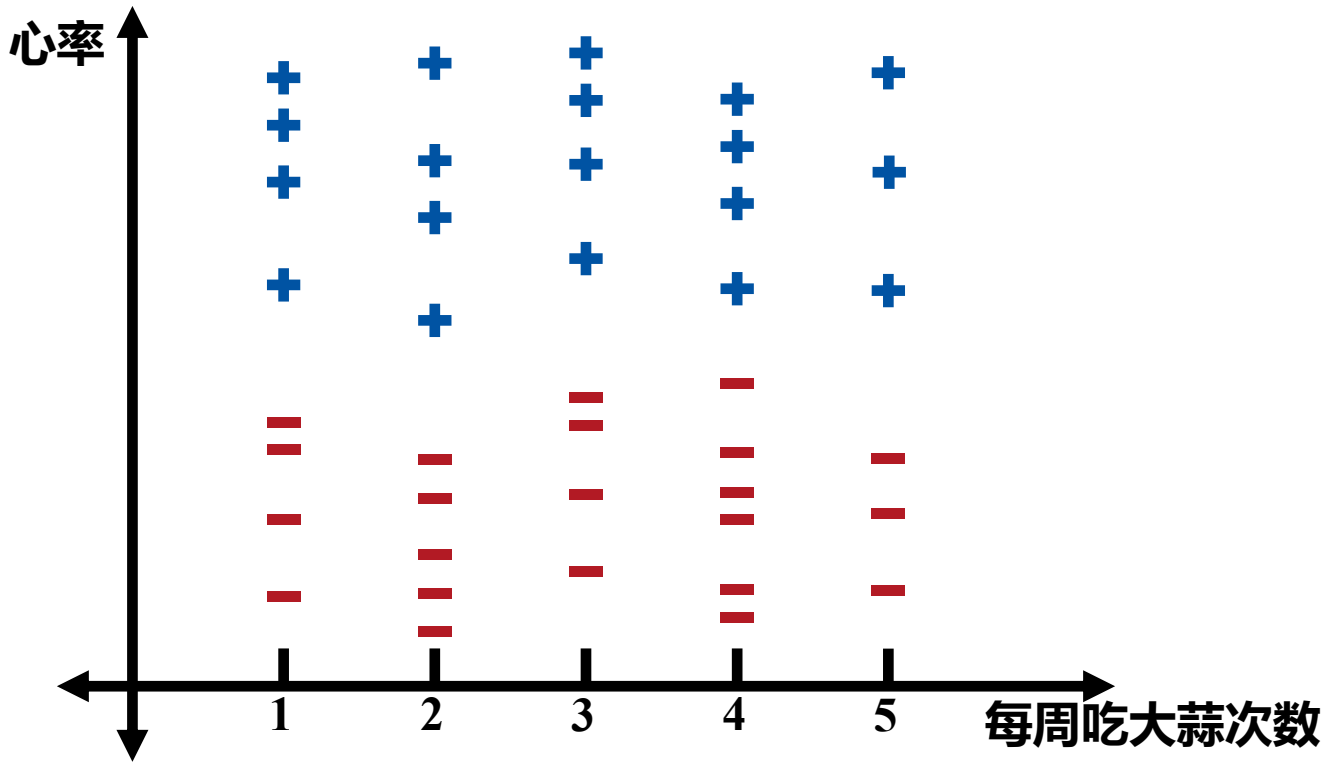
强烈否定	否定	中立	同意	强烈同意
1,0,0,0,0	1,1,0,0,0	1,1,1,0,0	1,1,1,1,0	1,1,1,1,1



