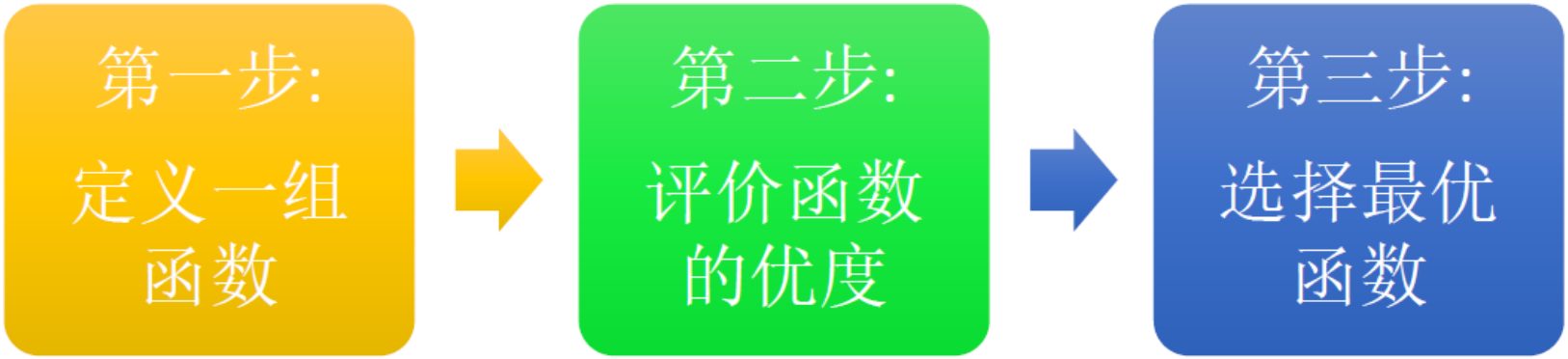
机器学习≈寻找一个函数



复习

监督学习

- 回归
- 分类
  - 二分类
  - 多分类

半监督学习

无监督学习

上节课后练习

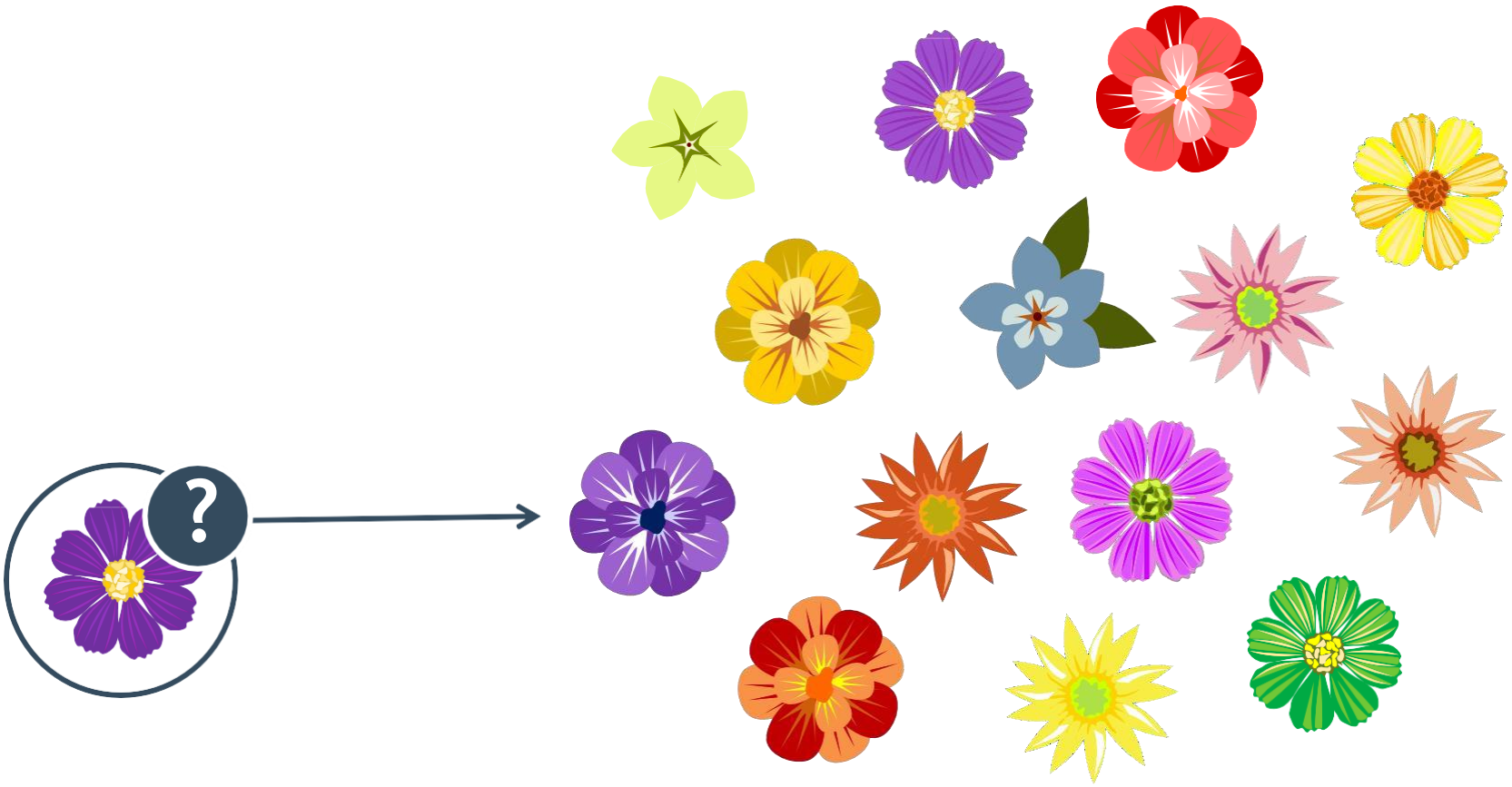
- 请在实际生活中找出以下各类问题（各两例）：
  - 回归
  - 二分类
  - 多分类
  - 聚类

