IDC

Machine Learning
introduction

华中科技大学计算机学院
王天江

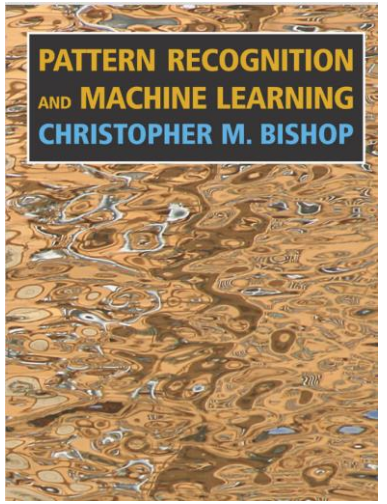
2024年11月13日星期三1时42分19秒
2010-10-11

IDC研究室

参考教材

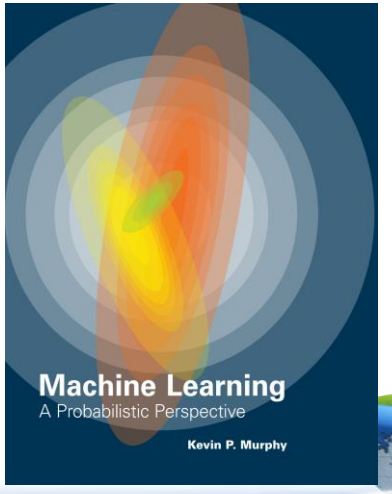
- Pattern Recognition and Machine Learning

  Christopher M. Bishop
  Springer



- Machine Learning A Probabilistic Perspective

  Kevin P Murphy
  The MIT Press Cambridge, Massachusetts London, England



机器学习课QQ群：

王天江-华中科技大学 邀请您参加腾讯会议
会议主题：机器学习
会议时间：2021/09/15 16:00-18:00
(GMT+08:00) 中国标准时间 - 北京
点击链接入会，或添加至会议列表：
https://meeting.tencent.com/dm/hB0haCd8mrdy
会议 ID：883 185 207
复制该信息，打开手机腾讯会议即可参与

群名称:机器学习课
群　号:618556482

第一章 机器学习概论

Machine learning introduction

……

contents

什么是机器学习

What are machine learning

……

机器学习定义

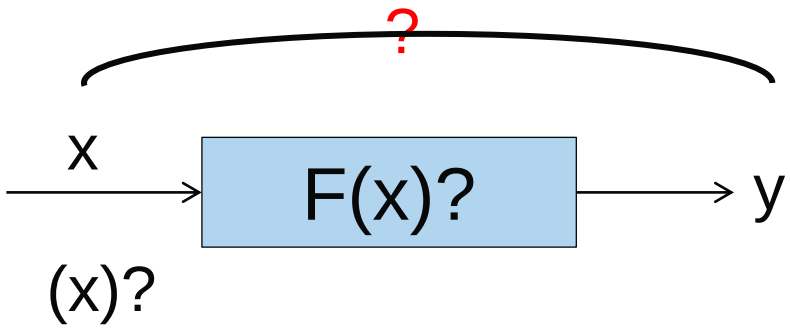
1. **如果一个计算机程序满足下面几条，则称它向经验 E学习：**
    1. 程序面向任务T运行,
    2. 定义一个度量方法 P,
    3. 加入经验 E 后发现程序效果得到了提高.　　T,
2. **机器学习定义为能实现下面要求的一组方法**
    1. 自动检测数据中的模式
    2. 用已掌握的模式来预测未来的数据
    3. 在不确定性条件下执行其他类型的决策

概率的视角看机器学习

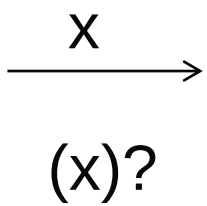
1. 将所有未知量视为随机变量
2. 它是在不确定性条件下进行决策的最佳方法
3. 概率建模是一种语言
   - 多数其他科学技术领域都在使用
   - 它为这些不同领域提供了统一的框架
   - 通过这种观点，将在机器学习中所做的事情与其他学科联系起来，比如像：
     - ⑩ 随机优化、
     - ⑩ 控制理论、
     - ⑩ 运筹学、
     - ⑩ 计量经济学、
     - ⑩ 信息论、统计物理学或者生物统计学

机器学习的类型

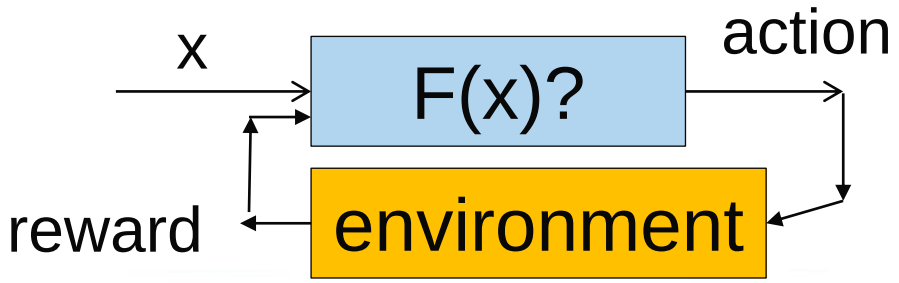
- 监督学习：
  - 强监督学习
  - 弱监督学习



- 无监督学习：
  - 无监督学习有时又称为知识发现.