- Idea：一元码（温度计编码）

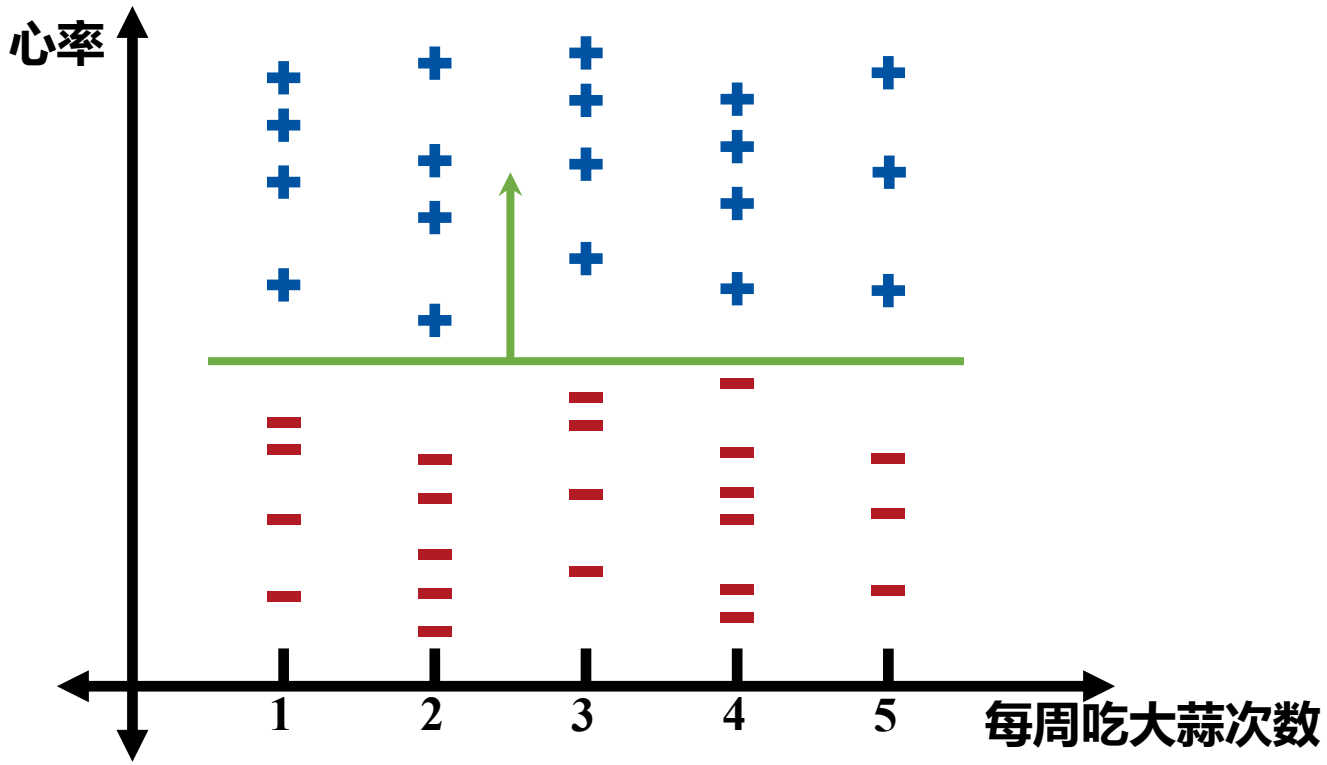
数据编码

- 观察线性分类器的输出



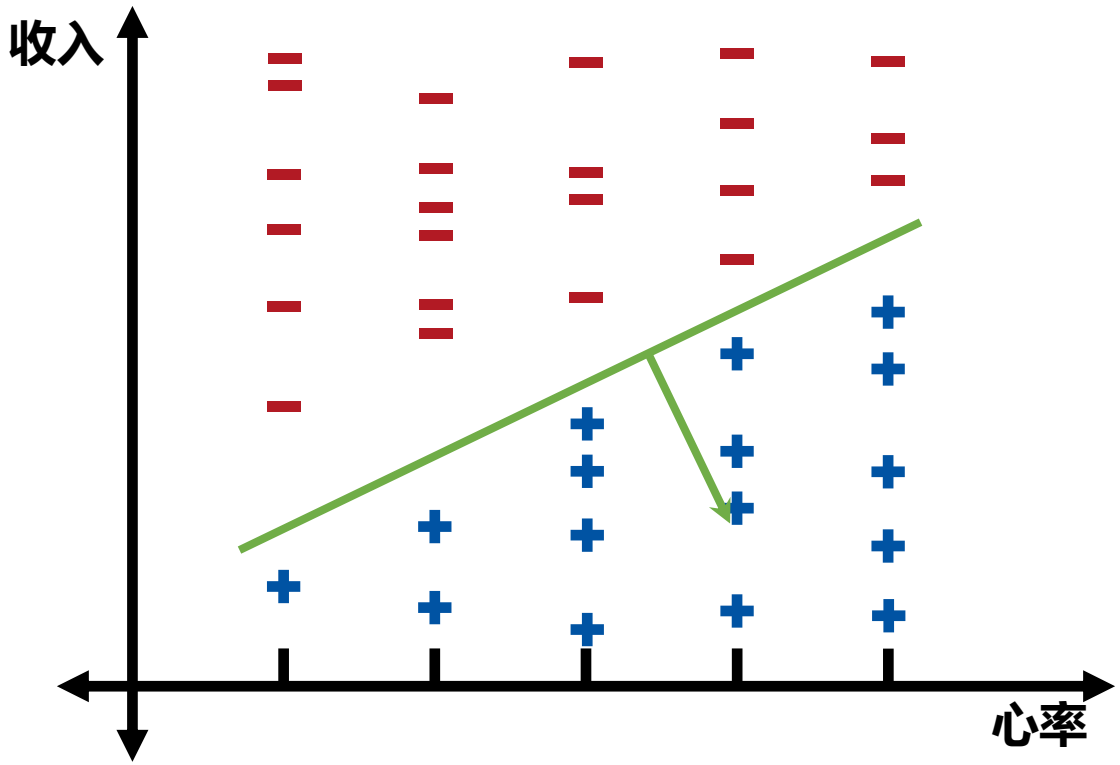
数据编码

- 观察线性分类器的输出



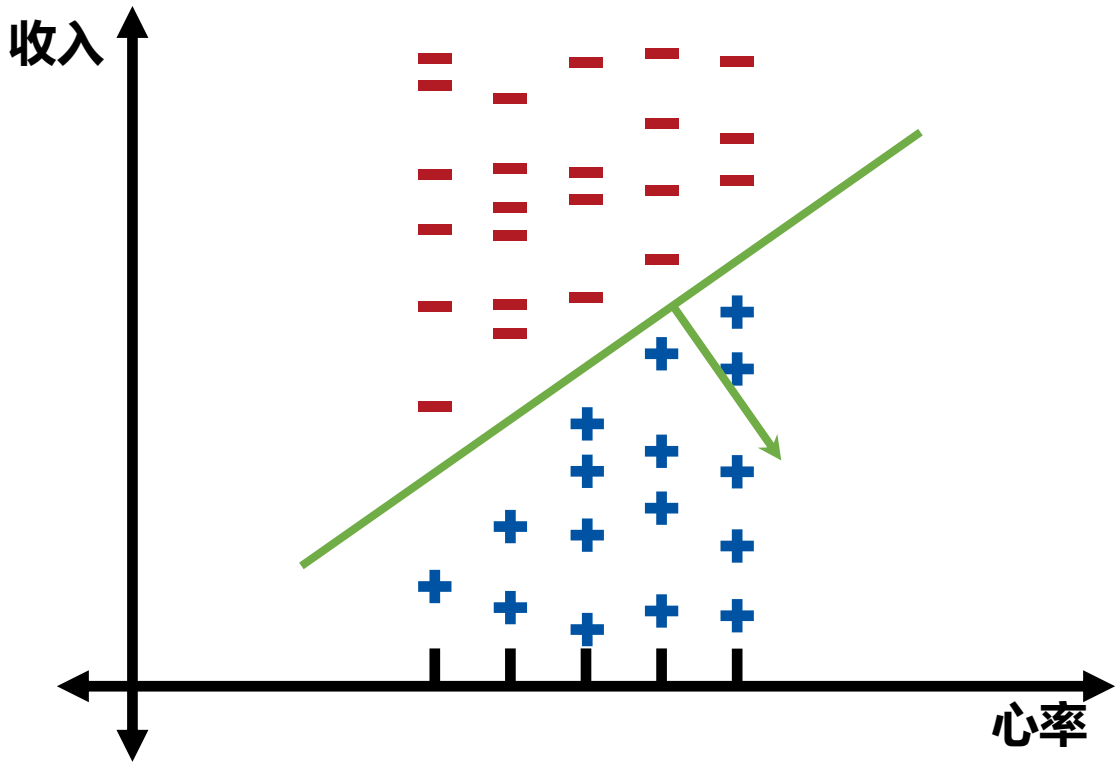
数据编码

- 观察线性分类器的输出



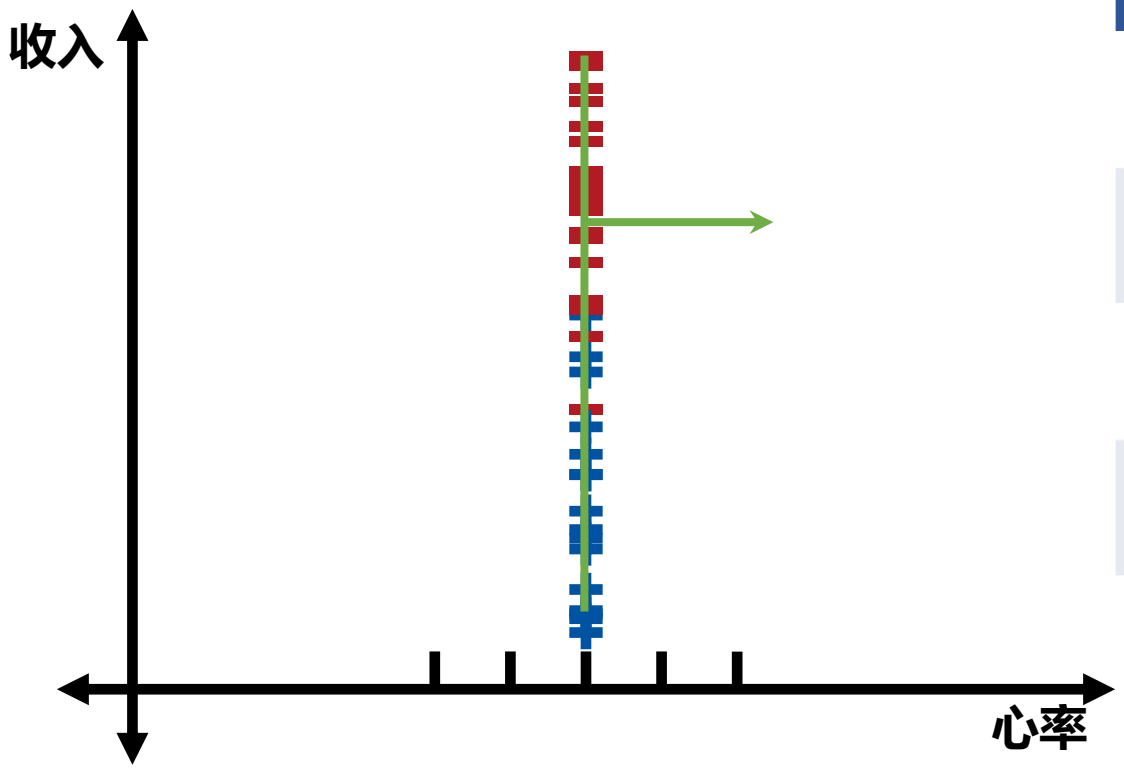
数据编码

- 观察线性分类器的输出



数据编码

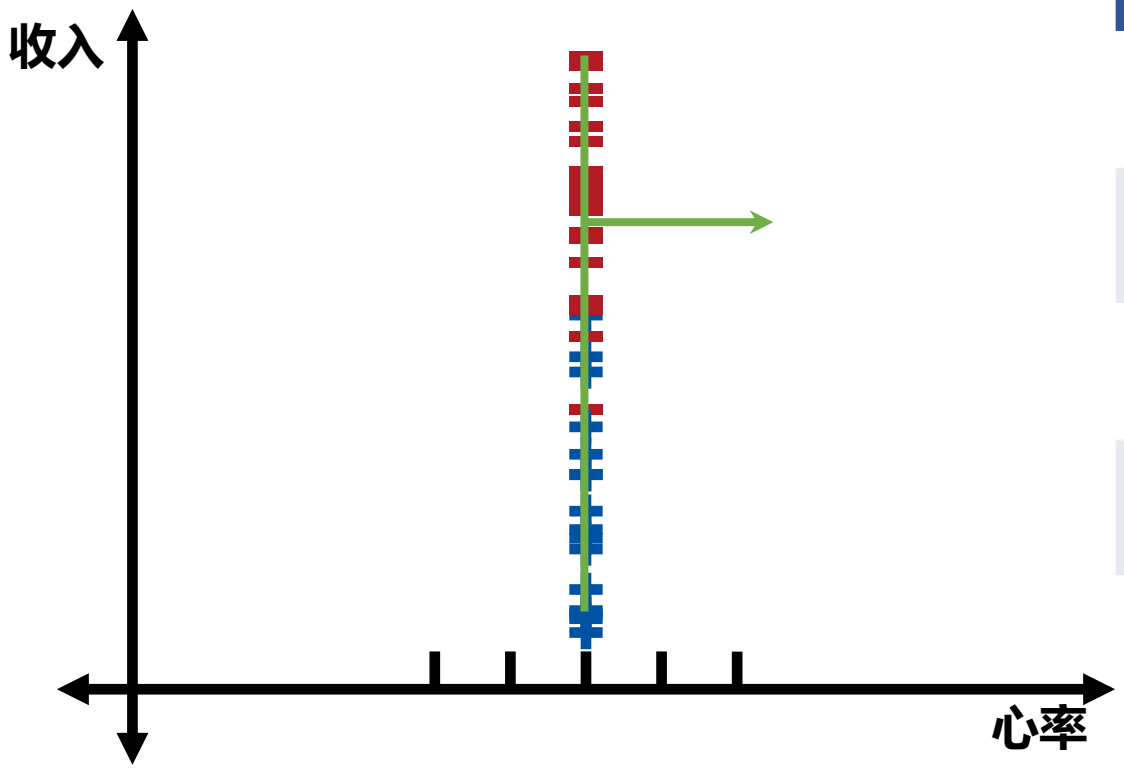
- 观察线性分类器的输出
- Idea：标准化数值数据



收入
123000
134000
140000
130000

数据编码

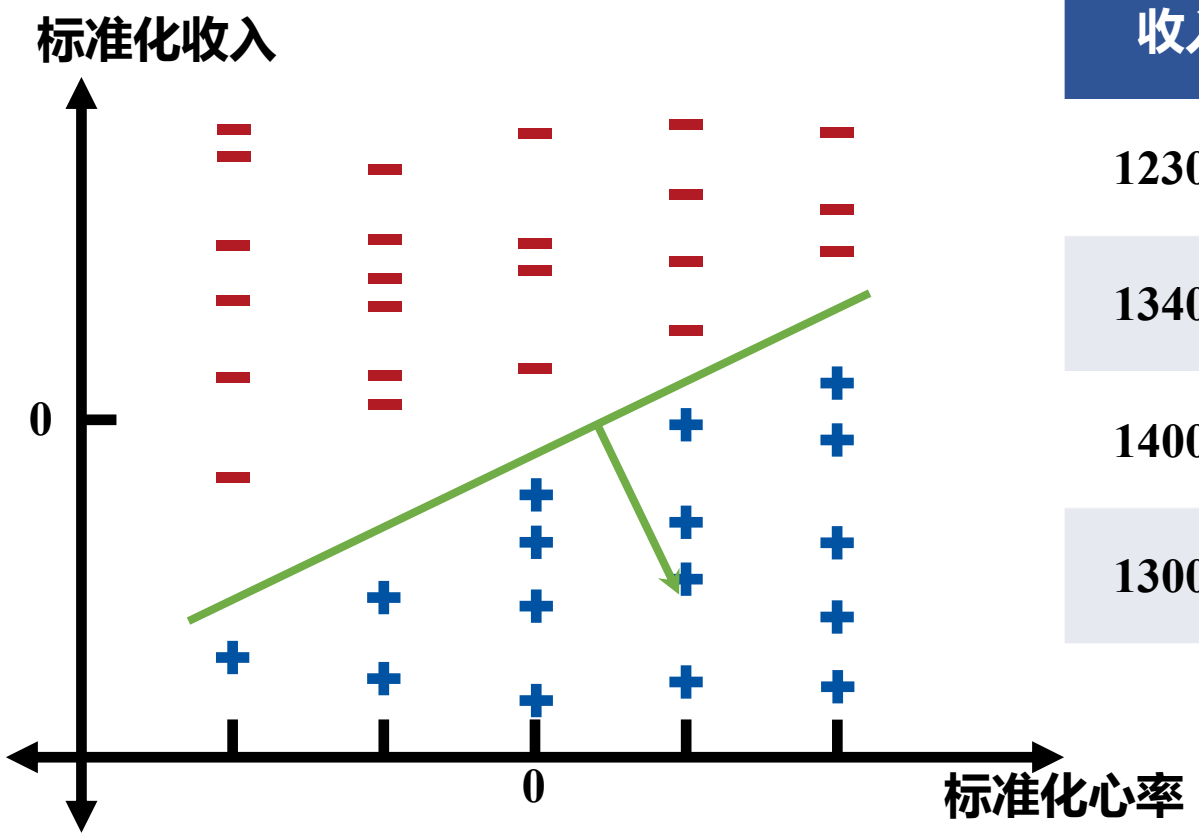
- 观察线性分类器的输出
- Idea：标准化数值数据
  - 对于d-th特征：$\phi_d^k = \frac{x_d^{(k)} - \text{mean}_d}{\text{stddev}_d}$