什么是分类？

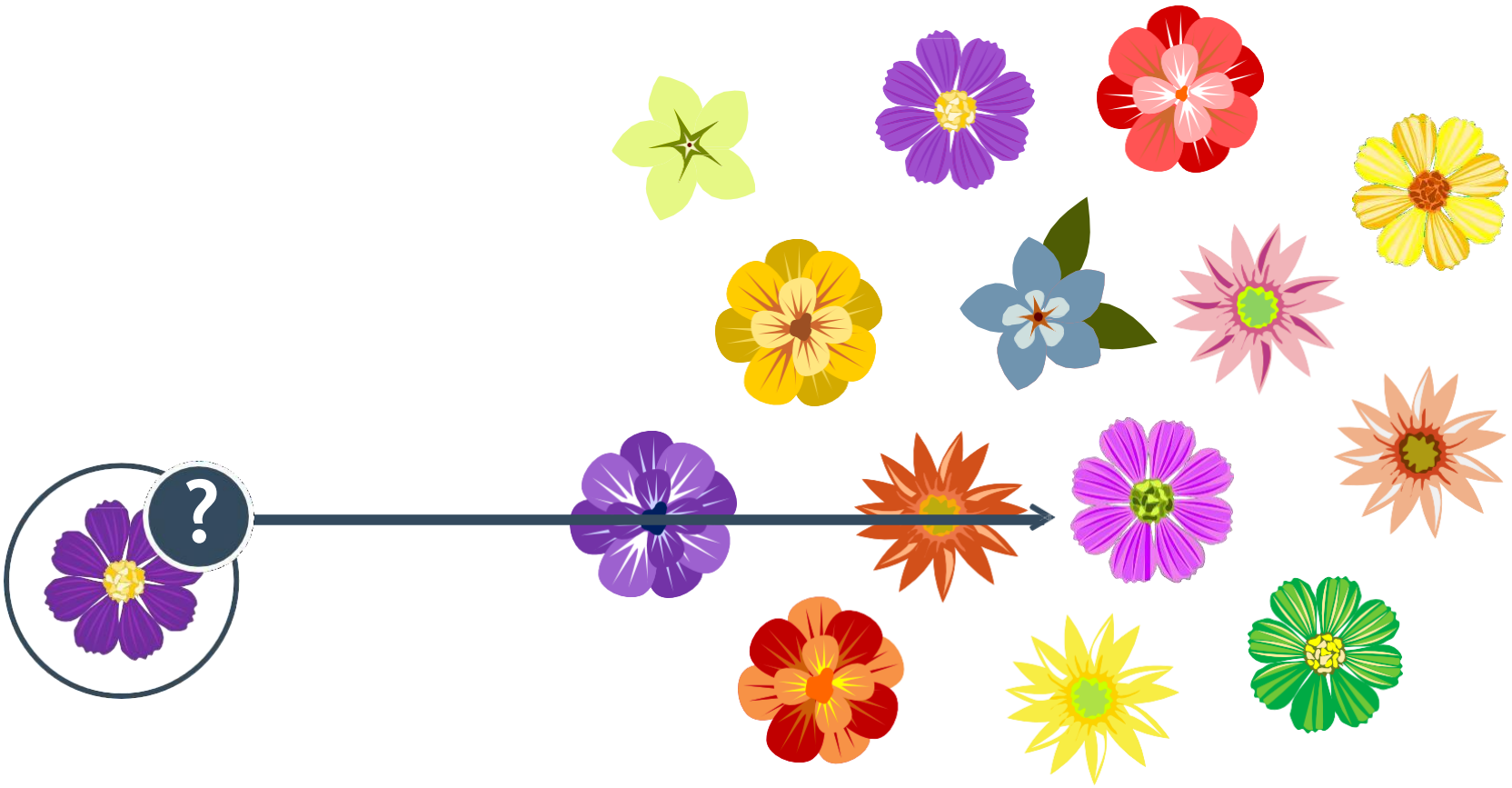
一家花店想根据某顾客最近买花的情况，来预测某种新来的花是否会被该顾客购买

什么是分类？

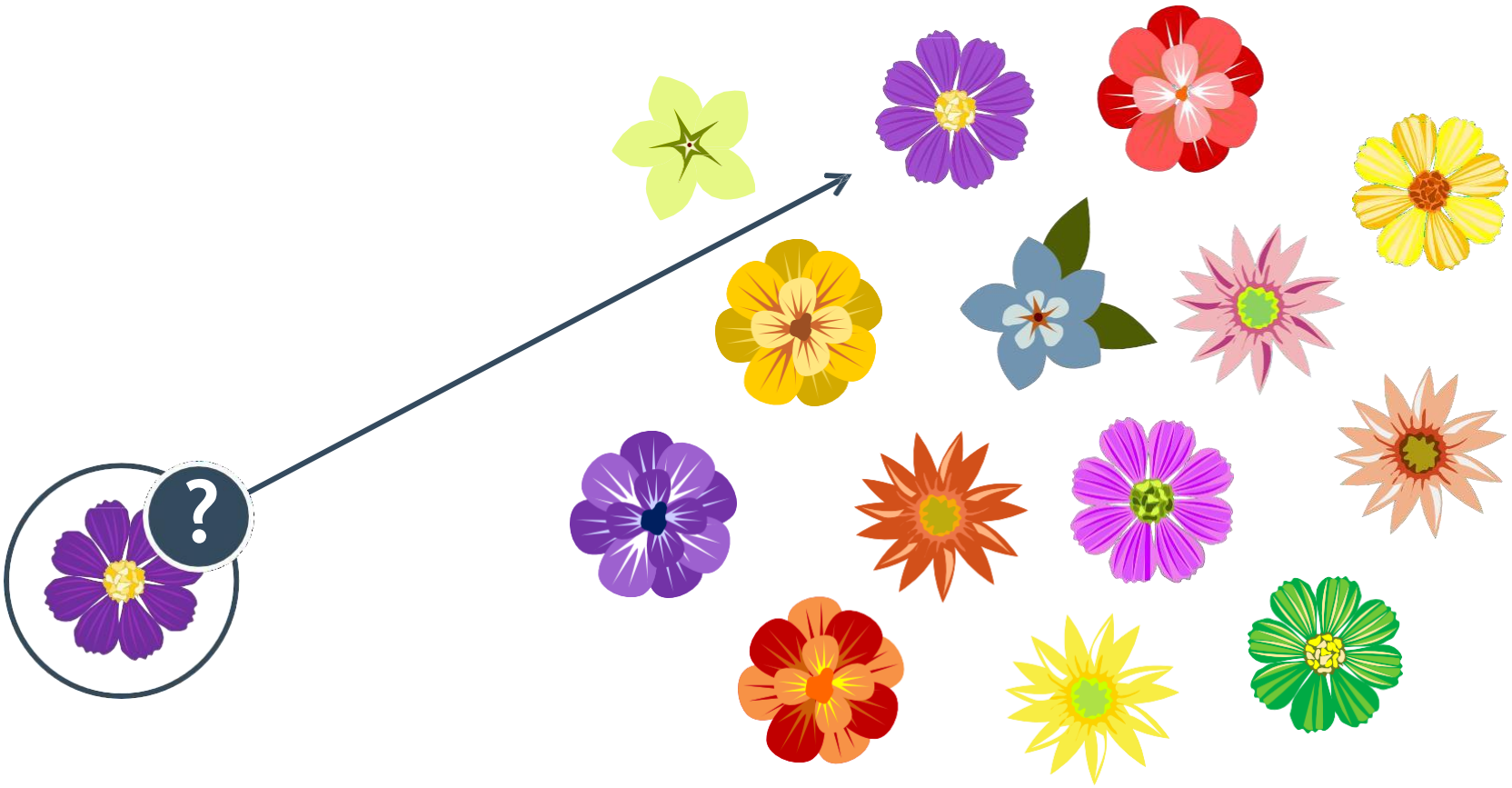
什么是分类？



什么是分类？



什么是分类？



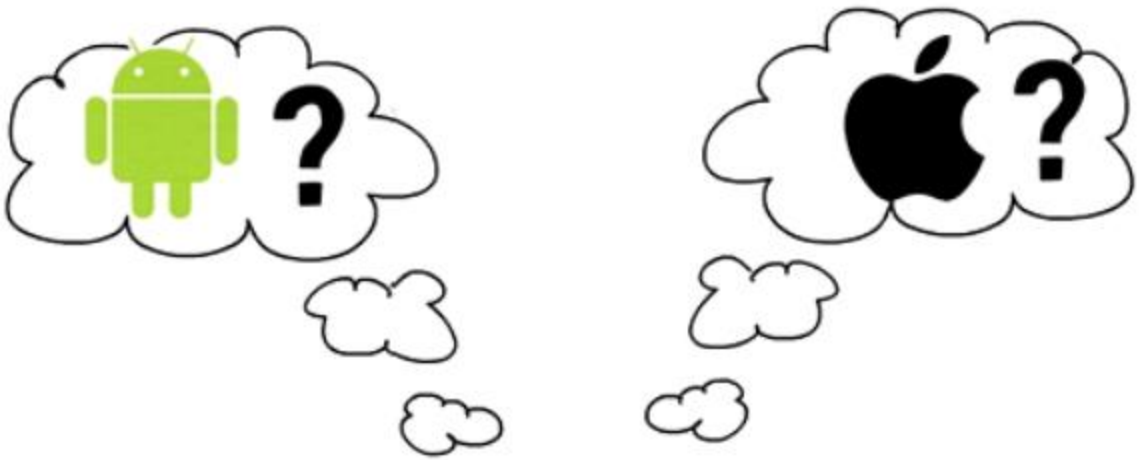
分类需要什么？

- 数据：
  - 将对象表示为量化的一组特征
  - 给定类别标签
- 对象间相似性的度量

引例1

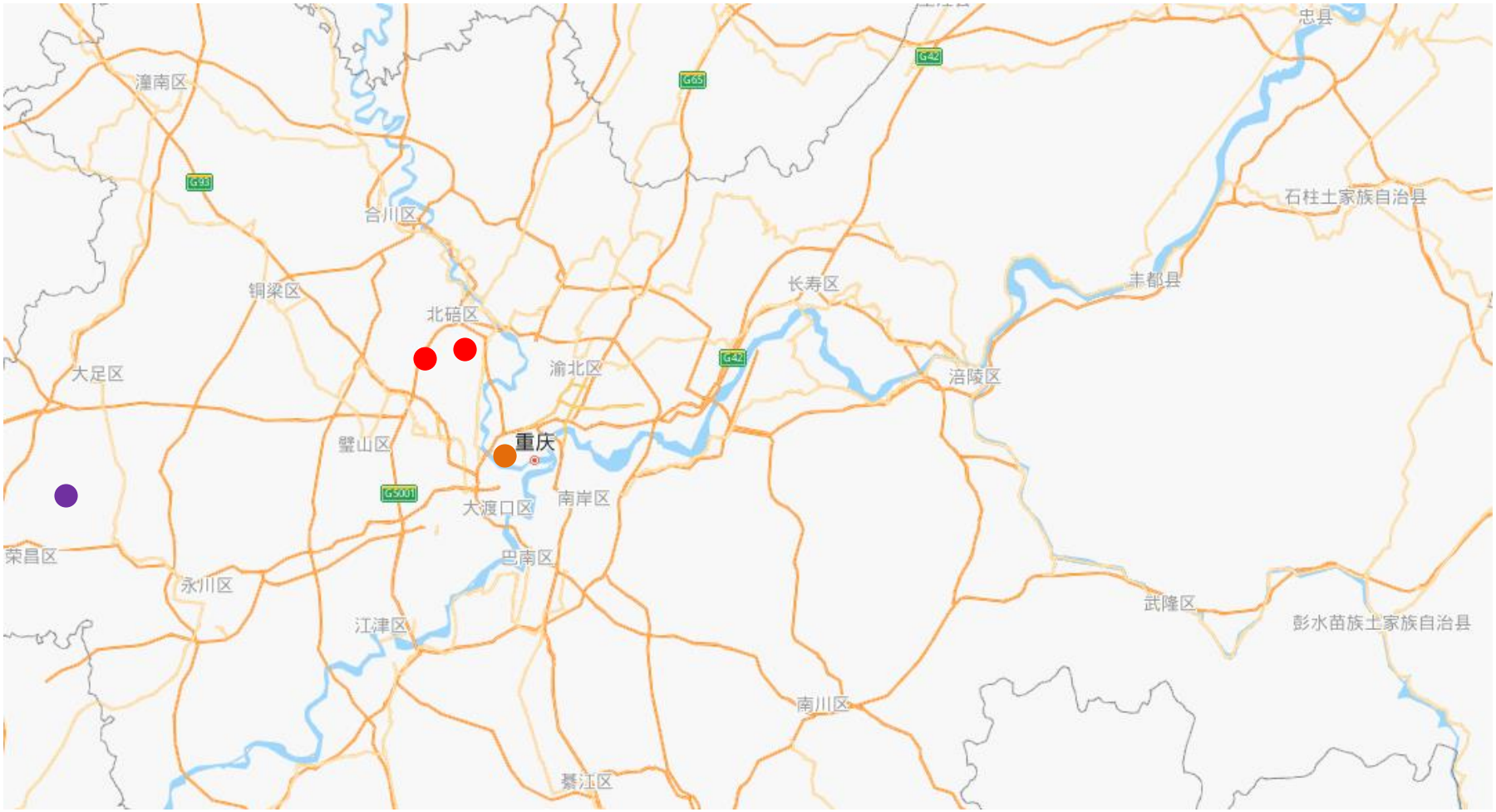
3个选项可参考，怎么选？

- 随便选一个
- 问问周围其他人的选择
- 做一份详细报告，根据自身情况量身定制选择方案



还是先问问其他人用什么吧?

引例2


潼南区
合川区
铜梁区
北碚区
渝北区
大足区
璧山区
重庆
南岸区
大渡口区
巴南区
荣昌区
永川区
江津区
南川区
綦江区
长寿区
涪陵区
丰都县
石柱土家族自治县
忠县
武隆区
彭水苗族土家族自治县
G93
G65
G42
G5001

分类--- K近邻

什么是K近邻？

K最近邻(k-Nearest Neighbor，KNN)分类算法，是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。

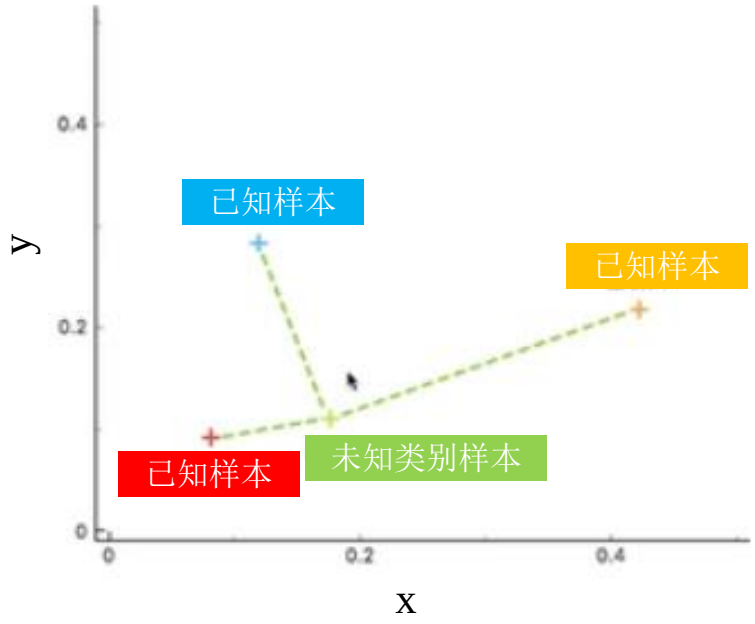
该方法的思路是：在特征空间中，如果一个样本附近的k个最近(即特征空间中最邻近)样本的大多数属于某一个类别，则该样本也属于这个类别。

K近邻算法采用测量不同特征值之间的距离的方法进行分类

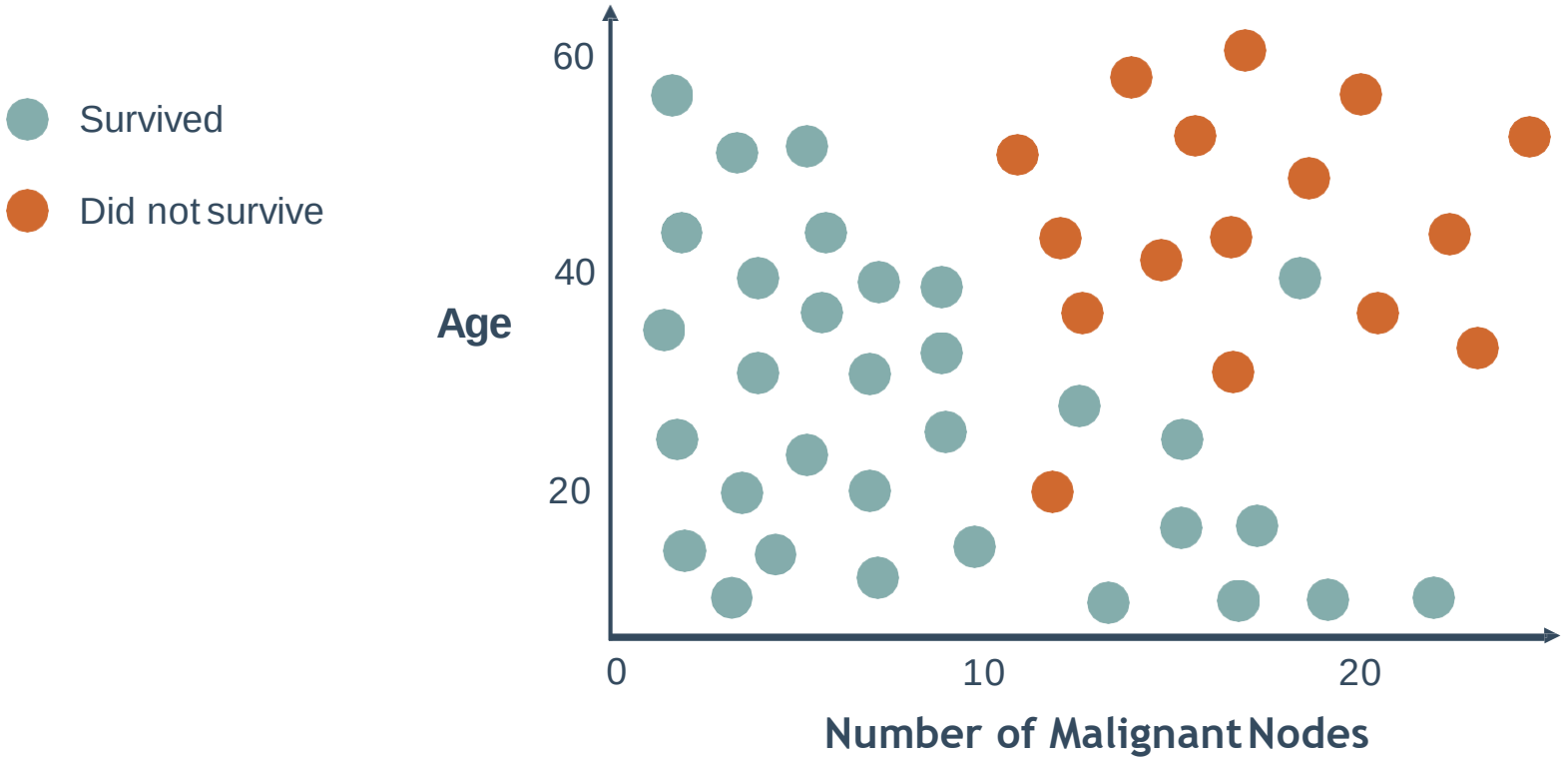
就是通过你的“邻居”来判断你是属于哪个类别

最近邻算法

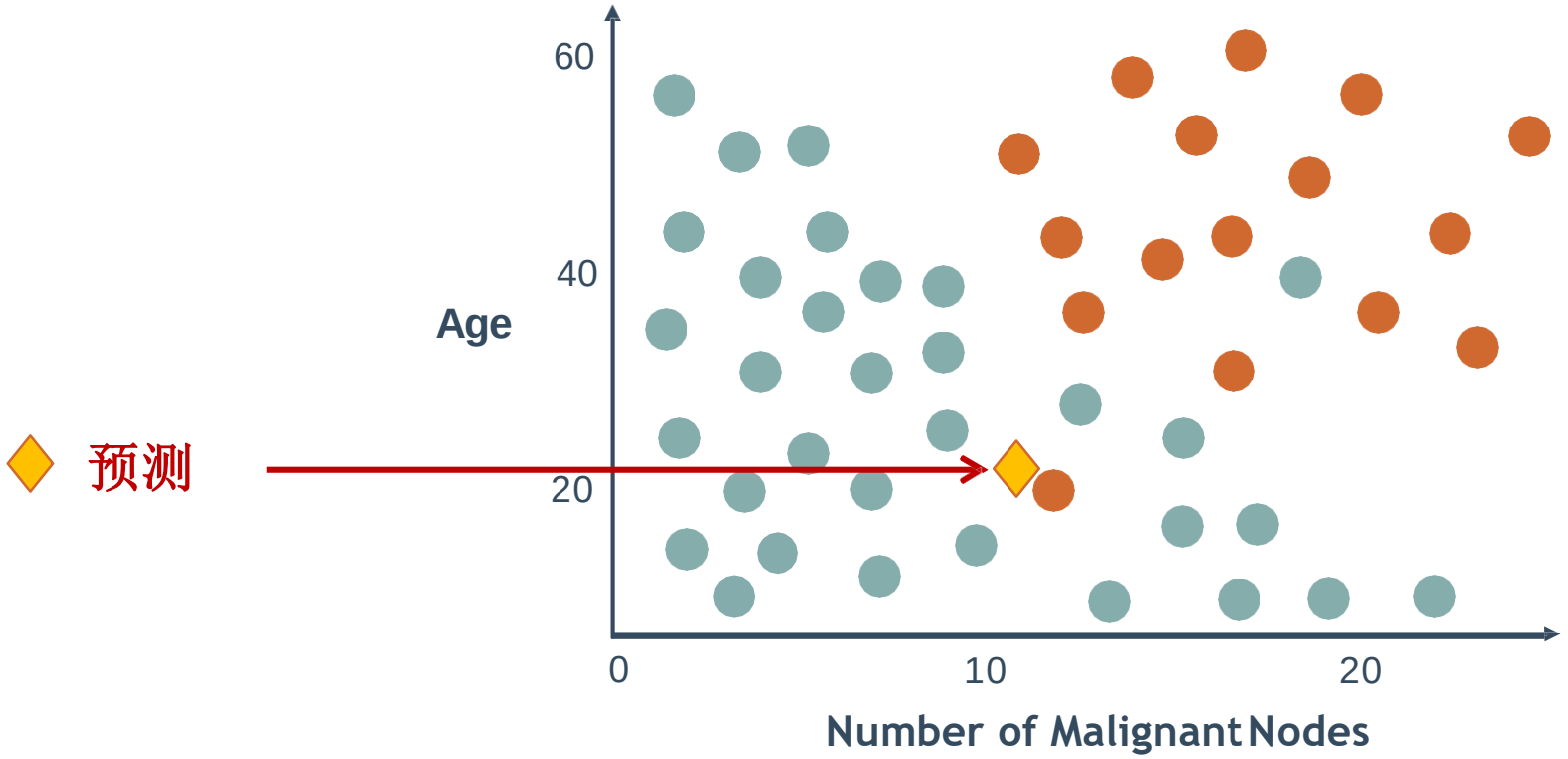
最近邻算法(Nearest Neighbor，NN)，其针对未知类别数据x，在训练集中找到与x最相似的训练样本y，用y的样本对应的类别作为未知类别数据x 的类别，从而达到分类的效果。