收入
123000
134000
140000
130000

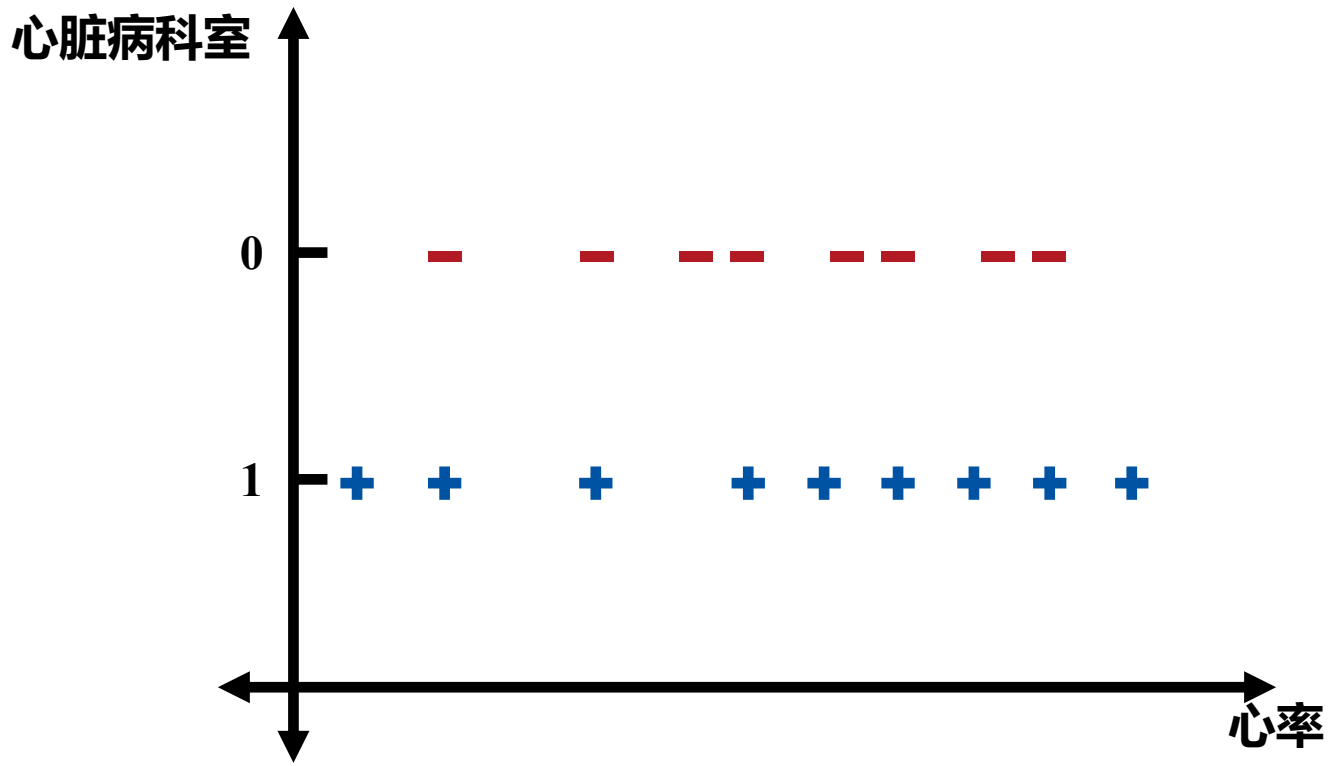
数据编码

- 观察线性分类器的输出
- Idea：标准化数值数据
  - 对于d-th特征：$\phi_d^k = \frac{x_d^{(k)} - \text{mean}_d}{\text{stddev}_d}$



收入
123000
134000
140000
130000

数据编码



数据编码

- 识别特征，并编码成实数
- 标准化数值数据

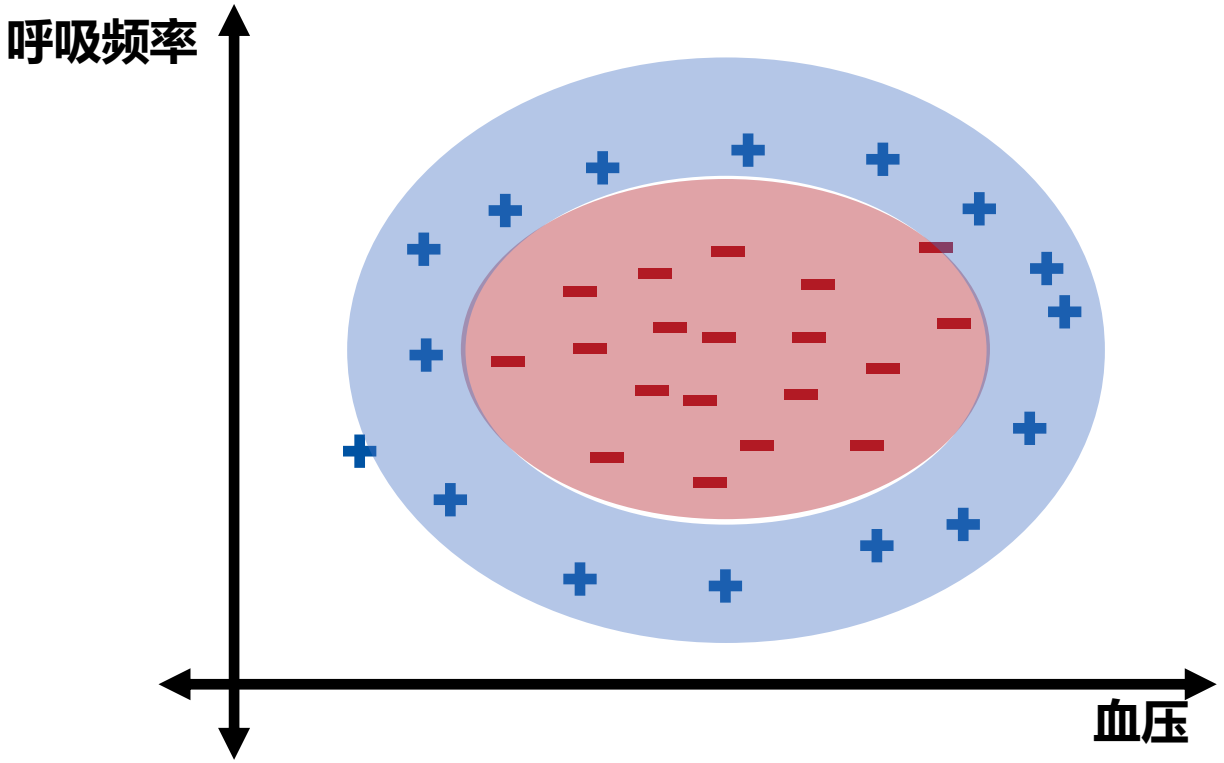
	心率	疼痛症状	j1,j2,j3,j4,j5	m1,m2	d	收入
1	55	0	1, 0, 0, 0, 0	1, 0	4	123000
2	71	0	0, 1, 0, 0, 0	1, 1	2	134000
3	89	1	0, 0, 1, 0, 0	0, 1	5	140000
4	67	0	0, 0, 0, 1, 0	0, 0	5	130000

数据编码

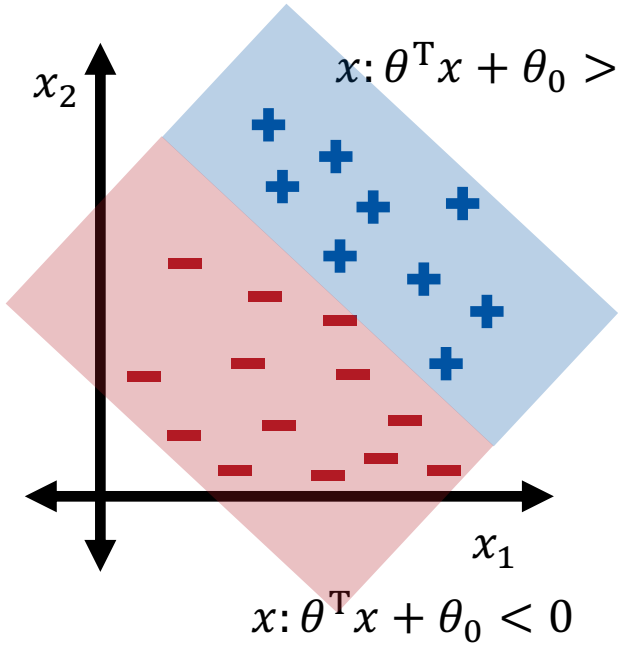
- 识别特征，并编码成实数
- 标准化数值数据

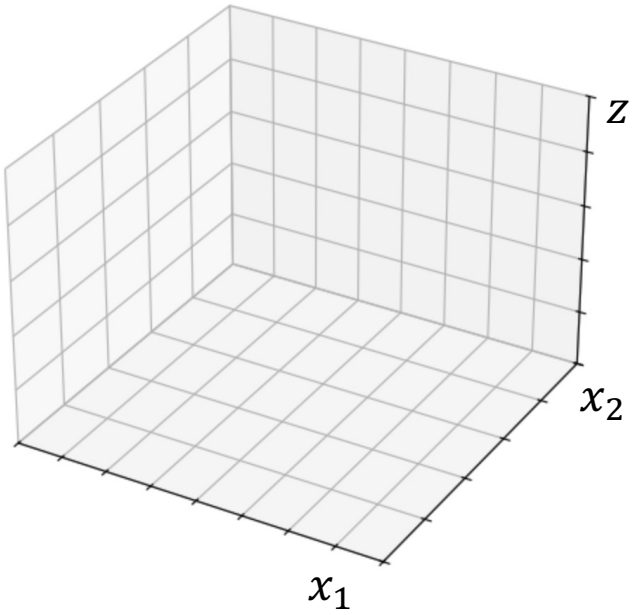
	心率	疼痛症状	j1,j2,j3,j4,j5	m1,m2	d	收入
1	-1.5	0	1, 0, 0, 0, 0	1, 0	1	2.075
2	0.1	0	0, 1, 0, 0, 0	1, 1	-1	-0.4
3	1.9	1	0, 0, 1, 0, 0	0, 1	2	-0.25
4	-0.3	0	0, 0, 0, 1, 0	0, 0	2	1.75

非线性边界

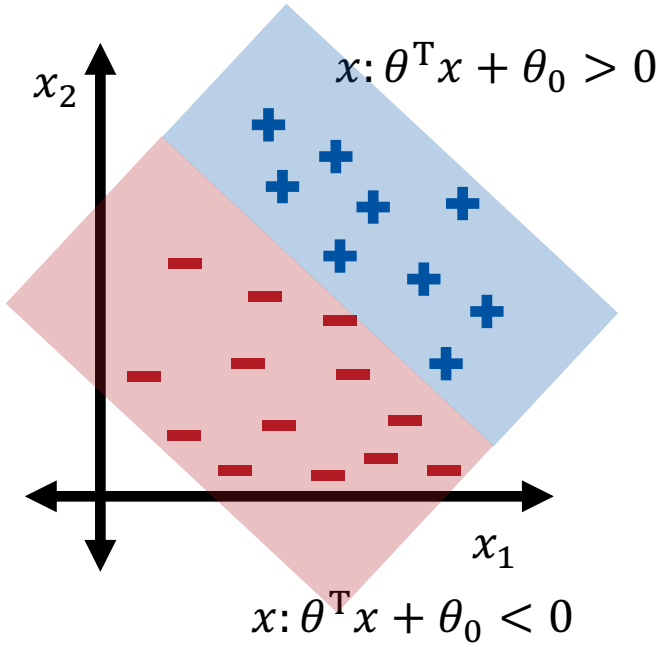


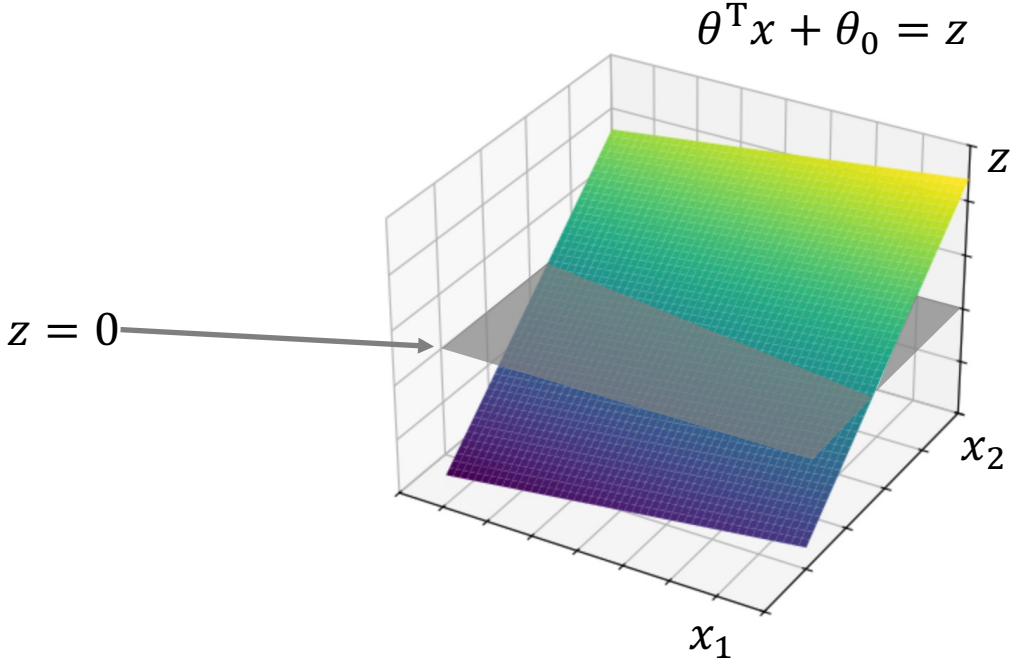
非线性边界


x_2
x_1
$x: \theta^{\mathrm{T}} x + \theta_0 > 0$
$x: \theta^{\mathrm{T}} x + \theta_0 < 0$

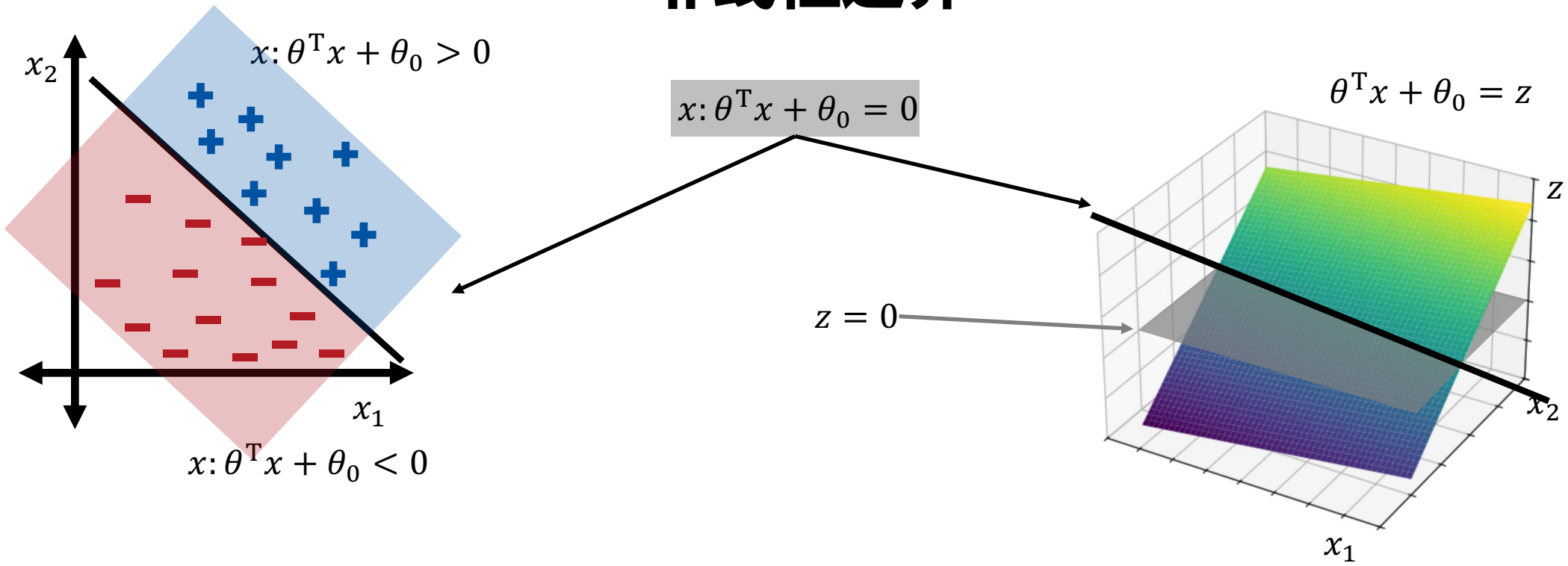

z
x_2
x_1

非线性边界

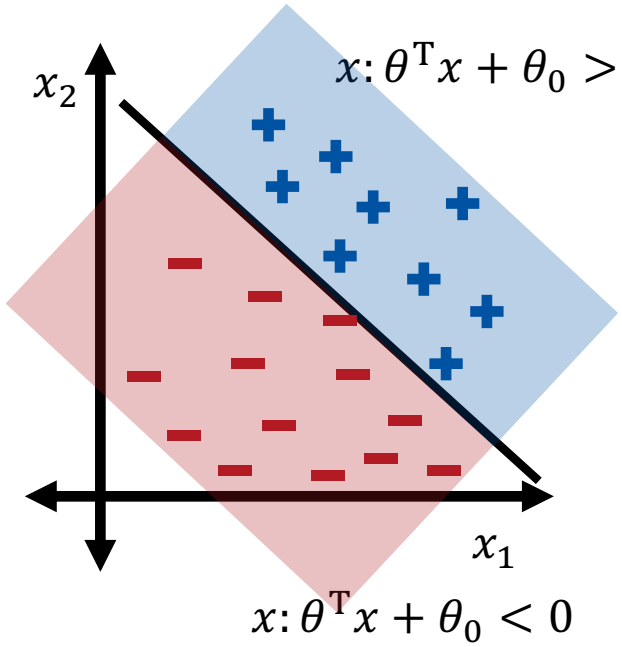

x_2
x_1
$x: \theta^{\mathrm{T}} x + \theta_0 > 0$
$x: \theta^{\mathrm{T}} x + \theta_0 < 0$


$\theta^{\mathrm{T}} x + \theta_0 = z$
$z = 0$
z
x_2
x_1

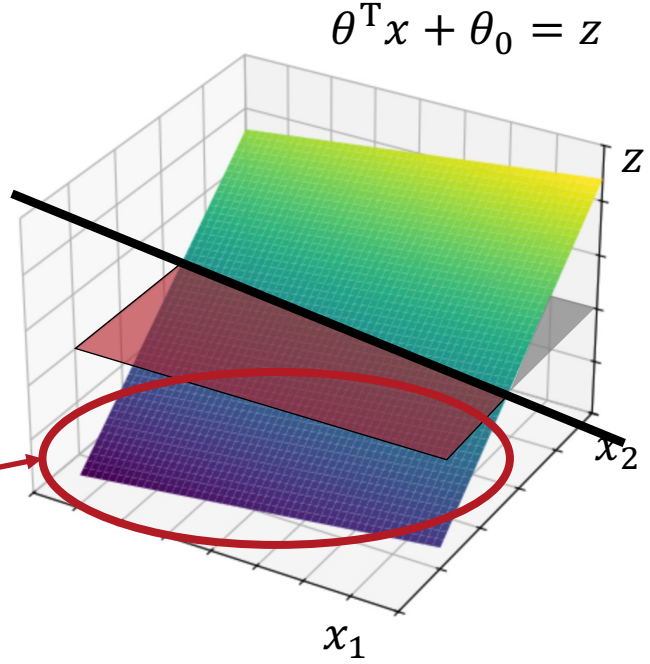
非线性边界


x_2
x_1
$x: \theta^{\mathrm{T}} x + \theta_0 > 0$
$x: \theta^{\mathrm{T}} x + \theta_0 < 0$
$x: \theta^{\mathrm{T}} x + \theta_0 = 0$
$z = 0$
$\theta^{\mathrm{T}} x + \theta_0 = z$
z
x_2
x_1

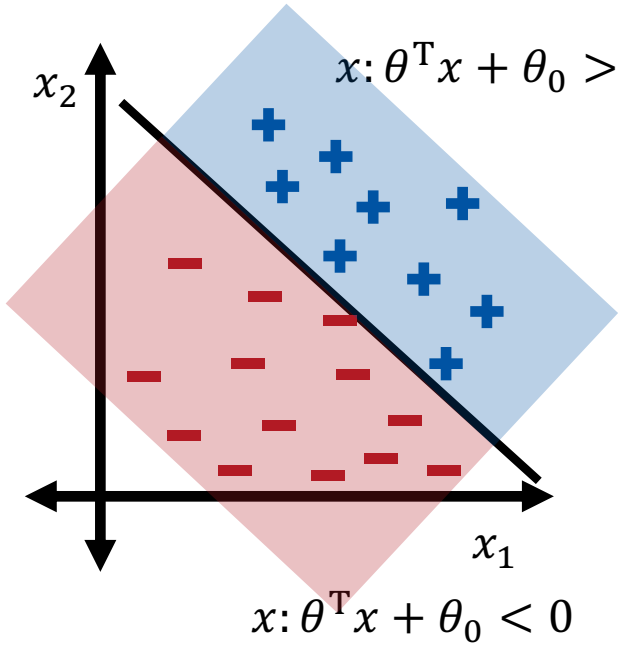
非线性边界



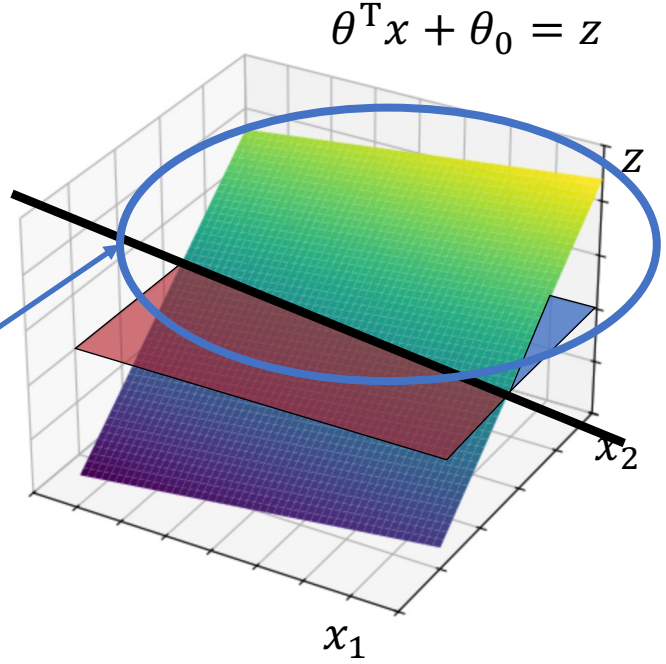
$x: \theta^{\mathrm{T}} x + \theta_0 < 0$