K近邻（KNN）分类


Survived
Did not survive
60
40
20
Age
0
10
20
Number of Malignant Nodes

K近邻（KNN）分类



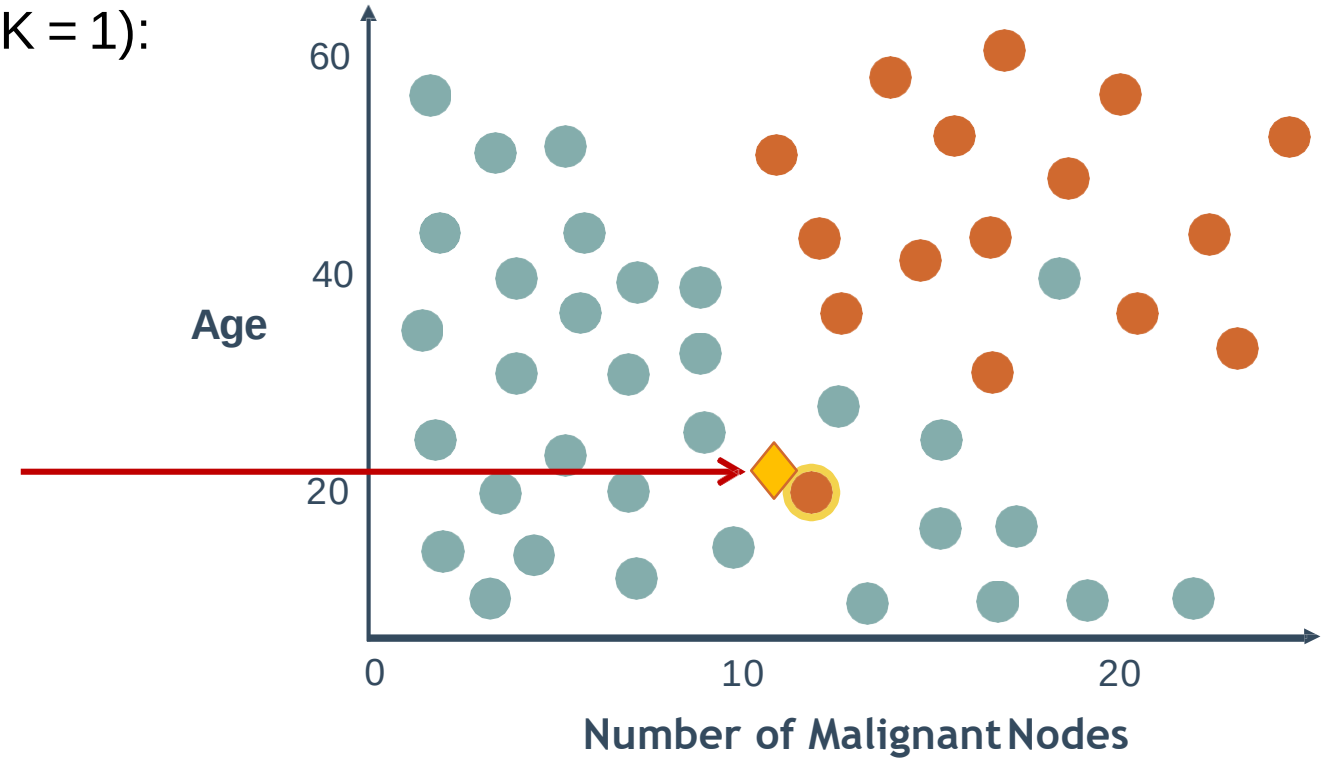
K近邻（KNN）分类

近邻数目(K = 1):

0

1

预测



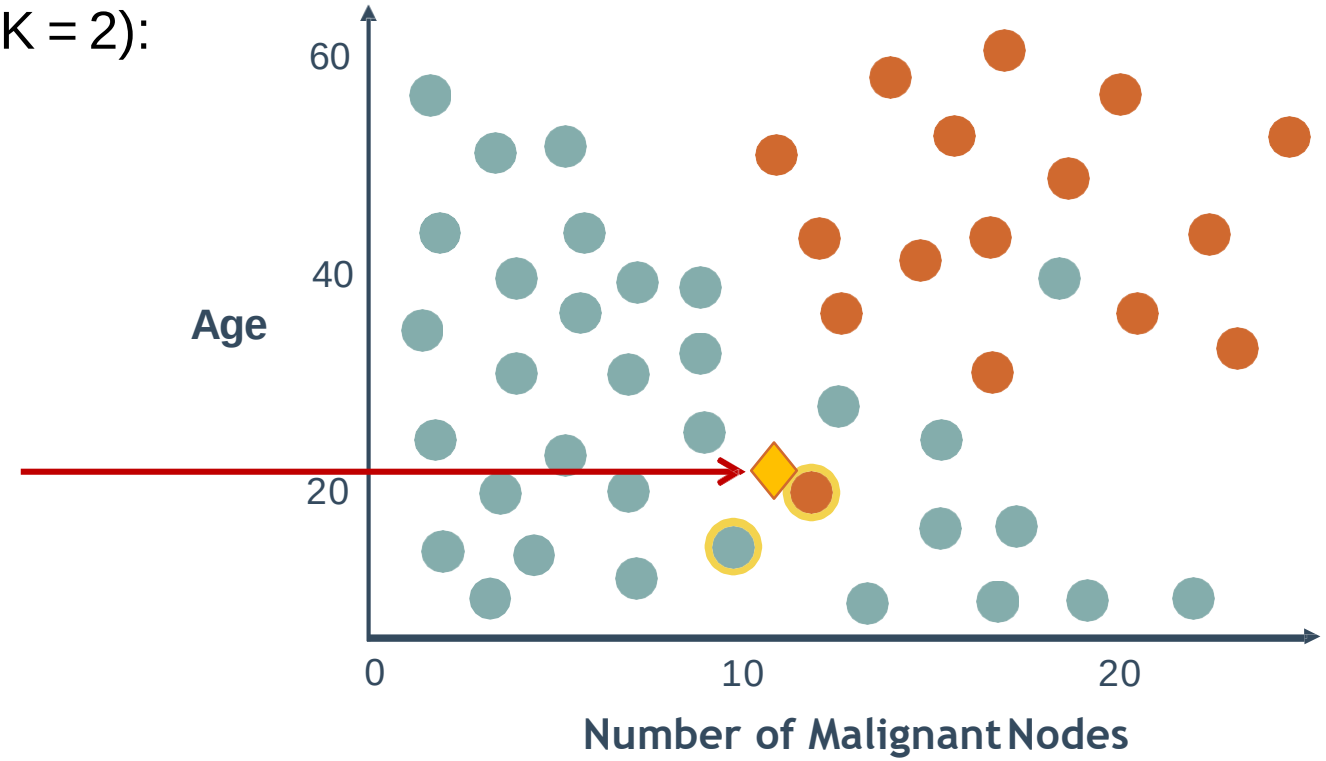
K近邻（KNN）分类

近邻数目(K = 2):

1

1

预测



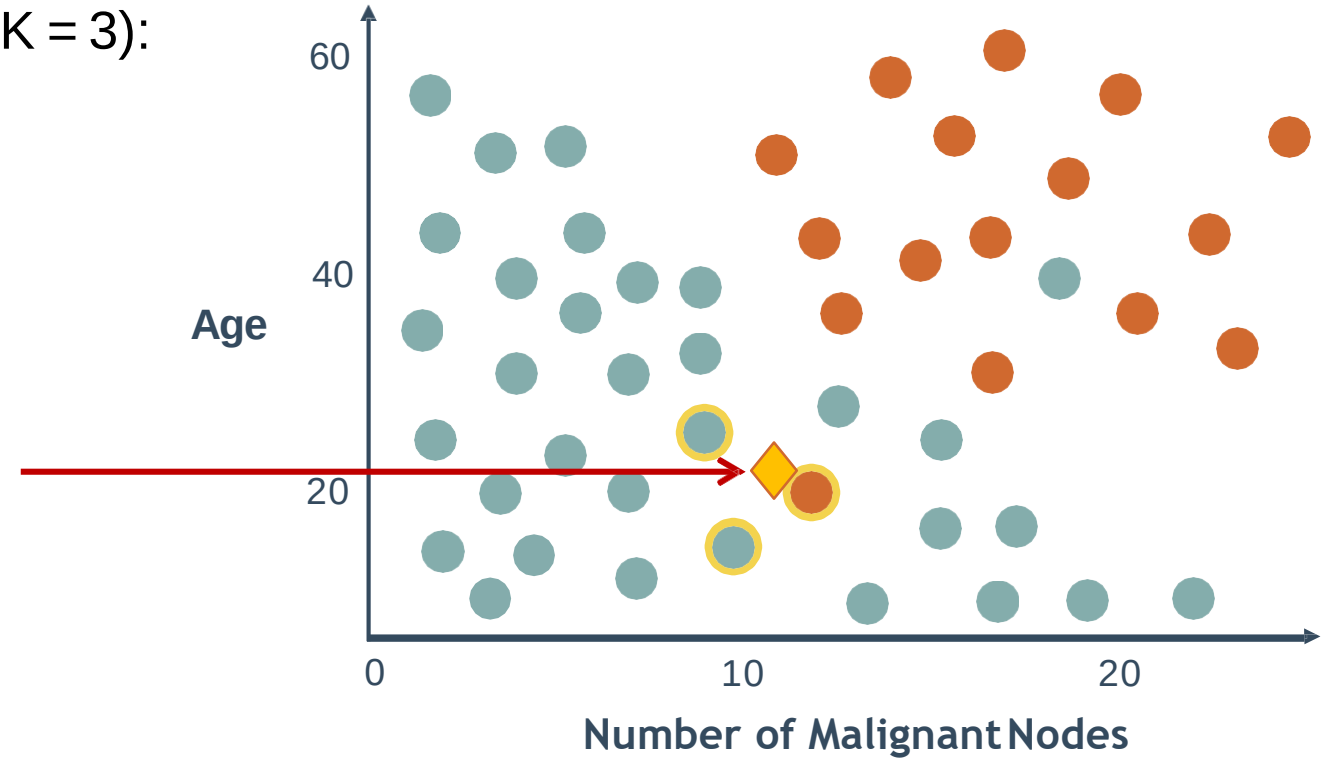
K近邻（KNN）分类

近邻数目(K = 3):

2

1

预测

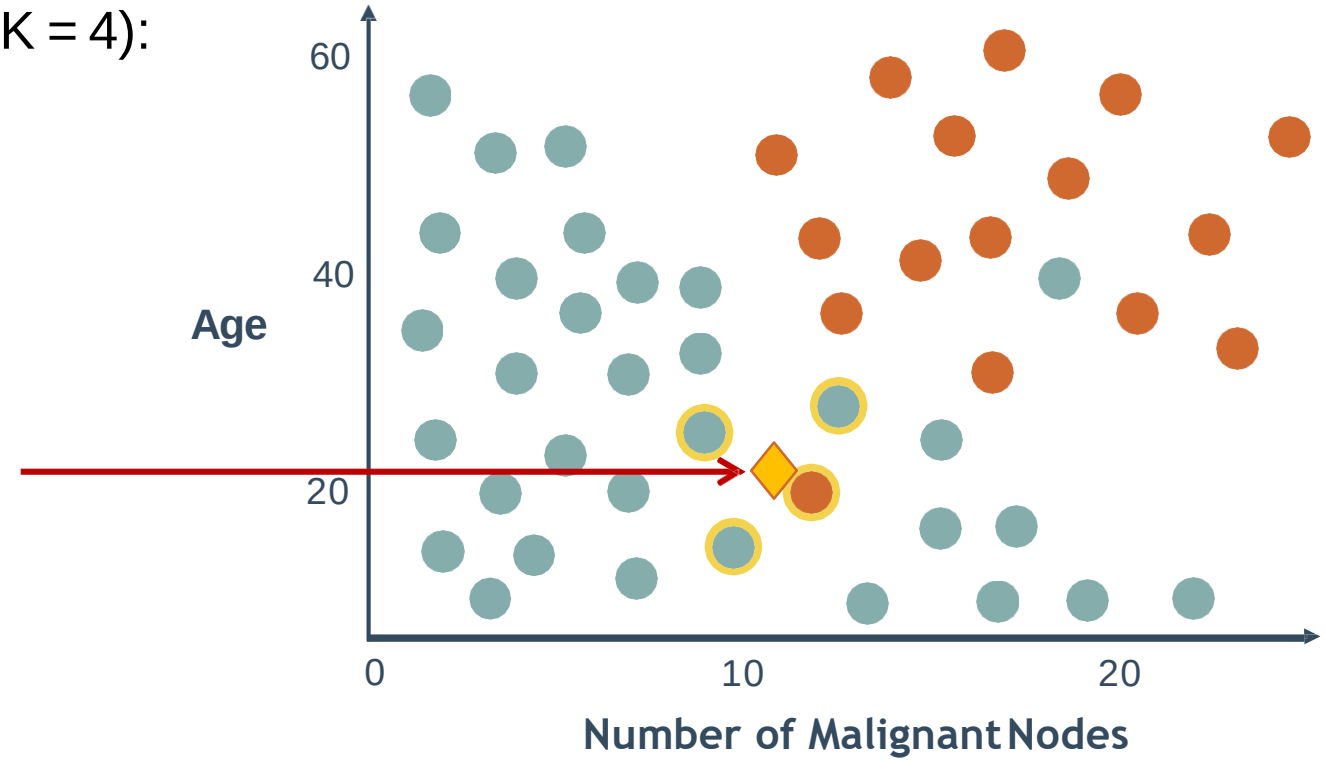


K近邻（KNN）分类

近邻数目(K = 4):