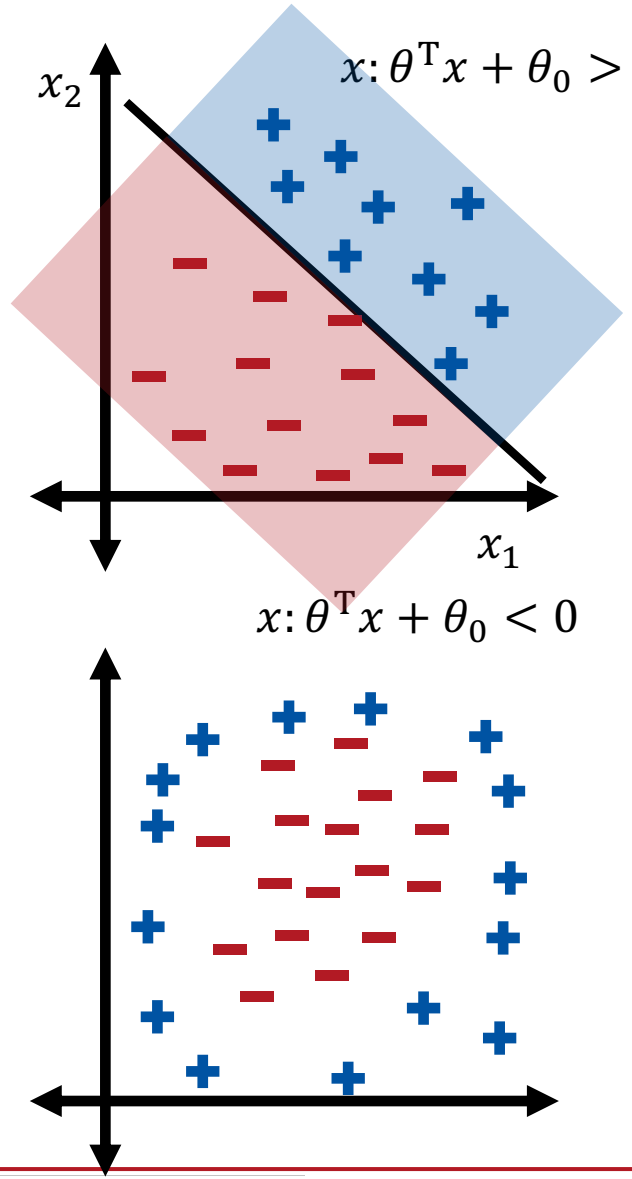
非线性边界

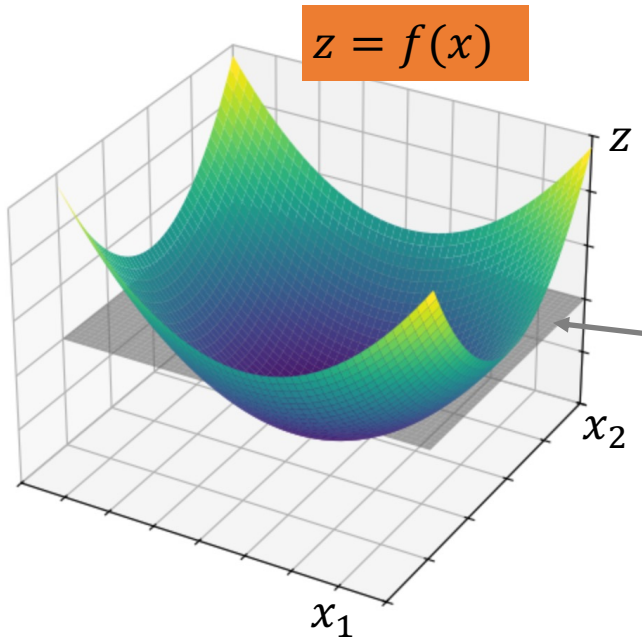


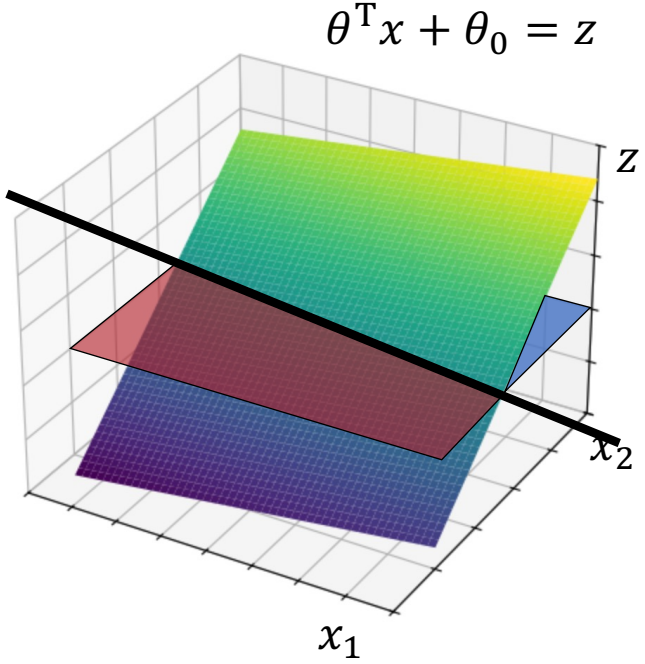
$x: \theta^{\mathrm{T}} x + \theta_0 > 0$



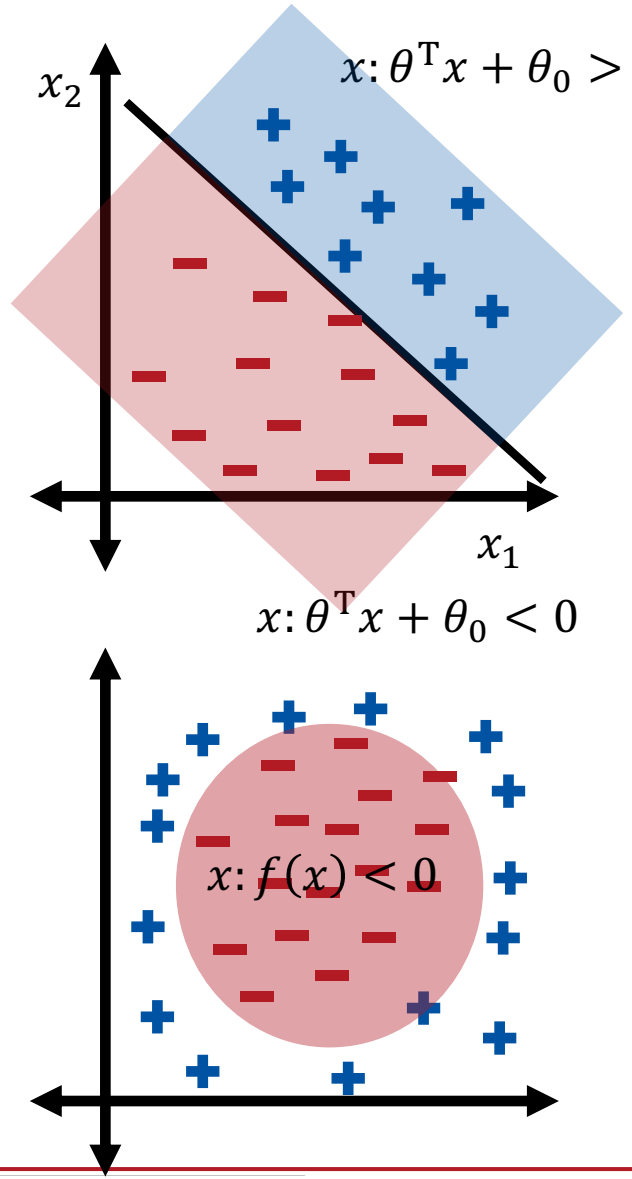
非线性边界

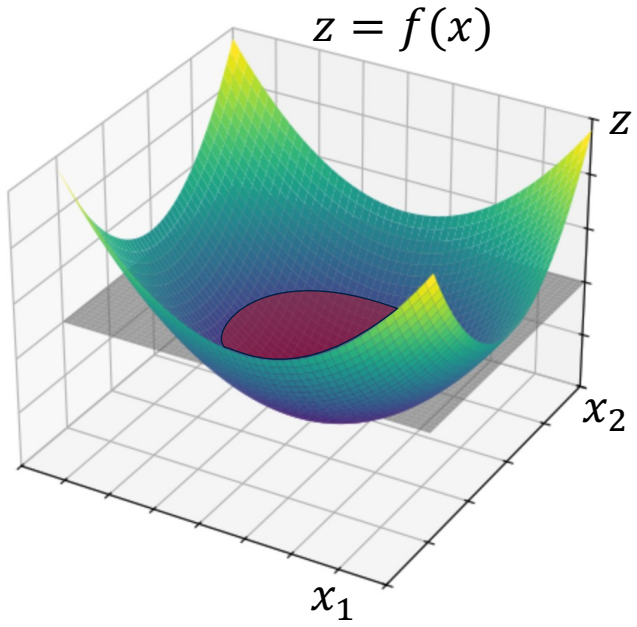

x_2
x_1
$x: \theta^{\mathrm{T}} x + \theta_0 > 0$
$x: \theta^{\mathrm{T}} x + \theta_0 < 0$