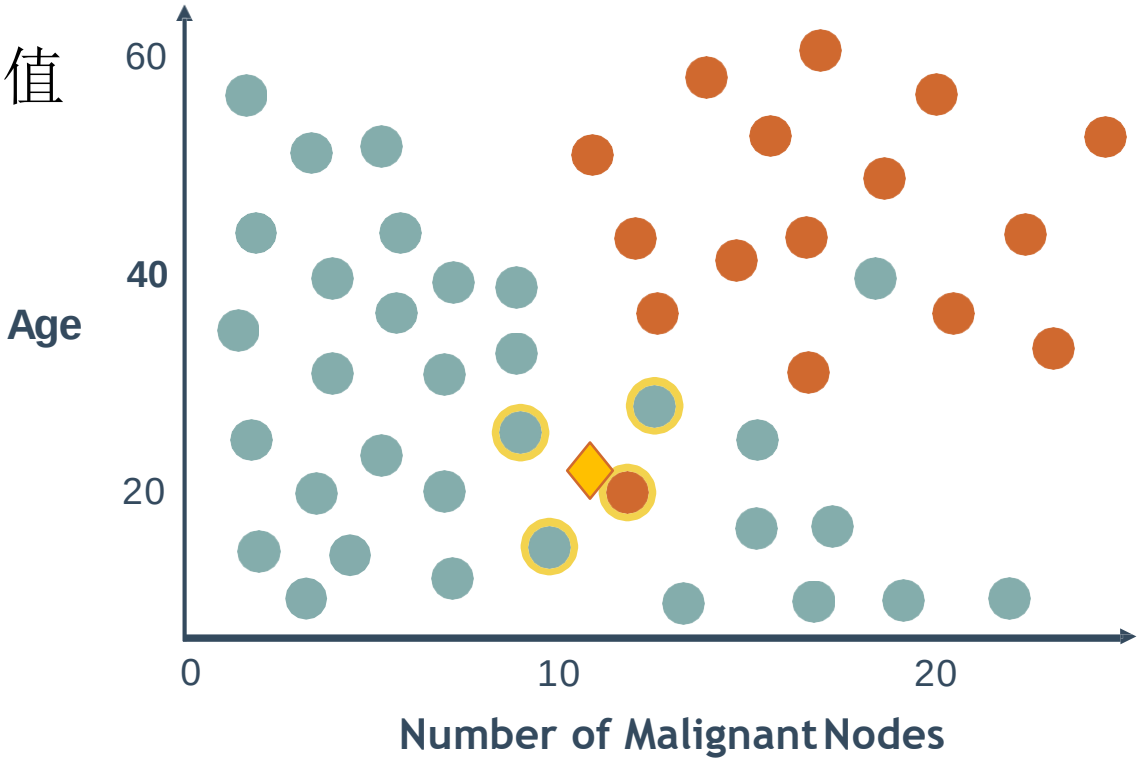
- 3
- 1

预测

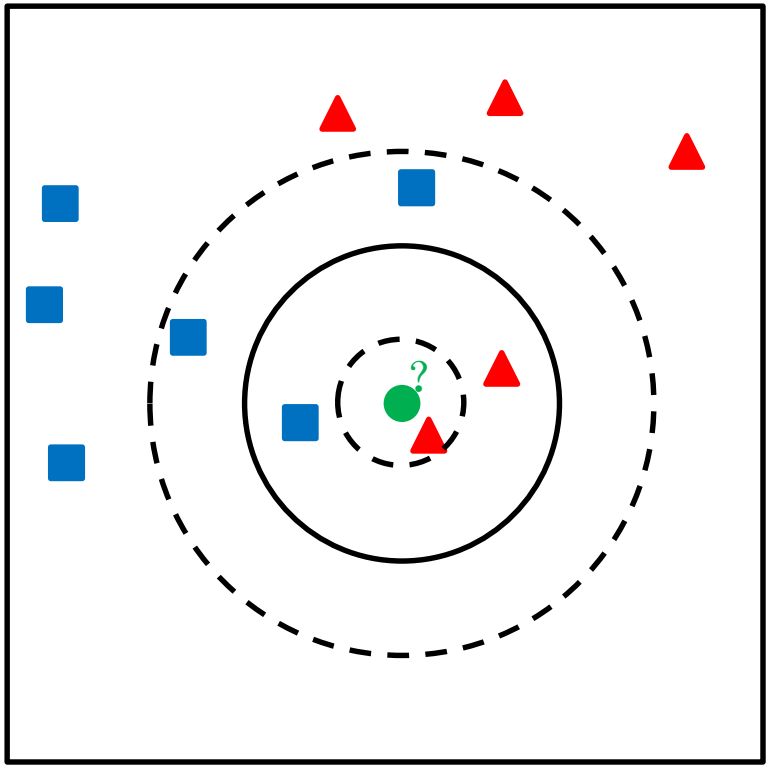


K近邻模型需要选择

- 正确的“K”值
- 如何度量相邻两点之间的相似性/距离？

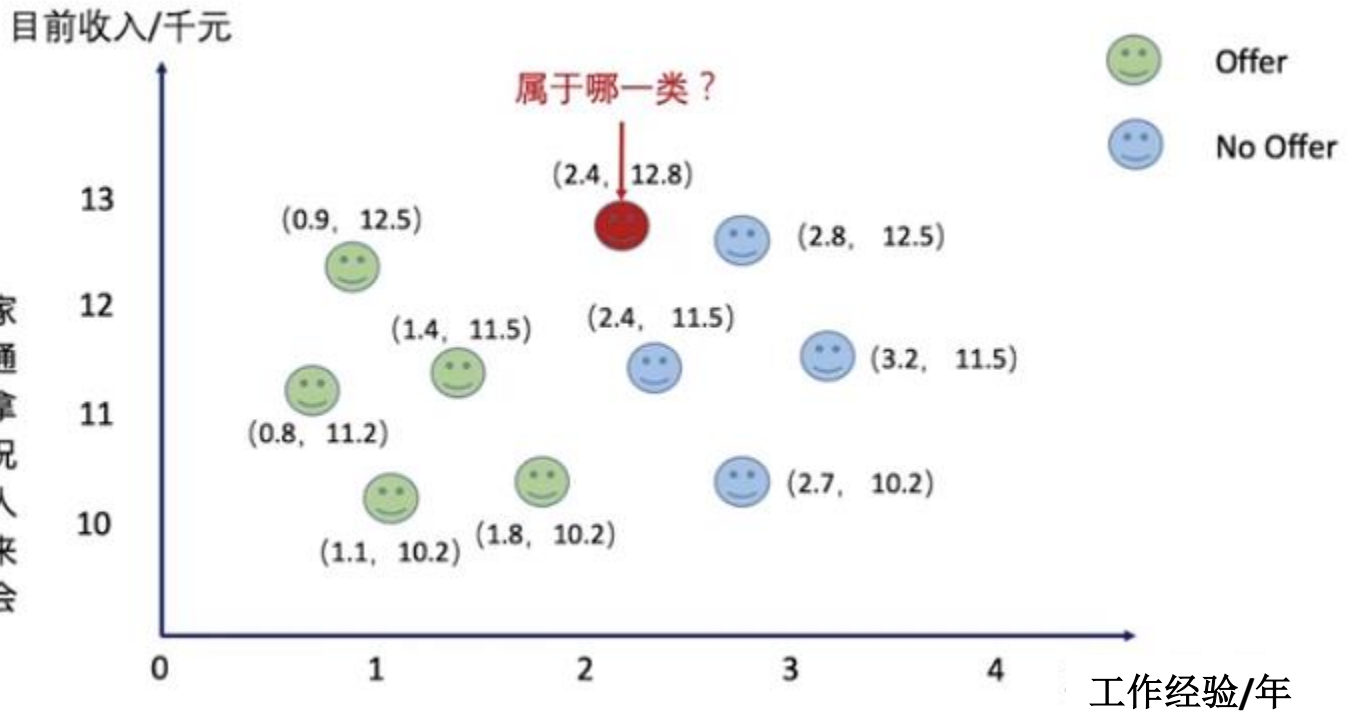


K近邻（KNN）算法

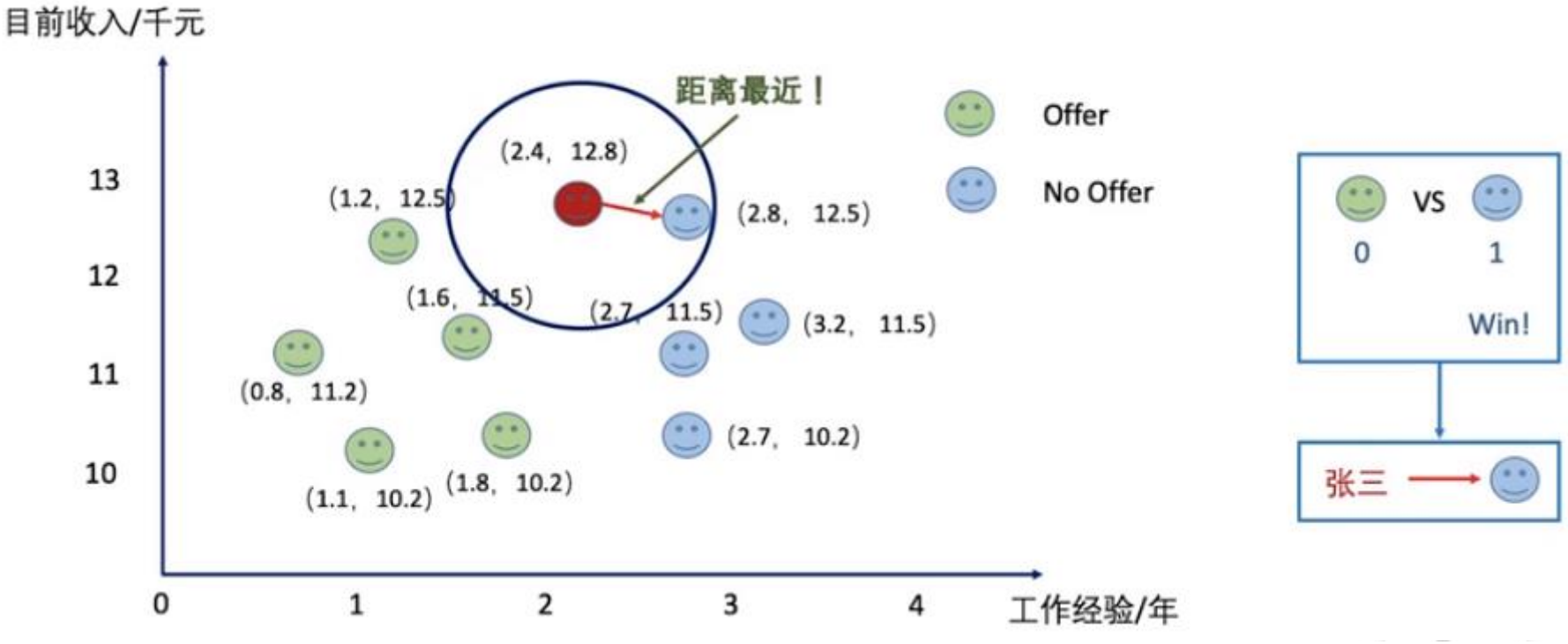


K近邻（KNN）算法

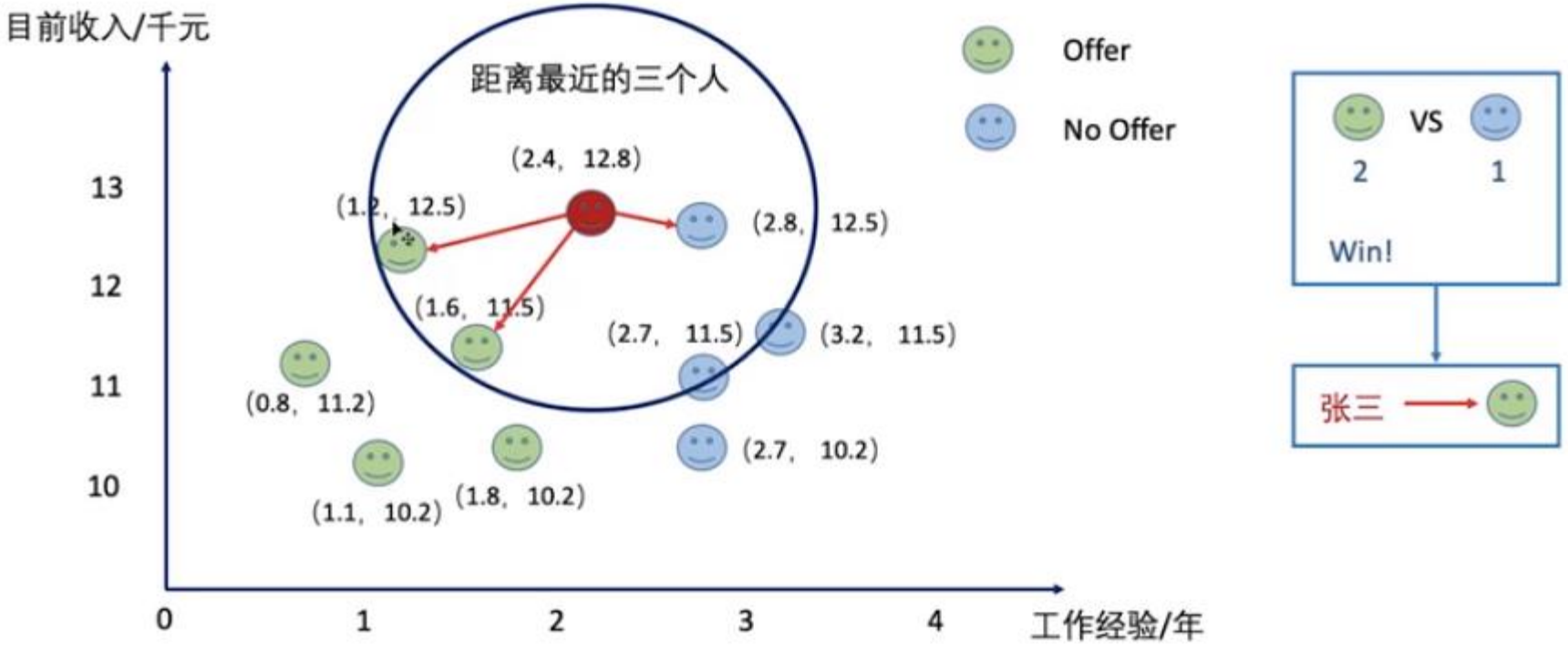
Q: 张三要参加一家公司的面试，但它通过各种渠道得知了拿到offer的人的情况和没有拿到offer人的情况。我们一起来预测一下张三是否会拿到offer?



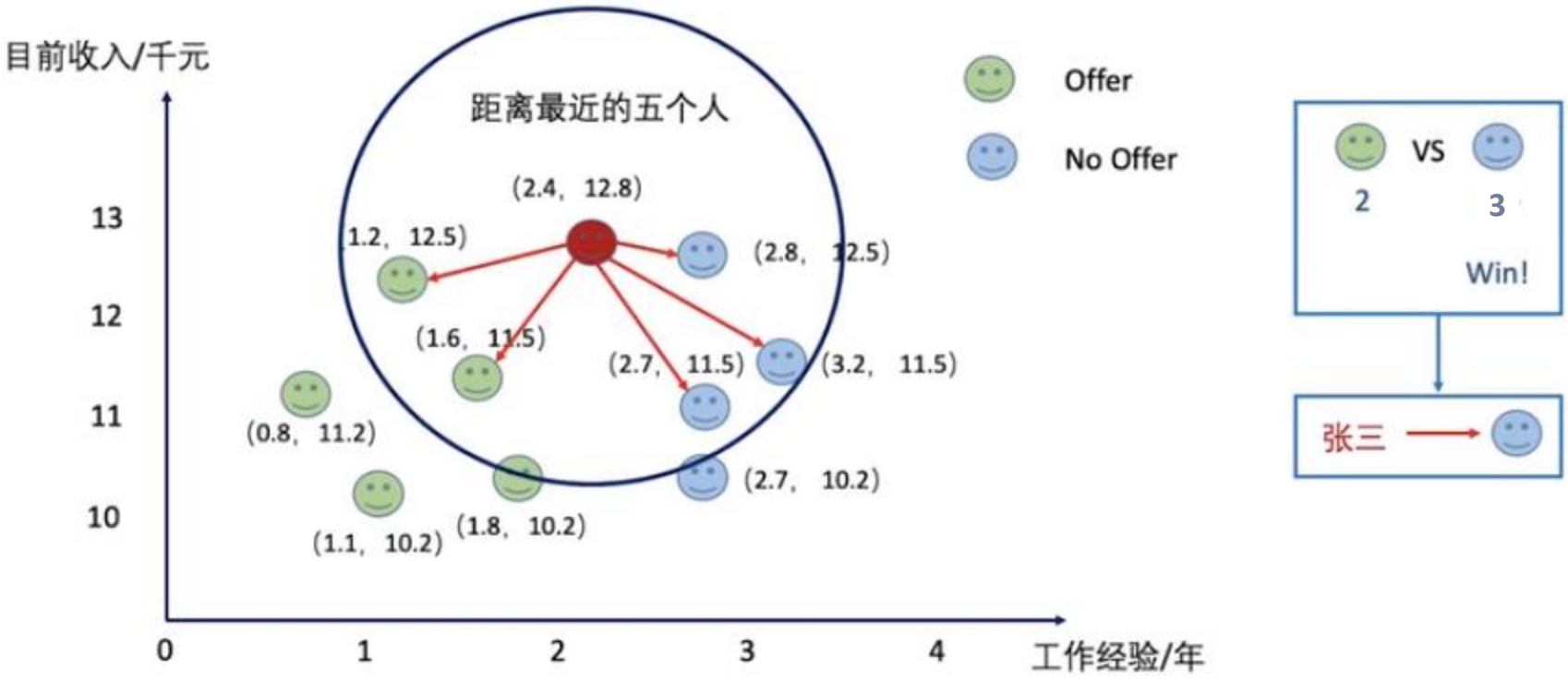
K近邻（KNN）算法（K=1）


目前收入/千元
13
12
11
10
0
1
2
3
4
工作经验/年
距离最近！
(2.4，12.8)
(1.2，12.5)
(2.8，12.5)
(1.6，11.5)
(2.7，11.5)
(3.2，11.5)
(0.8，11.2)
(2.7，10.2)
(1.1，10.2)
(1.8，10.2)
Offer
No Offer
VS
0
1
Win!
张三

K近邻（KNN）算法（K=3）


目前收入/千元
13
12
11
10
0
1
2
3
4
工作经验/年
距离最近的三个人
(2.4，12.8)
(1.2，12.5)
(2.8，12.5)
(1.6，11.5)
(2.7，11.5)
(3.2，11.5)
(0.8，11.2)
(2.7，10.2)
(1.1，10.2)
(1.8，10.2)
Offer
No Offer
VS
2
1
Win!
张三