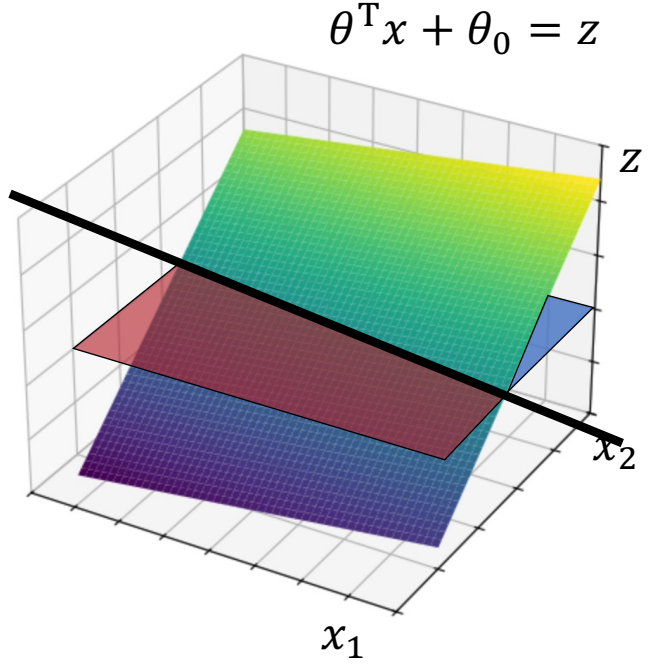

$z = f(x)$
z
x_1
x_2
$z = 0$


$\theta^{\mathrm{T}} x + \theta_0 = z$
z
x_1
x_2

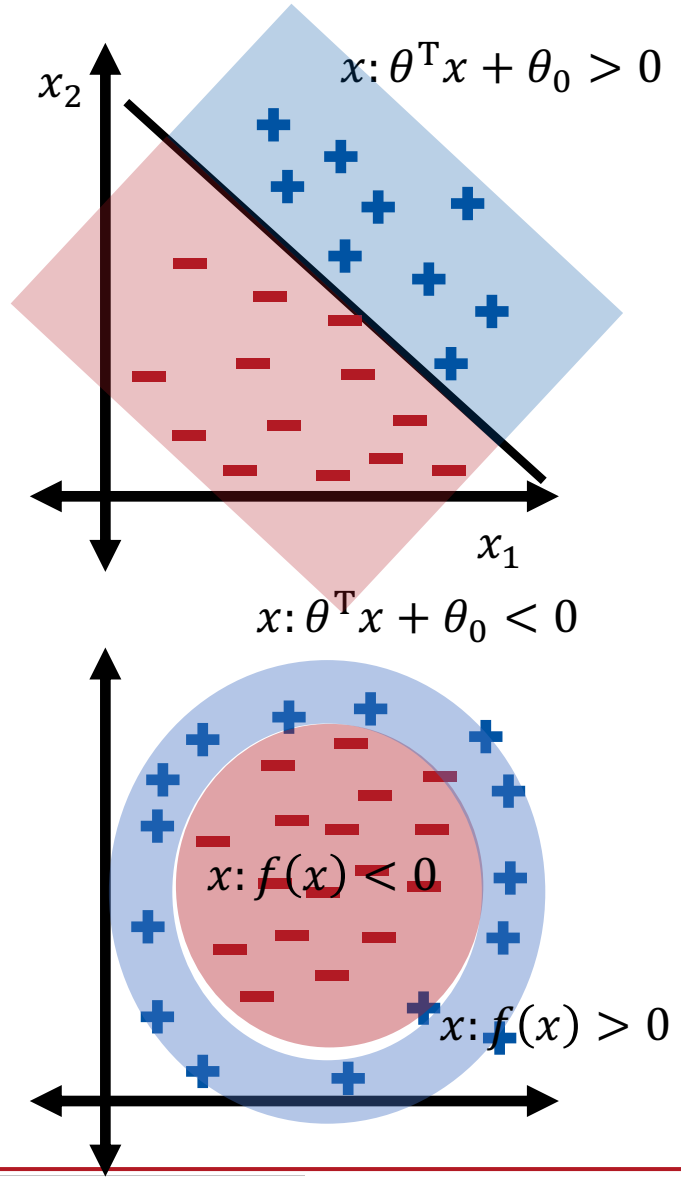
非线性边界

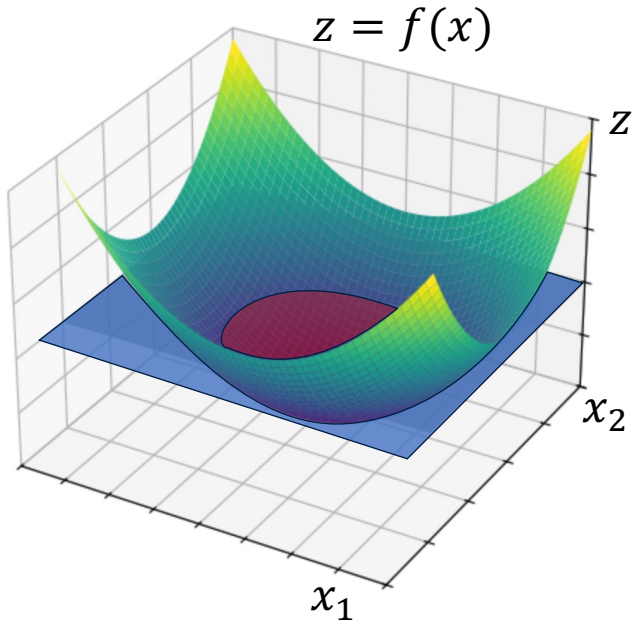

x_2
x_1
$x: \theta^{\mathrm{T}} x + \theta_0 > 0$
$x: \theta^{\mathrm{T}} x + \theta_0 < 0$
$x: f(x) < 0$

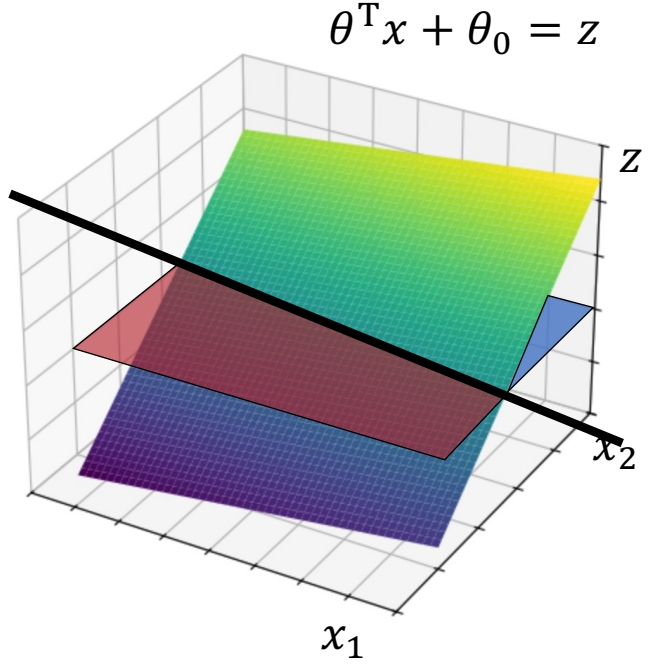

$z = f(x)$
z
x_2
x_1


$\theta^{\mathrm{T}} x + \theta_0 = z$
z
x_2
x_1

非线性边界


x_2
x_1
$x: \theta^{\mathrm{T}} x + \theta_0 > 0$
$x: \theta^{\mathrm{T}} x + \theta_0 < 0$
$x: f(x) < 0$
$x: f(x) > 0$


$z = f(x)$
z
x_2
x_1


$\theta^{\mathrm{T}} x + \theta_0 = z$
z
x_2
x_1

非线性边界

- Idea：k阶泰勒多项式近似非线性边界

阶	d=1时的项	一般情况下的项
0	$[1]$	
1	$[1, x_1]$	
2		
3		

非线性边界

- Idea：k阶泰勒多项式近似非线性边界

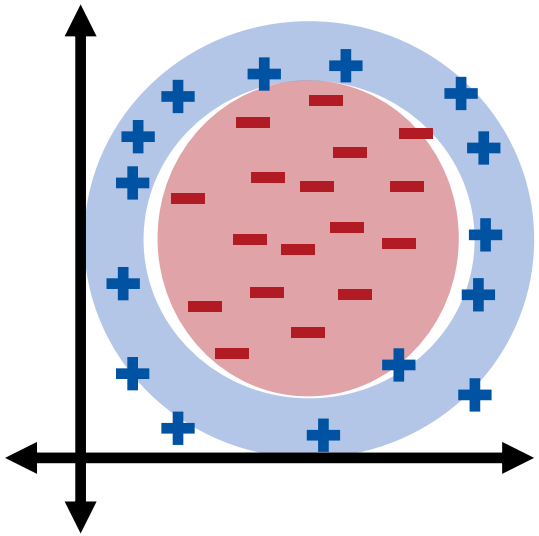
阶	d=1时的项	一般情况下的项
0	$[1]$	
1	$[1, x_1]$	
2	$[1, x_1, x_1^2]$	
3	$[1, x_1, x_1^2, x_1^3]$	

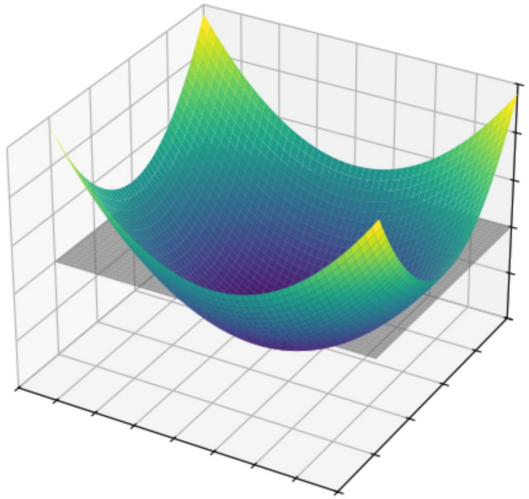
非线性边界

- Idea：k阶泰勒多项式近似非线性边界

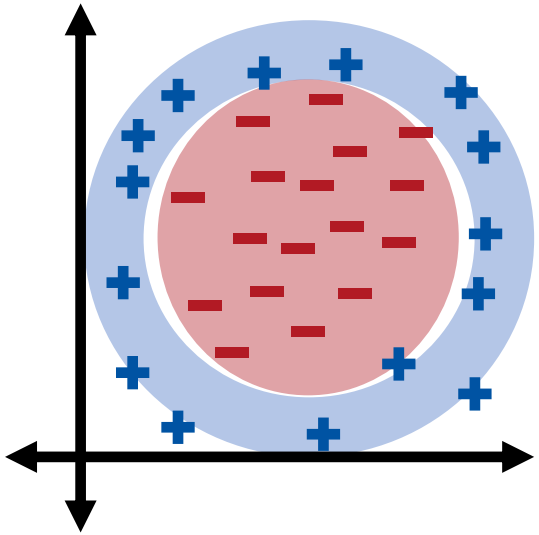
阶	d=1时的项	一般情况下的项
0	$[1]$	$[1]$
1	$[1, x_1]$	$[1, x_1, \dots, x_d]$
2	$[1, x_1, x_1^2]$	$[1, x_1, \dots, x_d,$ $x_1^2, x_1x_2, \dots, x_{d-1}x_d, x_d^2]$
3	$[1, x_1, x_1^2, x_1^3]$	$[1, x_1, \dots, x_d,$ $x_1^2, x_1x_2, \dots, x_{d-1}x_d, x_d^2,$ $x_1^3, x_1^2x_2, x_1x_2x_3, \dots, x_d^3]$