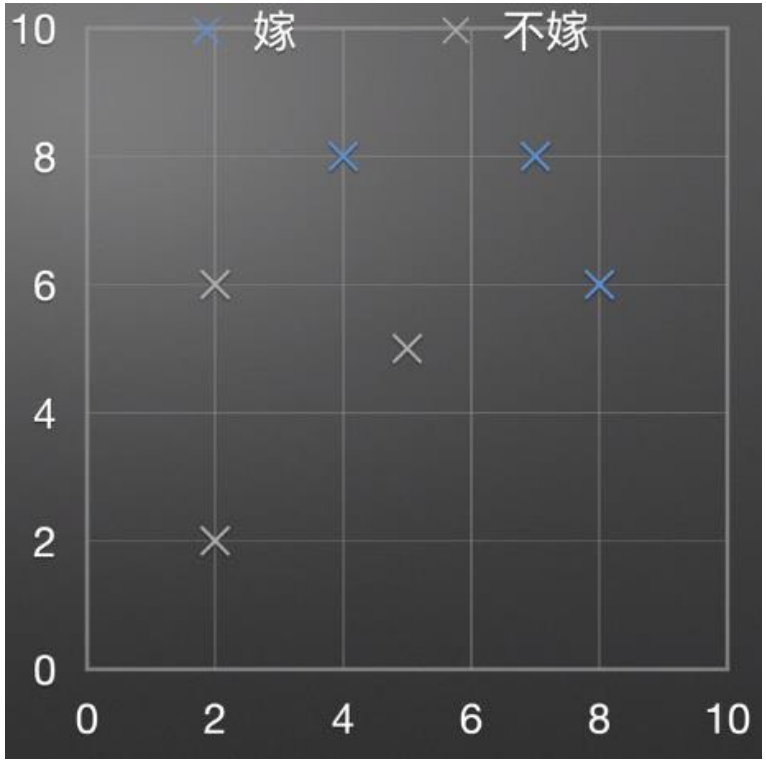
K近邻（KNN）算法（K=5）


目前收入/千元
13
12
11
10
0
1
2
3
4
工作经验/年
距离最近的五个人
(2.4，12.8)
(1.2，12.5)
(2.8，12.5)
(1.6，11.5)
(2.7，11.5)
(3.2，11.5)
(0.8，11.2)
(2.7，10.2)
(1.8，10.2)
(1.1，10.2)
Offer
No Offer
VS
2
3
Win!
张三

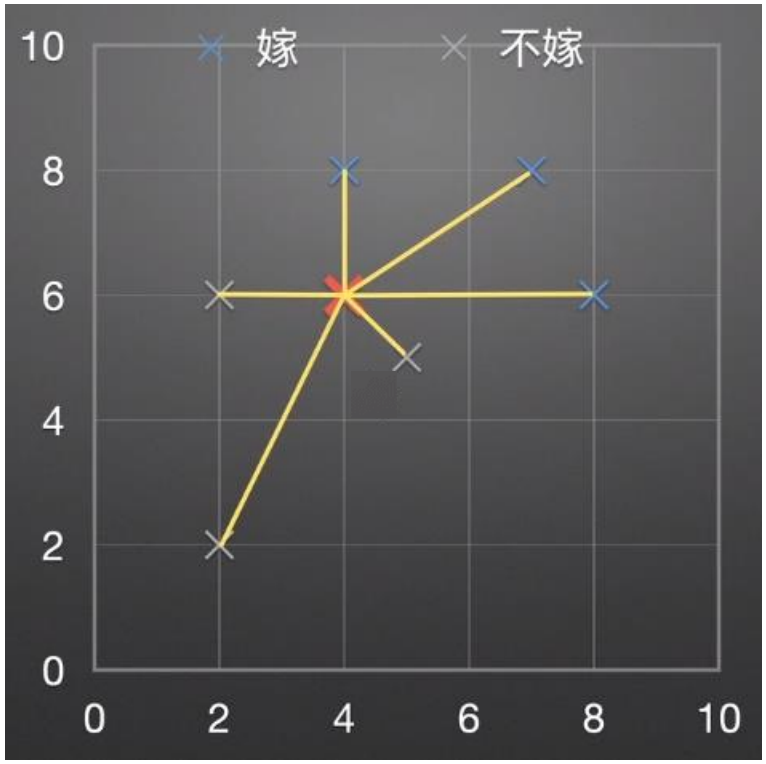
K近邻（KNN）算法

序号	财富	颜值	嫁吗
1	7	8	嫁
2	8	6	嫁
3	4	8	嫁
4	5	5	不嫁
5	2	2	不嫁
6	2	6	不嫁
7	4	6	?

K近邻（KNN）算法



K近邻（KNN）算法

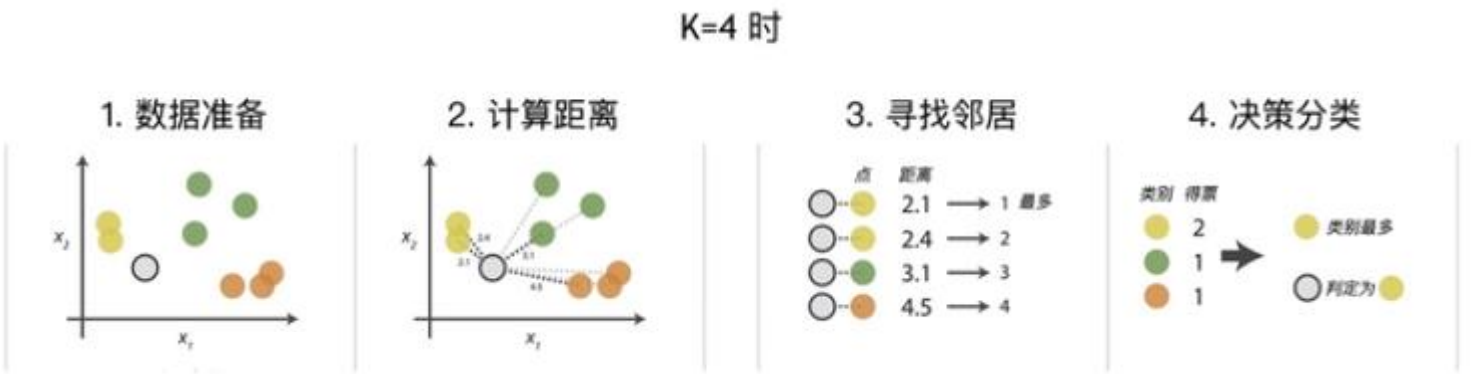


K近邻（KNN）算法

问题：

在使用K近邻算法的时候，我们一般会选择奇数的K值，为什么？

K近邻算法 实现步骤



把每一个物体表示成向量/矩阵

标记好每一个物体的标签

距离排序

选取K个距离最小的点

确定前K个点所在类别的出现频率

前K个点所出现频率最高的类别最为当前点的预测分类

K近邻算法 实现步骤

1. 把一个物体表示成向量

- 这也叫做 **“特征工程”** 英文叫**Feature Engineering**
- 模型的**输入一定是数量化的信息**，我们需要把现实生活中的物体表示成**向量/矩阵/张量**形式。

人  = ( )

图片  = ( )