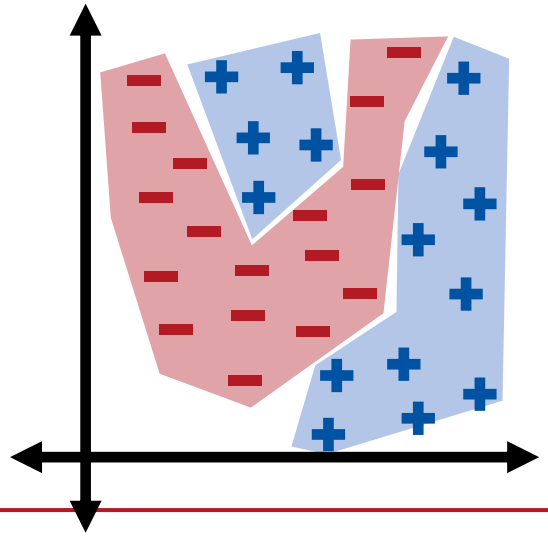
非线性边界

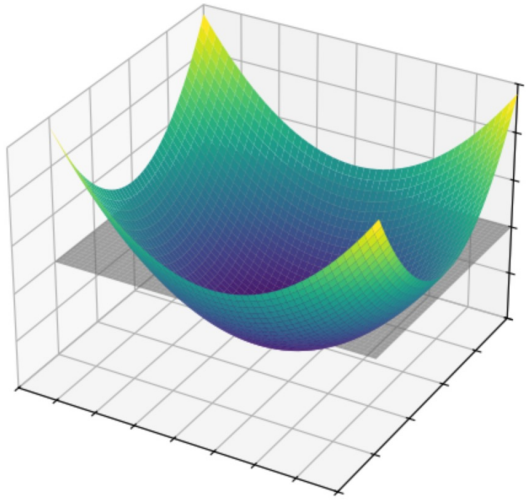


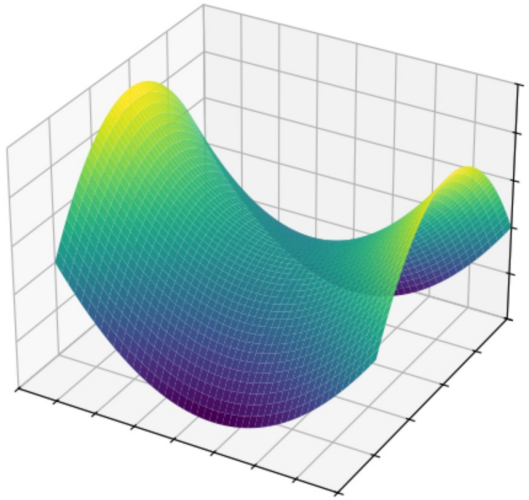


非线性边界

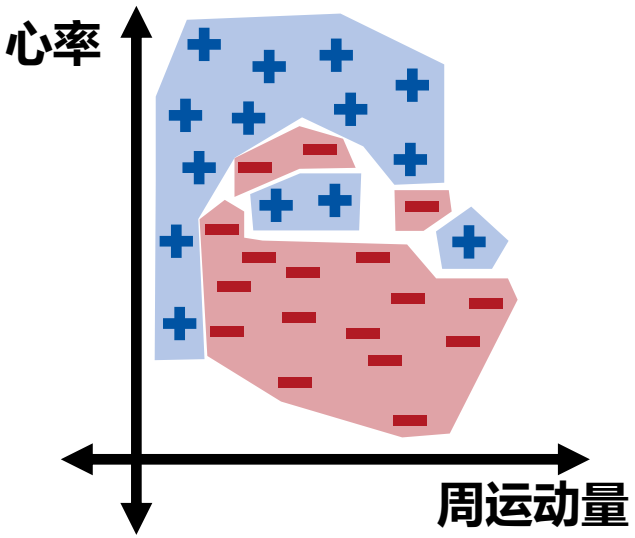






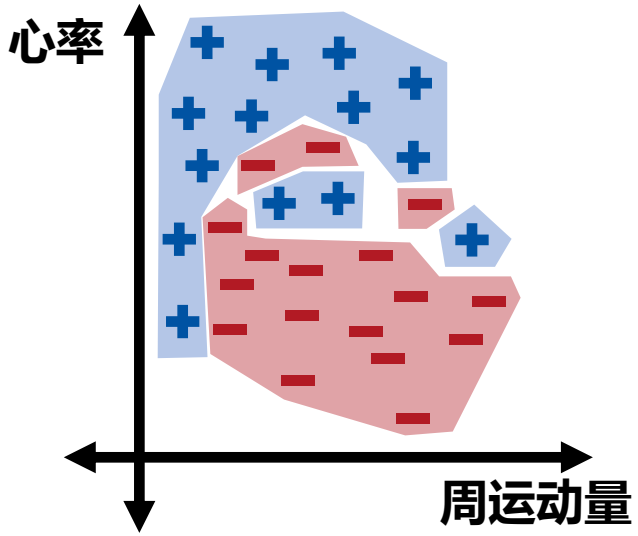


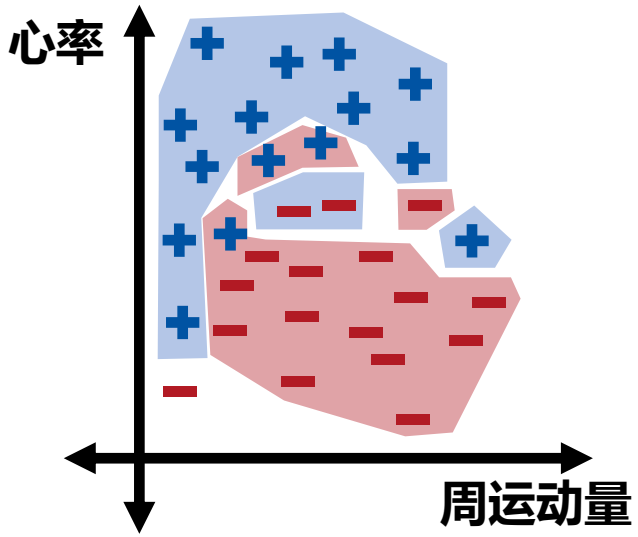
非线性边界



- 训练误差为0

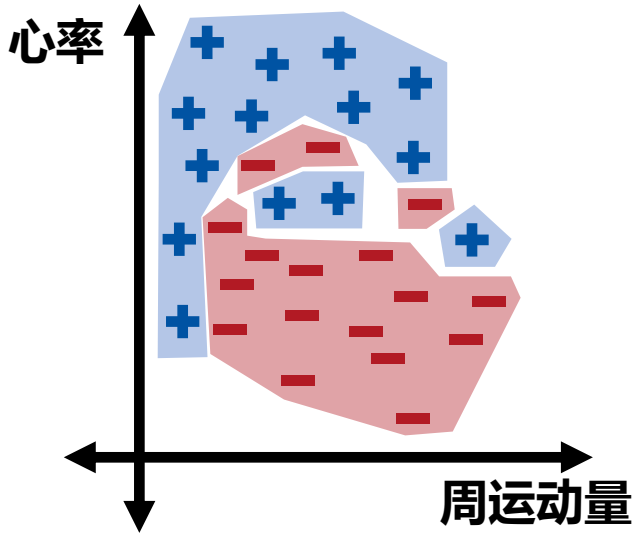
非线性边界

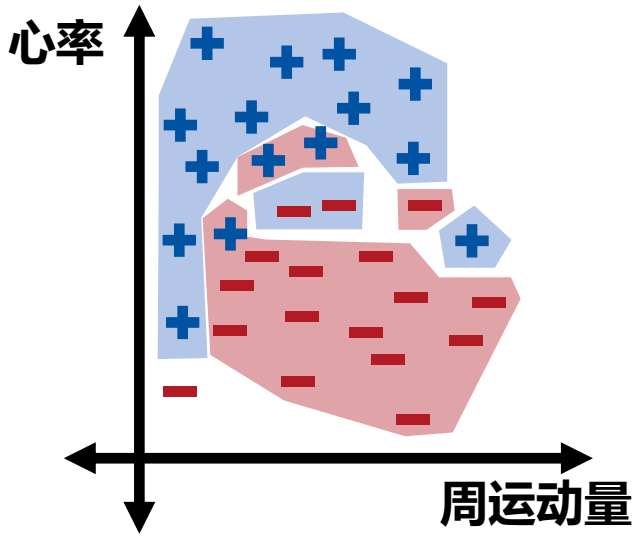




- 训练误差为0
- 过拟合！

非线性边界





- 训练误差为0
- 过拟合！
- 如何判断过拟合？
- 如何避免过拟合？

评估学习算法

- 如何评估学习算法在某个数据集上的优劣？

评估学习算法

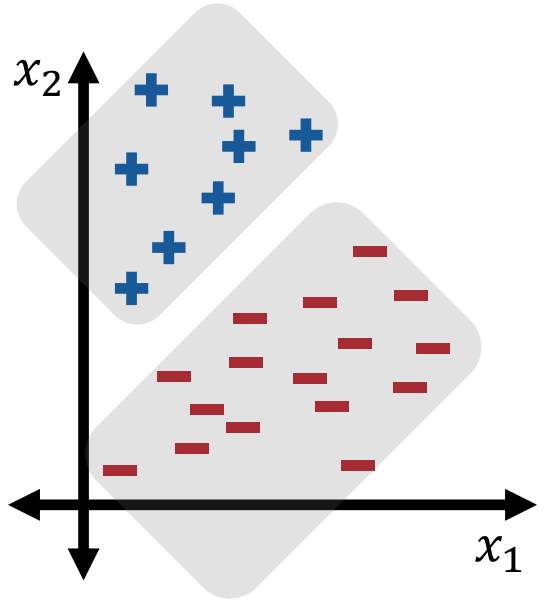
- 如何评估学习算法在某个数据集上的优劣？
- Idea 1：用整个数据集作为训练集，并评估训练误差

评估学习算法

- 如何评估学习算法在某个数据集上的优劣？
- Idea 1：用整个数据集作为训练集，并评估训练误差
- Idea 2：保留部分数据用作测试

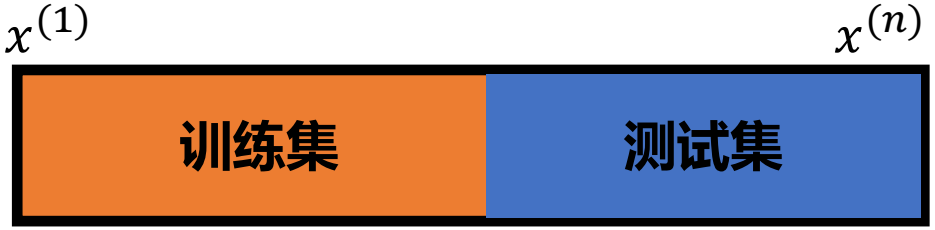
评估学习算法

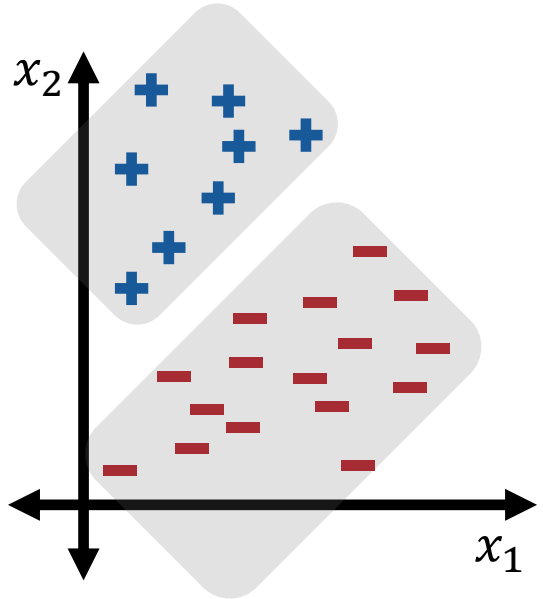
- 如何评估学习算法在某个数据集上的优劣？
- Idea 1：用整个数据集作为训练集，并评估训练误差
- Idea 2：保留部分数据用作测试
  - 更多的训练数据：近似全数据训练
  - 更多的测试数据：稳定性能评估
  - 单个分类器的性能不代表学习算法性能



评估学习算法

- 如何评估学习算法在某个数据集上的优劣？
- Idea 1：用整个数据集作为训练集，并评估训练误差
- Idea 2：保留部分数据用作测试
  - 更多的训练数据：近似全数据训练
  - 更多的测试数据：稳定性能评估
  - 单个分类器的性能不代表学习算法性能
  - trick：打乱数据集

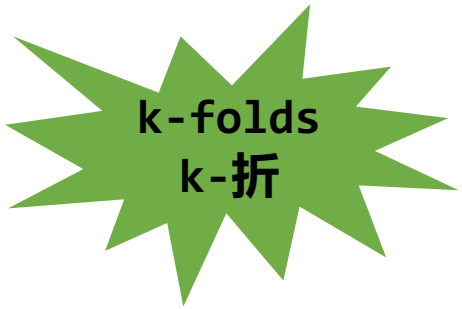




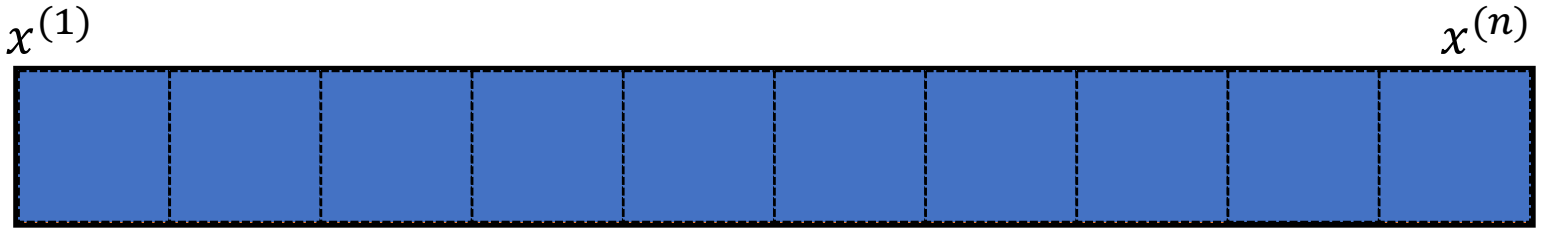
评估学习算法

Cross-validate($\mathcal{D}_n$, k)

将$\mathcal{D}_n$分为相同尺寸的k折$\mathcal{D}_{n,1}, \mathcal{D}_{n,2}, \ldots, \mathcal{D}_{n,k}$



评估学习算法



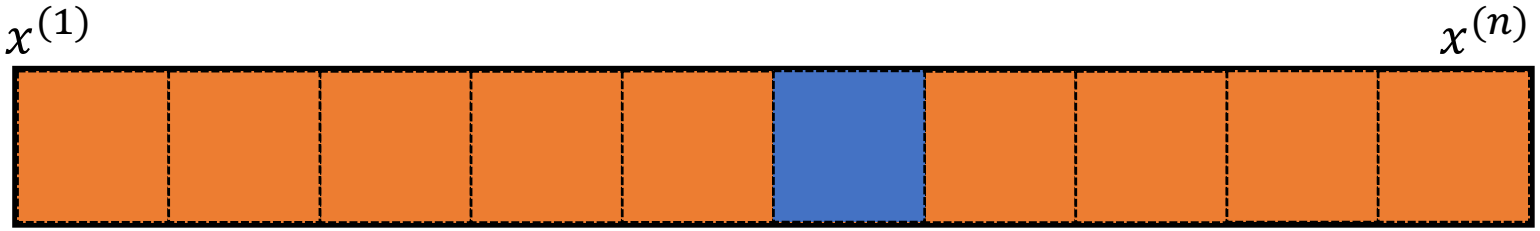
Cross-validate($\mathcal{D}_n$, k)

将$\mathcal{D}_n$分为相同尺寸的k折$\mathcal{D}_{n,1}, \mathcal{D}_{n,2}, \ldots, \mathcal{D}_{n,k}$

for i = 1 to k

k-folds
k-折

评估学习算法



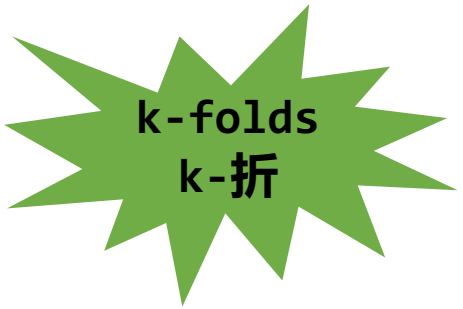
Cross-validate($\mathcal{D}_n$, k)

将$\mathcal{D}_n$分为相同尺寸的k折$\mathcal{D}_{n,1}, \mathcal{D}_{n,2}, \ldots, \mathcal{D}_{n,k}$

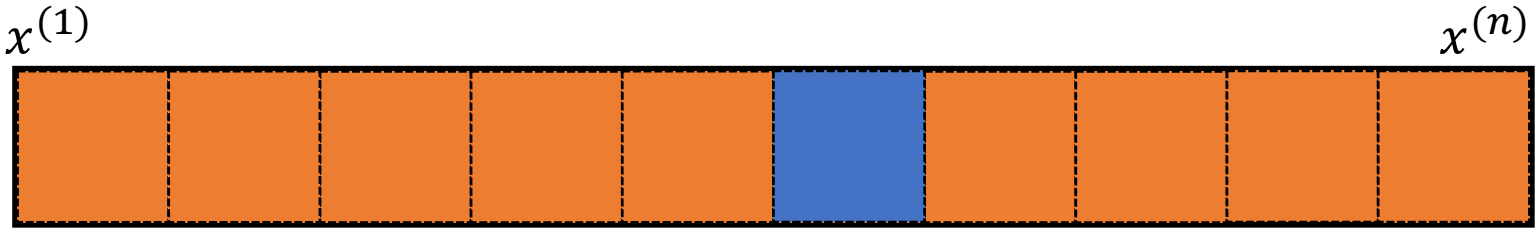
for i = 1 to k

在$\mathcal{D}_n \backslash \mathcal{D}_{n,i}$上训练分类器$h_i$(即除了第i折)

计算h_i在$\mathcal{D}_{n,i}$上的测试误差$\mathcal{E}(h_i, \mathcal{D}_{n,i})$



评估学习算法



Cross-validate($\mathcal{D}_n$, k)

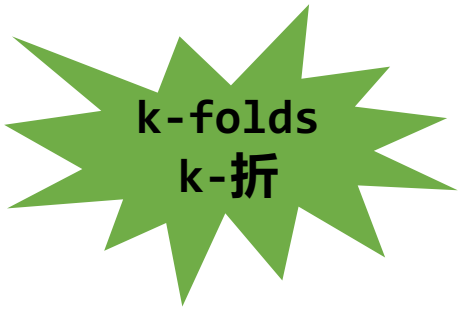
将$\mathcal{D}_n$分为相同尺寸的k折$\mathcal{D}_{n,1}, \mathcal{D}_{n,2}, \ldots, \mathcal{D}_{n,k}$

for i = 1 to k

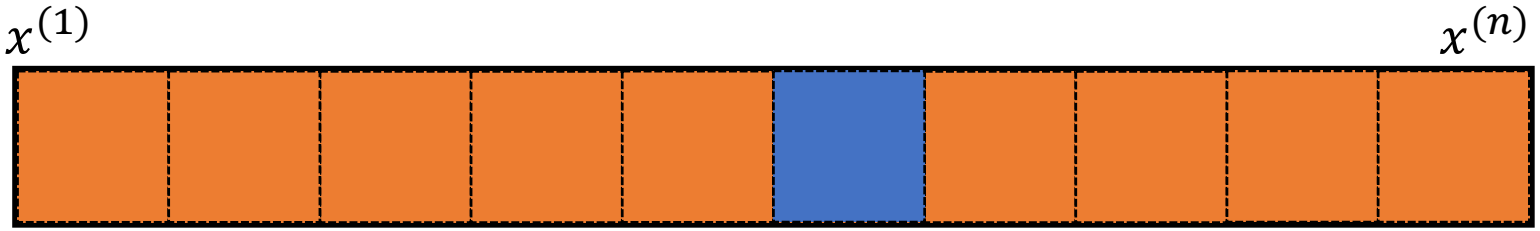
在$\mathcal{D}_n \backslash \mathcal{D}_{n,i}$上训练分类器$h_i$(即除了第i折)

计算h_i在$\mathcal{D}_{n,i}$上的测试误差$\mathcal{E}(h_i, \mathcal{D}_{n,i})$

Return $\frac{1}{k}\sum_{i=1}^{k}\mathcal{E}(h_i, \mathcal{D}_{n,i})$



评估学习算法



Cross-validate($\mathcal{D}_n$, k)

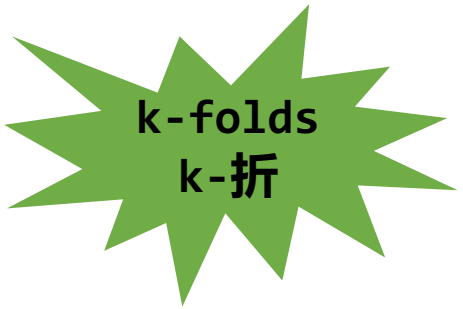
将$\mathcal{D}_n$分为相同尺寸的k折$\mathcal{D}_{n,1}, \mathcal{D}_{n,2}, \ldots, \mathcal{D}_{n,k}$

for i = 1 to k

在$\mathcal{D}_n \backslash \mathcal{D}_{n,i}$上训练分类器$h_i$(即除了第i折)

计算h_i在$\mathcal{D}_{n,i}$上的测试误差$\mathcal{E}(h_i, \mathcal{D}_{n,i})$

Return $\frac{1}{k}\sum_{i=1}^{k}\mathcal{E}(h_i, \mathcal{D}_{n,i})$



- trick：打乱数据集