K近邻算法 实现步骤

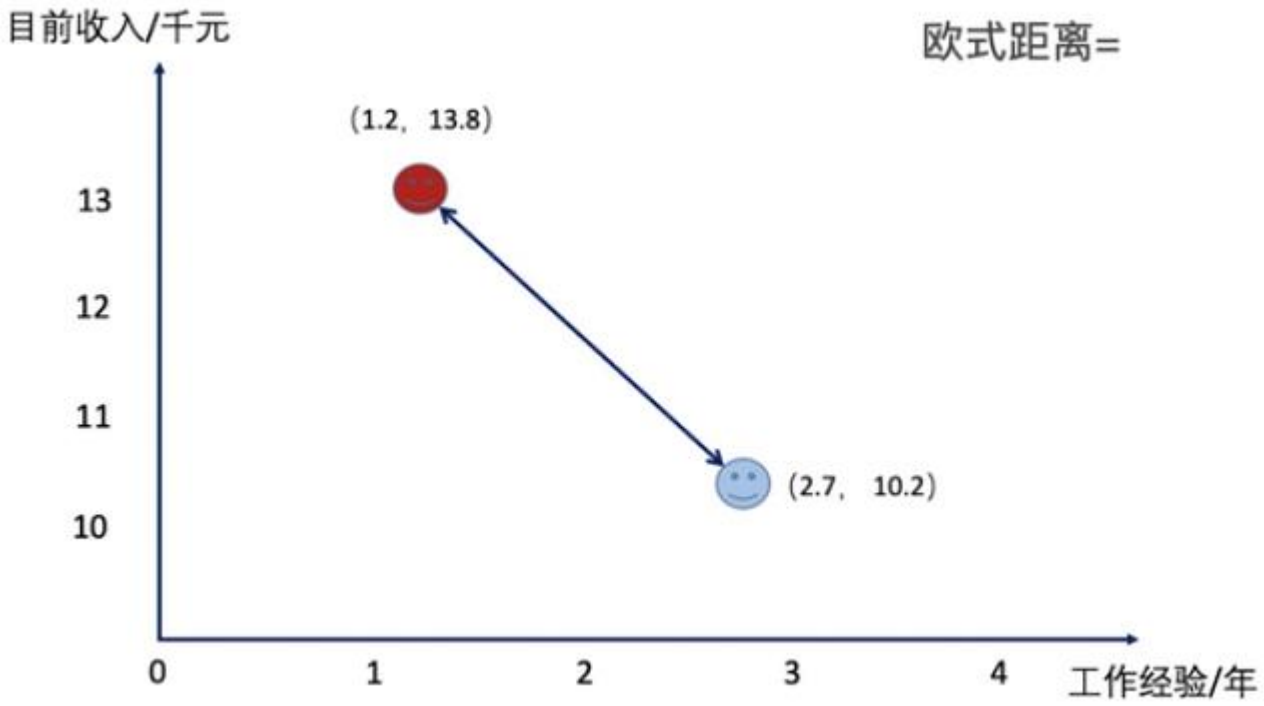
2. 标记好每个物体的标签

好人/坏人识别

水果种类识别

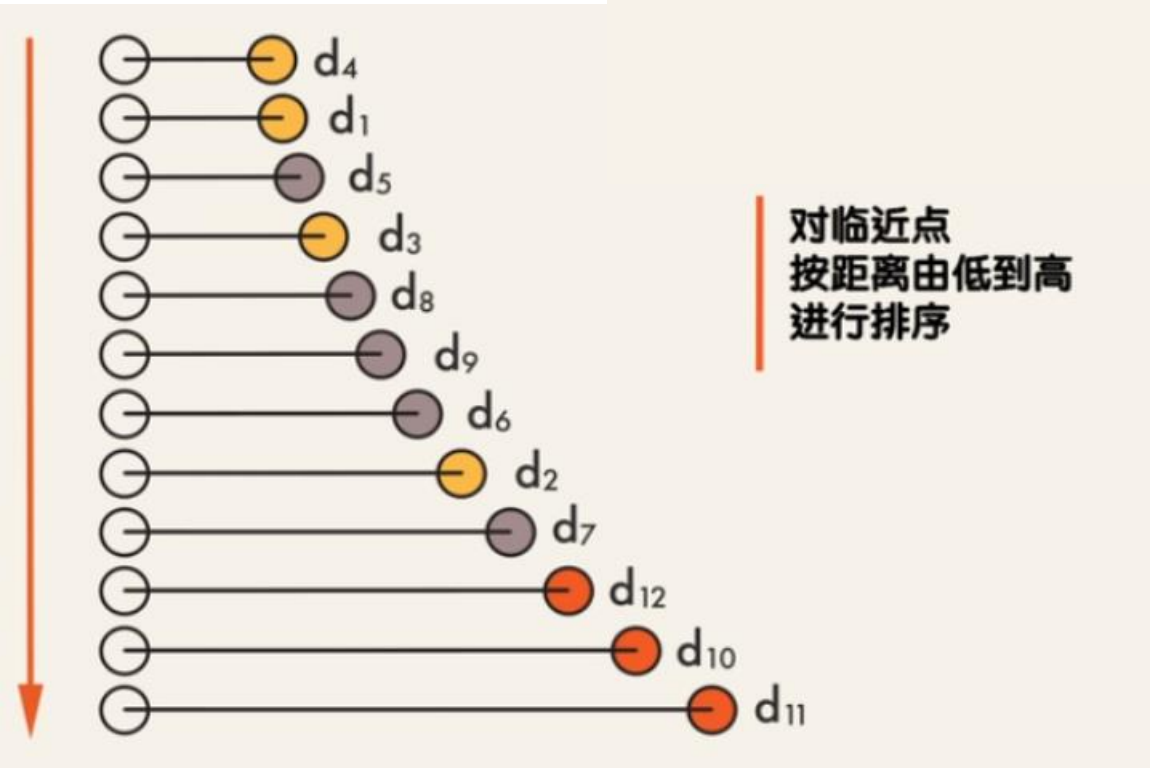
K近邻算法 实现步骤

3. 计算两个物体之间的距离/相似度



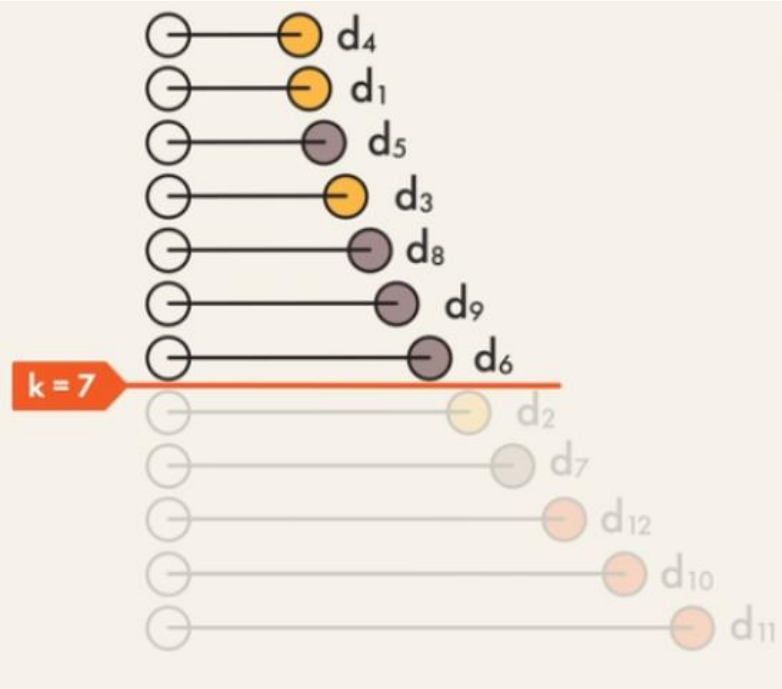
K近邻算法 实现步骤

4. 按距离排序



K近邻算法 实现步骤

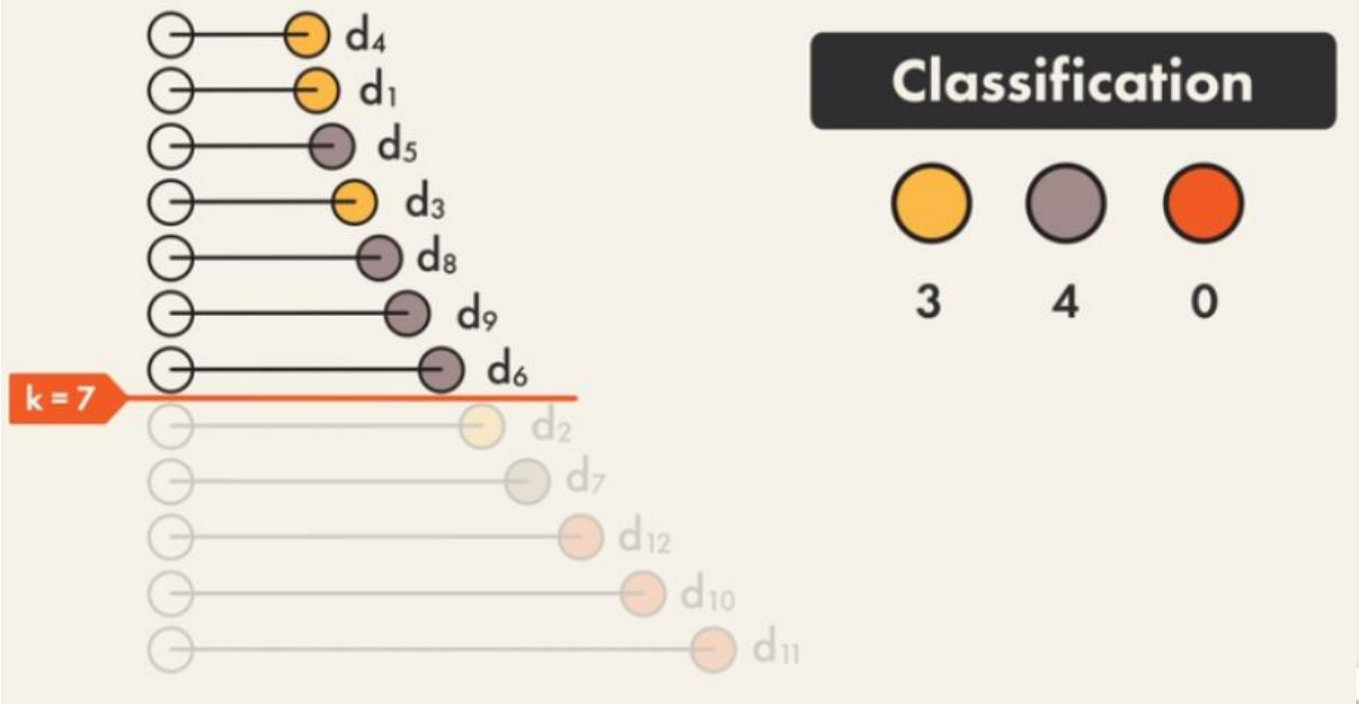
5. 选择合适的K



6. 选取K个距离最小的点

K近邻算法 实现步骤

7. 确定前K个点所在类别的出现频率



8. 前K个点所出现频率最高的类别最为当前点的预测分类

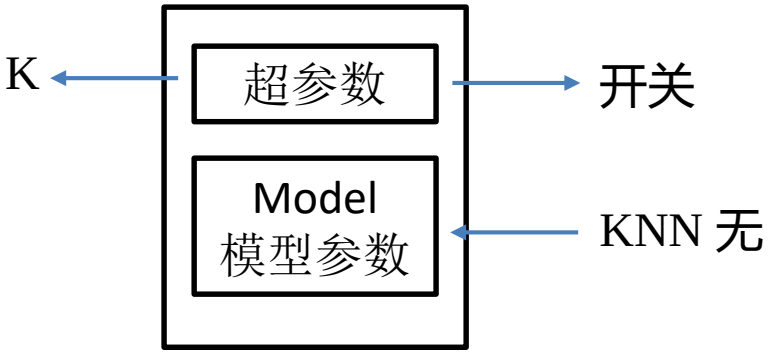
K近邻算法 三个要素

1. K值大小选择（数据交叉验证）
2. 距离的度量方法（欧氏距离等）
3. 分类决策规则（多数表决）

K值的选择

参数

- 模型参数（Model parameter） ← 训练数据
- 超参数（Hyper parameter） ← 不属于模型



K值的选择

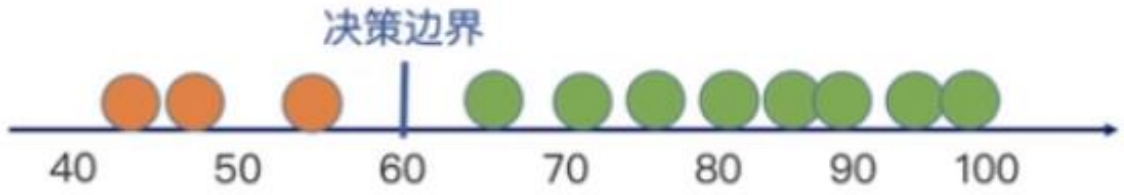
为了选择合理的K, 首先需要去理解**K对算法的影响**

↓

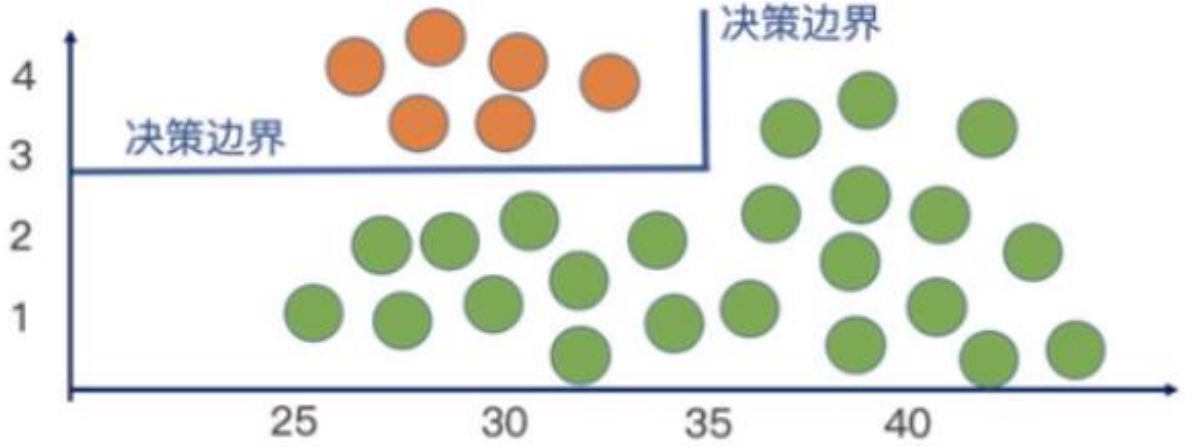
为了理解K对算法的影响，需要先理解什么叫算法的**决策边界**

决策边界

例： 大学里60分以上作为及格，60分以下作为不及格



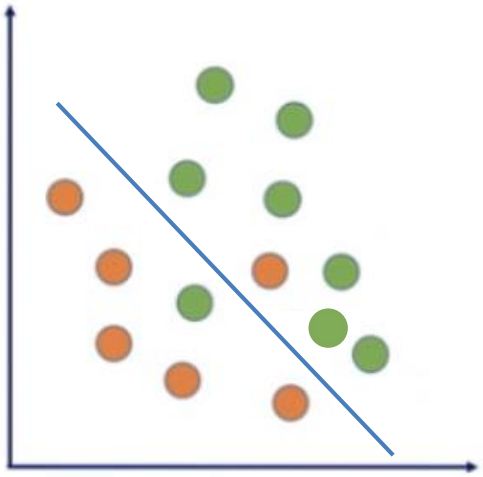
例： 今年某高校要计划引入35岁以下，具有海外研究经验3年以上的学者。



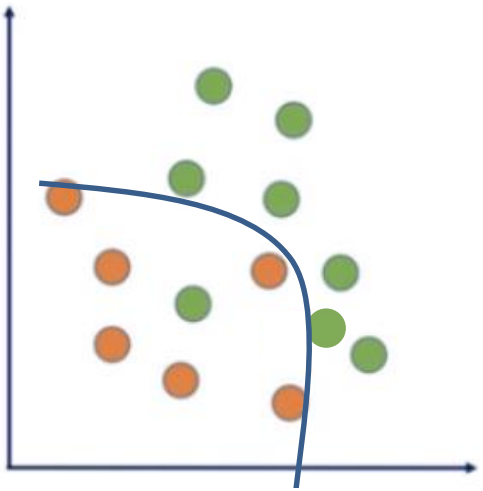
决策边界

- 准确
- 稳定

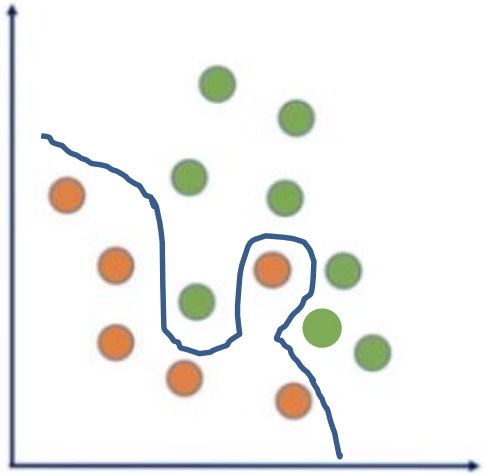
- 决策边界决定“线性分类器”或者“非线性分类器”



线性



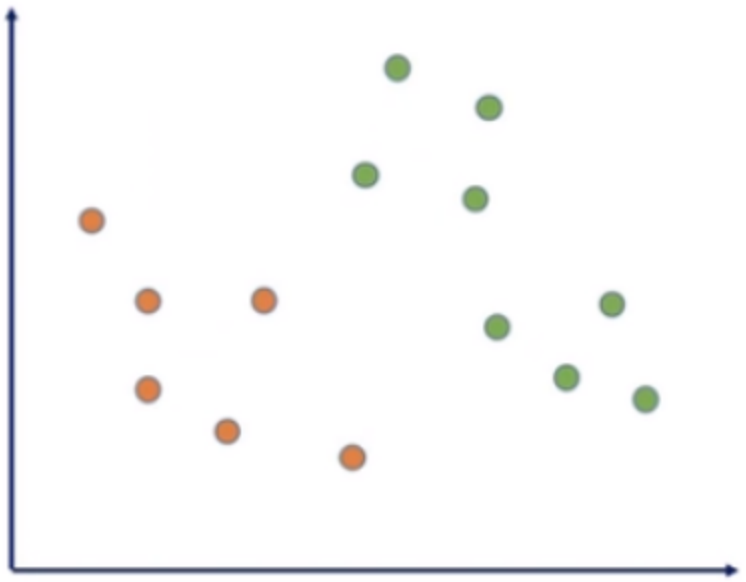
非线性



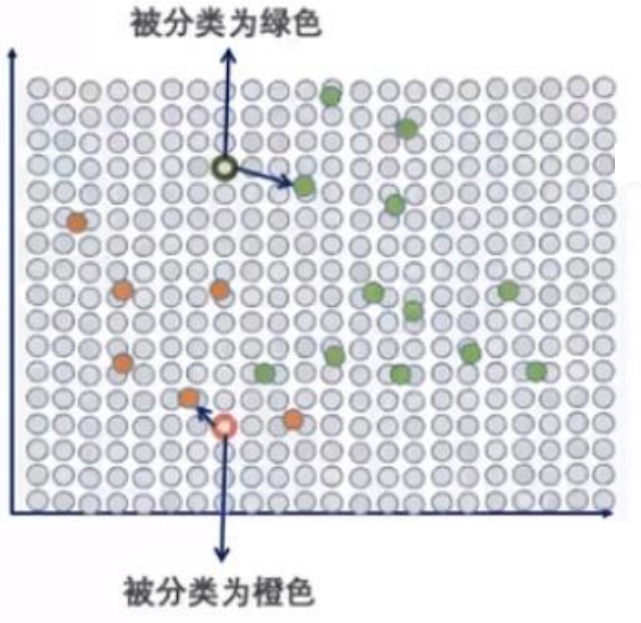
非线性

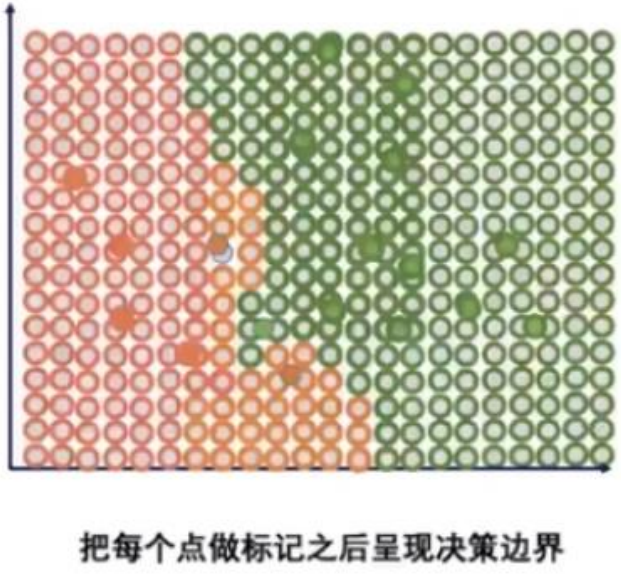
哪一个最好?

KNN的决策边界：怎么寻找（i.e.，K=1）？

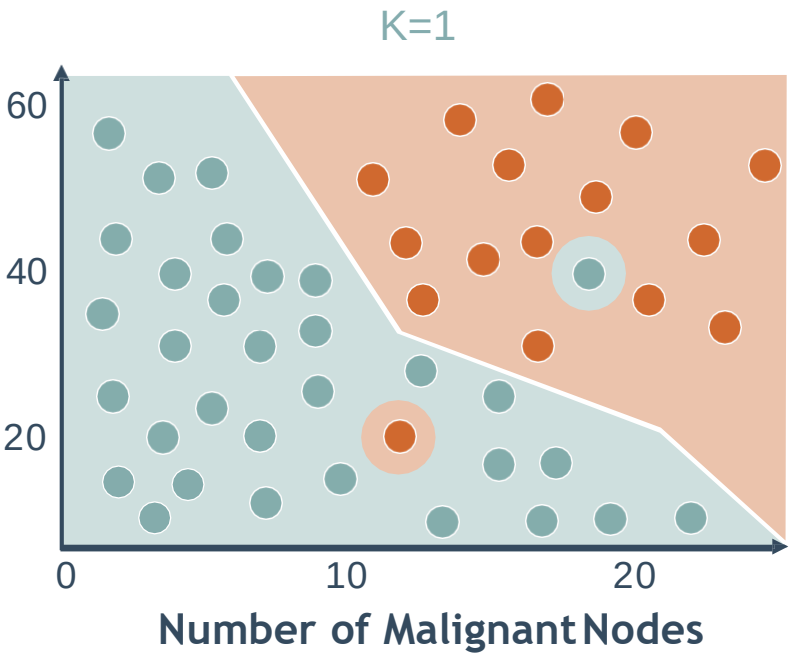


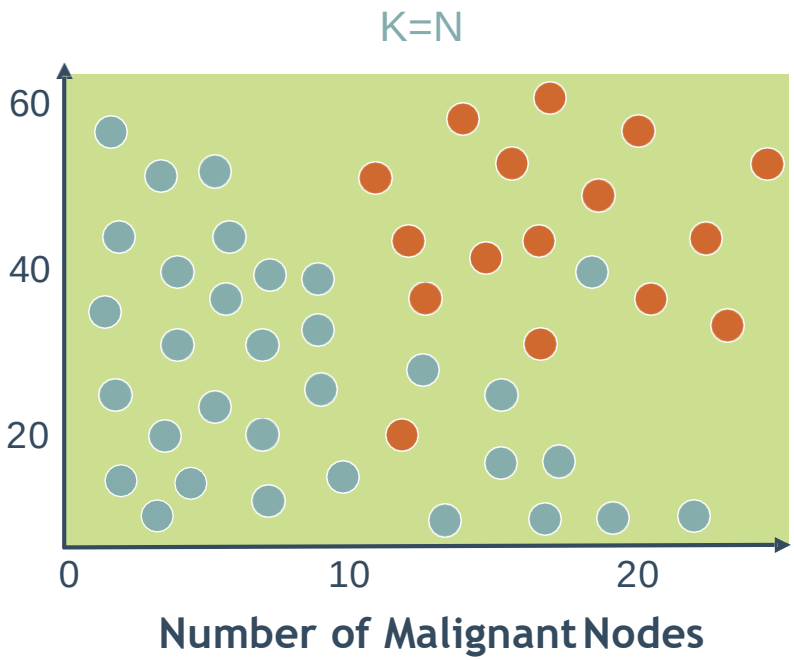
KNN的决策边界：怎么寻找（i.e.，K=1）？





K值的大小会影响决策/判定边界





- K增加 ⟶ 决策边界变得平滑

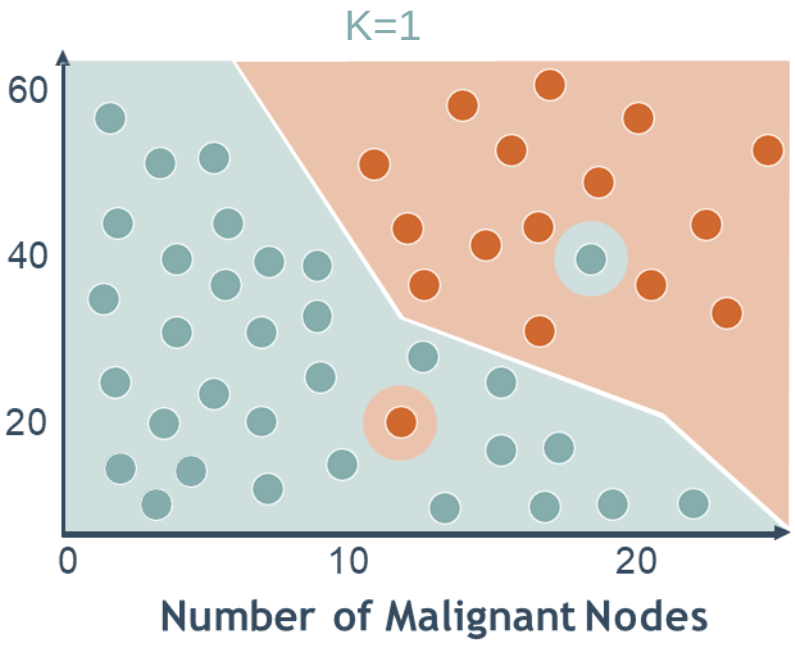
K值的选择

- 选择较小的K值，相当于用较小的领域内的训练实例进行预测，“学习”的近似误差会减小。只有与输入实例较近或相似的训练实例才会对预测结果起作用，与此同时带来的问题是“学习”的估计误差会增大。
  - 近似误差：对现有训练集的训练误差。如果近似误差过小可能会出现过拟合的现象，对现有的训练集能有很好的预测，但对未知的测试样本的预测将会出现较大的偏差。模型本身不是最接近最佳模型。
  - 估计误差：对测试集的测试误差。关注测试集，估计误差小说明对未知数据的预测能力好，模型本身最接近最佳模型。

近似误差小，容易过拟合 ←—在训练集上表现好，测试集上表现不好

估计误差好才是真的好！

K值的选择

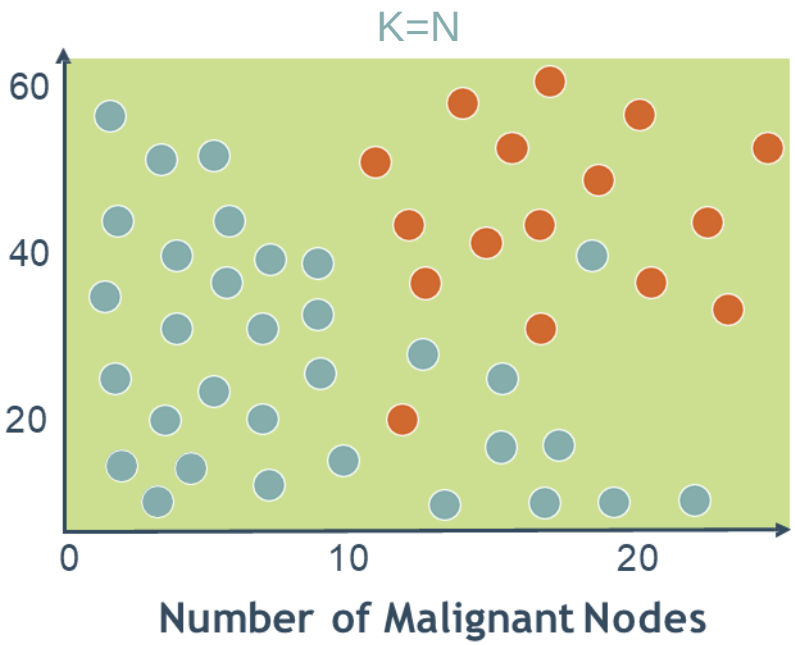


- K值的减小意味着整体模型变得复杂
- 容易受到异常点的影响
- 容易发生过拟合

K值的选择

- 选择较小的K值，相当于用较大的领域内的训练实例进行预测。优点是可以减少“学习”的估计误差，缺点是“学习”的近似误差会增大。此时，与输入实例较远或不相似的训练实例也会对预测结果起作用，使预测发生错误。
- K=N，完全不可取。此时无论输入实例是什么，都只是简单地预测它属于在训练实例中最多的类。模型过于简单，忽略了训练实例中大量的有用信息。

K值的选择



- K值的增大意味着整体模型变得简单
- 容易受到样本均衡的影响
- 容易发生欠拟合

K值的选择

K值过小：

- 容易受到异常点的影响
- 偏差低，方差大
- 过拟合

K值过大：

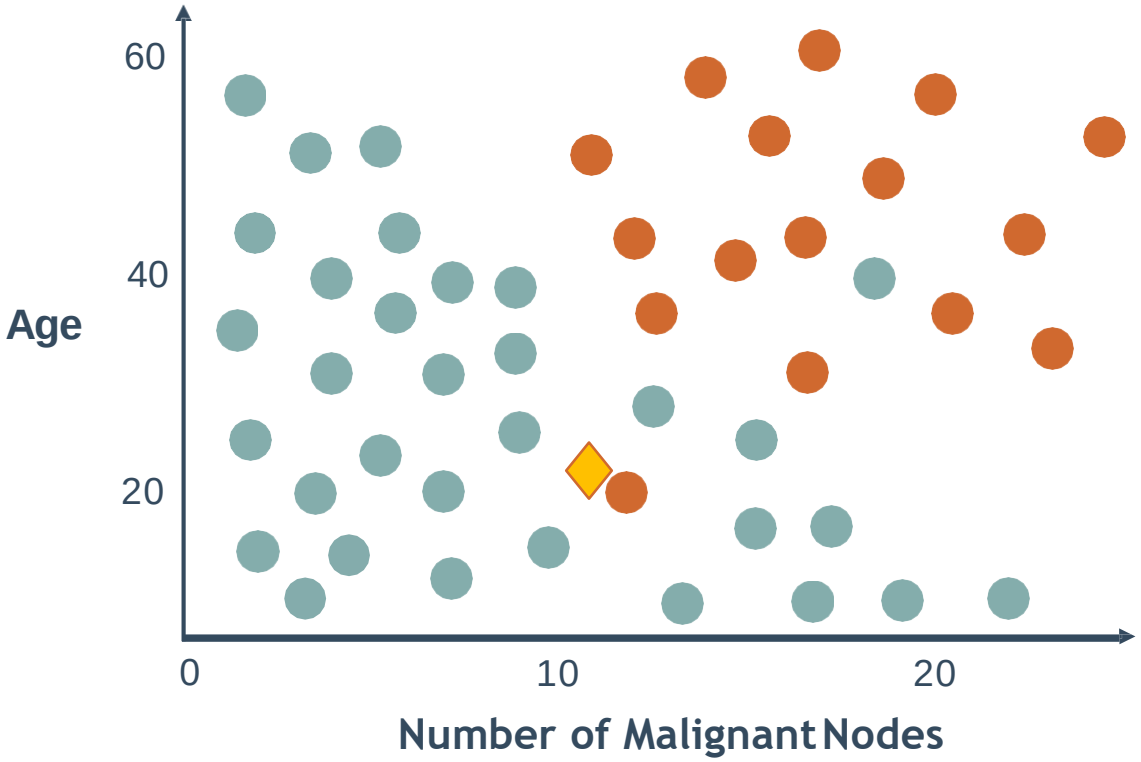
- 容易受到样本均衡的影响
- 偏差高，方差低
- 欠拟合

最好的K值：

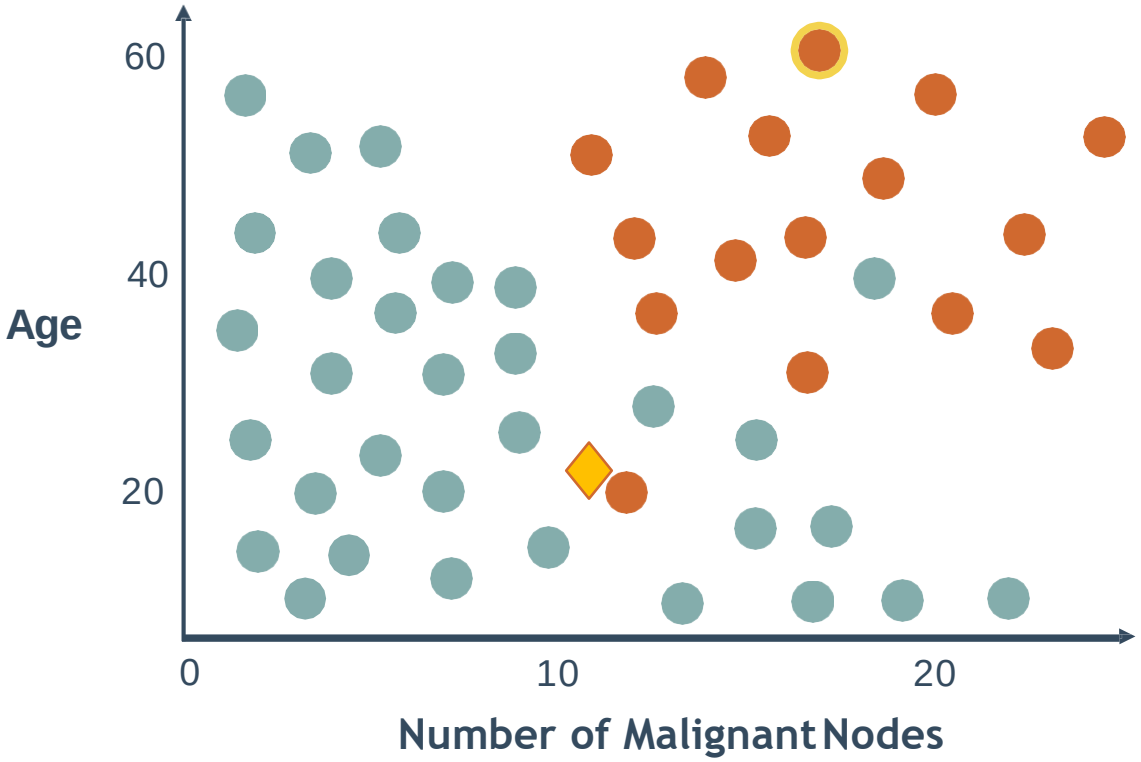
- 需要控制过拟合和欠拟合间的平衡

决定最佳K的方法——交叉验证将在下一章讨论

距离的度量



距离的度量



1. 欧氏距离（Euclidean Distance）：

也叫欧几里得距离， L2距离

二维：

$$d_{12}=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$$

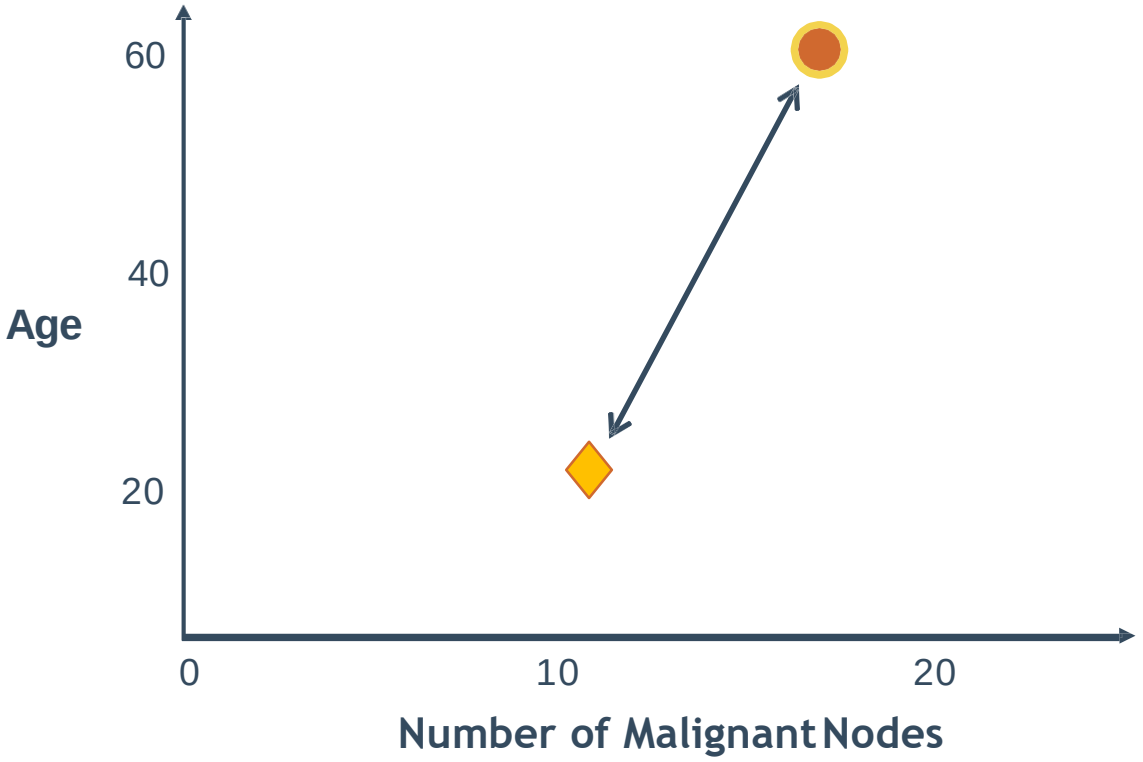
三维：

$$d_{12}=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}$$

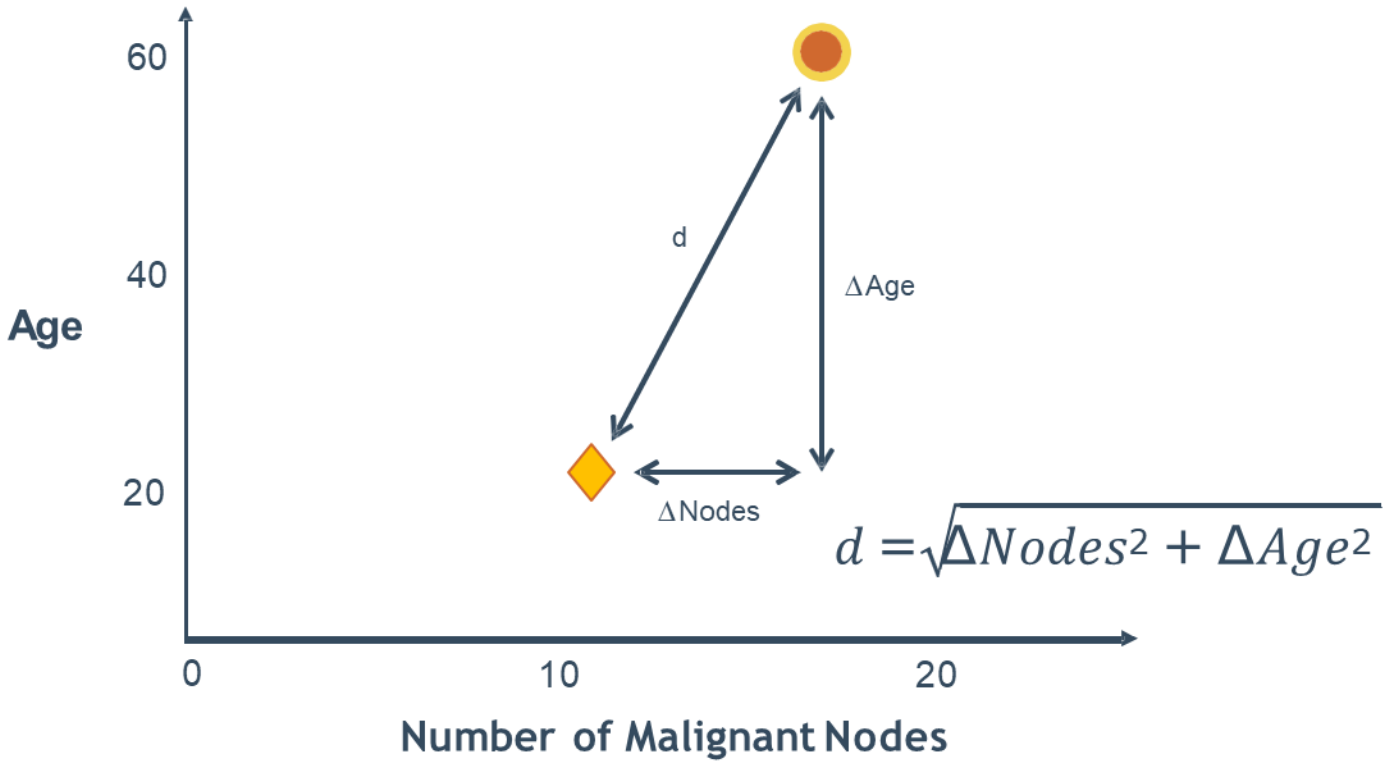
n维：

$$d_{12}=\sqrt{\sum_{k=1}^{n}(x_{1k}-x_{2k})^2}$$

1. 欧氏距离

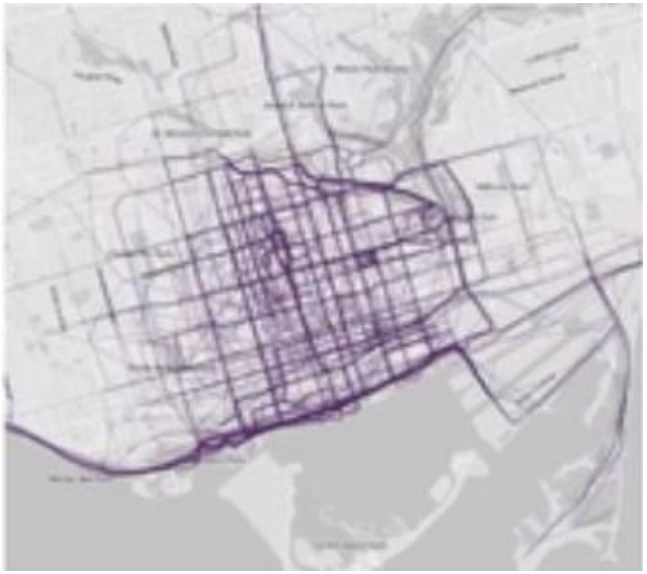


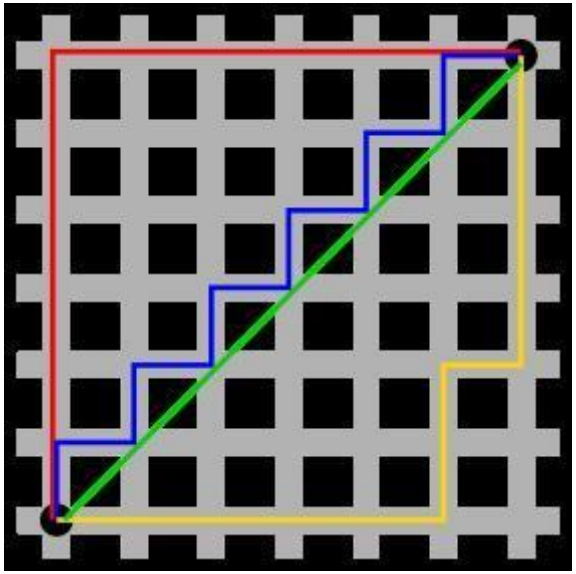
1. 欧氏距离


60
40
20
Age
d
ΔAge
ΔNodes
$d = \sqrt{\Delta Nodes^2 + \Delta Age^2}$
0
10
20
Number of Malignant Nodes

2. 曼哈顿距离（Manhattan Distance）：

也叫出租车距离，街区距离， L1距离

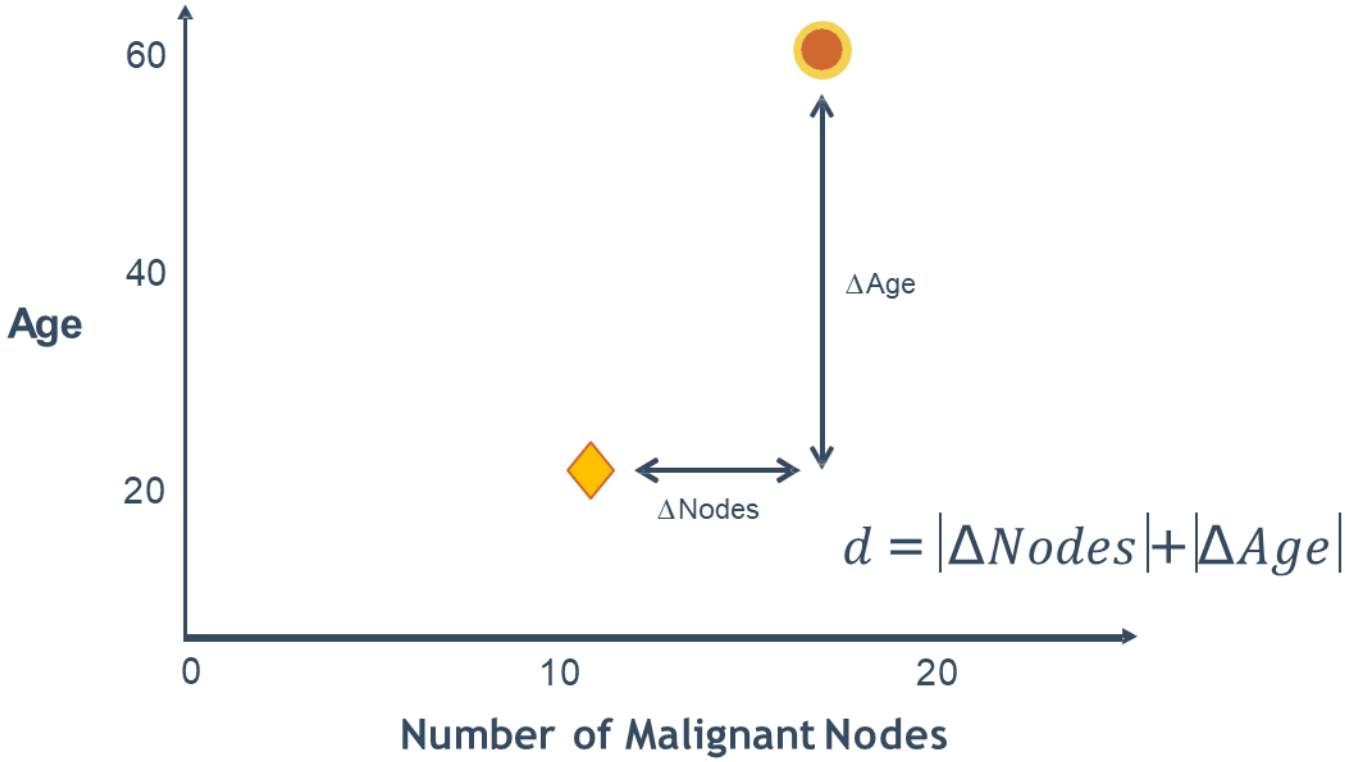




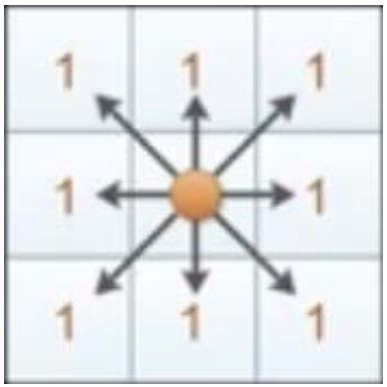
二维： $d_{12}= |x_1 - x_2| + |y_1 - y_2|$

n维： $d_{12}= \sum_{k=1}^{n} |x_{1k} - x_{2k}|$

2. 曼哈顿距离


60
40
20
Age
ΔAge
ΔNodes
$d = |\Delta Nodes| + |\Delta Age|$
0
10
20
Number of Malignant Nodes

3. 切比雪夫距离（Chebyshev Distance）：



二维： $d_{12}= \max(|x_1 - x_2|, |y_1 - y_2|)$

n维： $d_{12}=\max(|x_{1k} - x_{2k}|)$

4. 闵可夫斯基距离（Minkowski Distance）：

$$d_{12}=\sqrt[p]{\sum_{k=1}^{n}|x_{1k}-x_{2k}|^{p}}$$

其中p是一个变参数：

当p=1时，曼哈顿距离；

当p=2时，欧氏距离；

当$p\to\infty$时，切比雪夫距离；

根据的p不同，闵氏距离可以表示某一类/种距离。

5. 标准化欧氏距离（Standardized Euclidean Distance）： 也叫加权欧氏距离

假设样本集X的均值(mean)为m，标准差(standarddeviation)为s，那么X的“标准化变量”表示为：

$$X^*=\frac{X-m}{s}$$

标准化变量的数学期望为0，方差为1。

因此样本集的标准化过程(standardization)用公式描述就是：

标准化后的值 = (标准化前的值 - 分量的均值) /分量的标准差

经过简单的推导就可以得到两个n维向量a(x11,x12,…,x1n)与b(x21,x22,…,x2n)间的标准化欧氏距离的公式：

$$d_{12}=\sqrt{\sum_{k=1}^{n}\left(\frac{x_{1k}-x_{2k}}{s_k}\right)^2}$$

6. 余弦距离（Cosine Distance）：

二维空间中，向量*A*(x1,y1)与*B*(x1,y1)的夹角余弦：

$$\cos(\theta)=\frac{x_1x_2+y_1y_2}{\sqrt{{x_1}^2+{y_1}^2}\times\sqrt{{x_2}^2+{y_2}^2}}$$

n维：

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}.$$

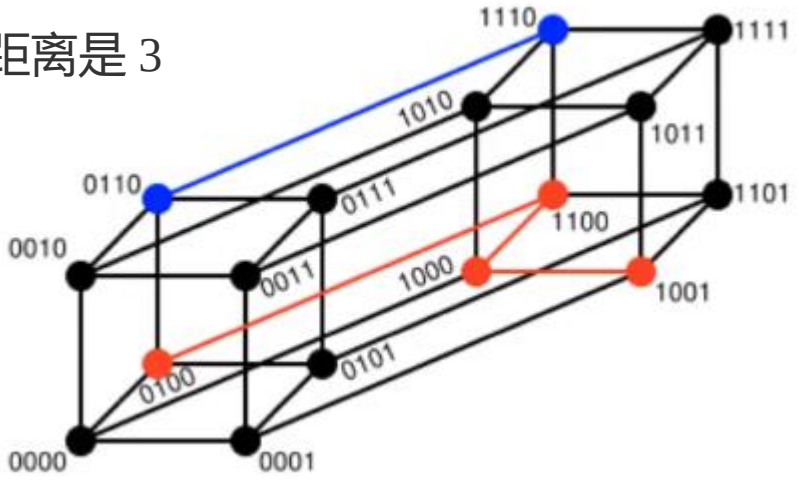
7. 汉明距离（Hamming Distance）：

两个等长字符串的汉明距离为：将一个字符串变换成另外一个字符串所需要替换的字符个数：

1011101 与 1001001 之间的汉明距离是 2

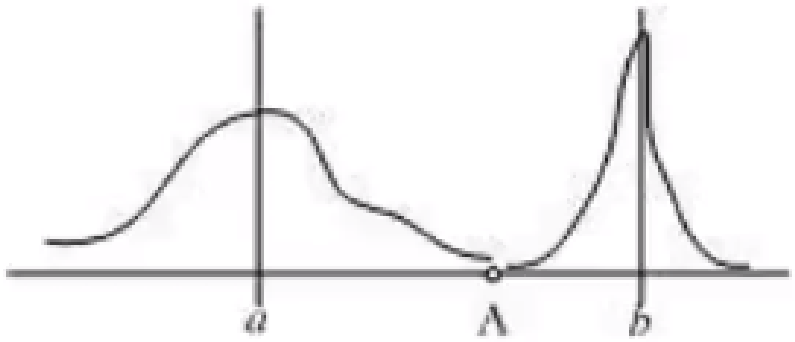
2143896 与 2233796 之间的汉明距离是 3

"toned" 与 "roses" 之间的汉明距离是 3



8. 马氏距离（Mahalanobis Distance）：

两个正态分布图，均值为a和b，方差不一样。图中A点离哪个总体更近？或者说A有更大的概率属于谁？



显然，A点离左边更近，A属于左边总体的概率更大，尽管A与a的欧氏距离远一些。这就是马氏距离的直观解释。

8. 马氏距离（Mahalanobis Distance）：

马氏距离是由印度统计学家马哈拉诺比斯 (P. C. Mahalanobis)提出的，表示点与一个分布之间的距离（协方差距离）。它是一种有效的计算两个未知样本集的相似度的方法。与欧氏距离不同的是，它考虑到各种特性之间的联系（例如：一条关于身高的信息会带来一条关于体重的信息，因为两者是有关联的），并且是尺度无关的(scale-invariant)，即独立于测量尺度。对于一个均值为$\mu=(\mu_1, \mu_2, \mu_3, \ldots, \mu_p)^T$，协方差矩阵为$\Sigma$的多变量向量$x=(x_1, x_2, x_3, \ldots, x_p)^T$，其马氏距离为 $D_M(x)=\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$

如果协方差矩阵为单位矩阵，马氏距离就简化为欧氏距离；如果协方差矩阵为对角矩阵，马氏距离可称为正规化的欧氏距离。

$$d(\vec{x},\vec{y})=\sqrt{\sum_{i=1}^{p}\frac{(x_i-y_i)^2}{\sigma_i^2}}$$

σ_i是x_i的标准差

马氏距离是基于样本分布的一种距离。