- 增强学习（强化学习）



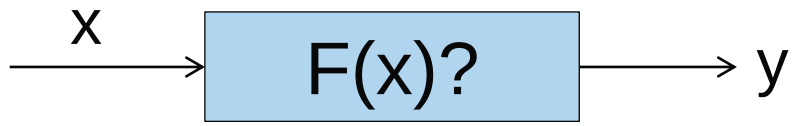
监督学习

Supervised Learning

……

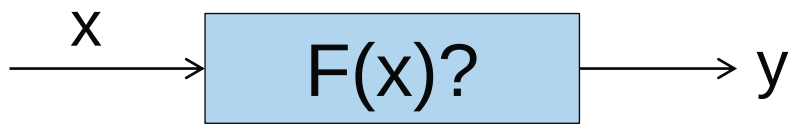
监督学习的概念

- 对于任务 T
  - 要学习一个从输入 $x \in X$ 到输出 $y \in Y$ 的映射 f
  - 输入 x 又称为**特征**
- 经验 E 表示为：
  - 输入-输出二元组的集合：D={ (x1,y1), (x1,y2),…(xN,yN)}
- 度量 P 依赖于预测输出的类型



监督学习包括两大问题

- **分类**：输出y是一个有限离散值, $y \in \{1, 2, \ldots, N\}$
  - 这种离散输出y，又称为类标签
  - 在给定输入的情况下预测类标签的问题也称为模式识别.
  - 如果问题仅有两类，$y \in \{0, 1\}$ 或者 $y \in \{-1, +1\}$, 又称为二分类问题(**binary classification)**
- **回归**：输出y是一个连续变量，$y \in [R1,R2]$



分类问题

Classification Problem

……

一个监督学习的例子

- 三种鸢(yuan)尾花:

Setosa

Versicolor

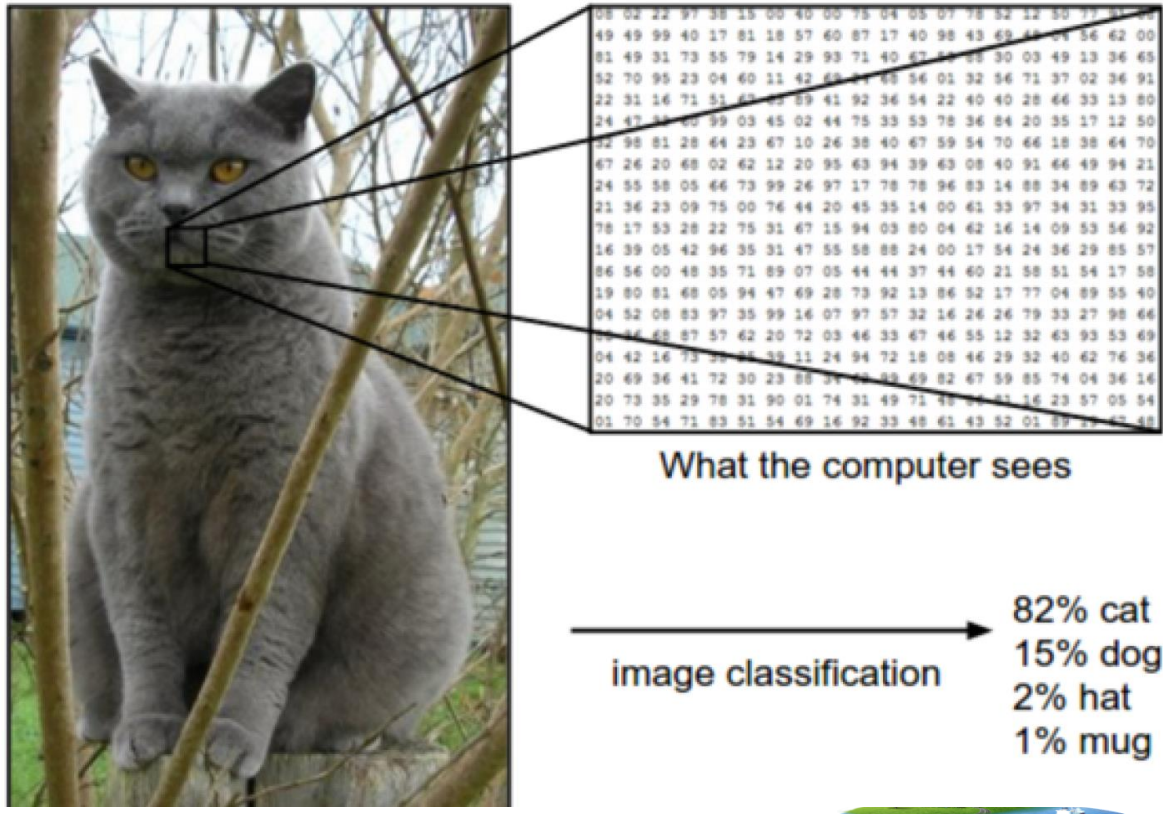
Virginica

- 将鸢尾花按种类分开
  - $Y = \{1,2,3\}$

Index	萼片长度	萼片宽度	花瓣长度	花瓣宽度	label
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
…					
50	7.0	3.2	4.7	1.4	Versicolor
…					
149	5.9	3.0	5.1	1.8	Virginica

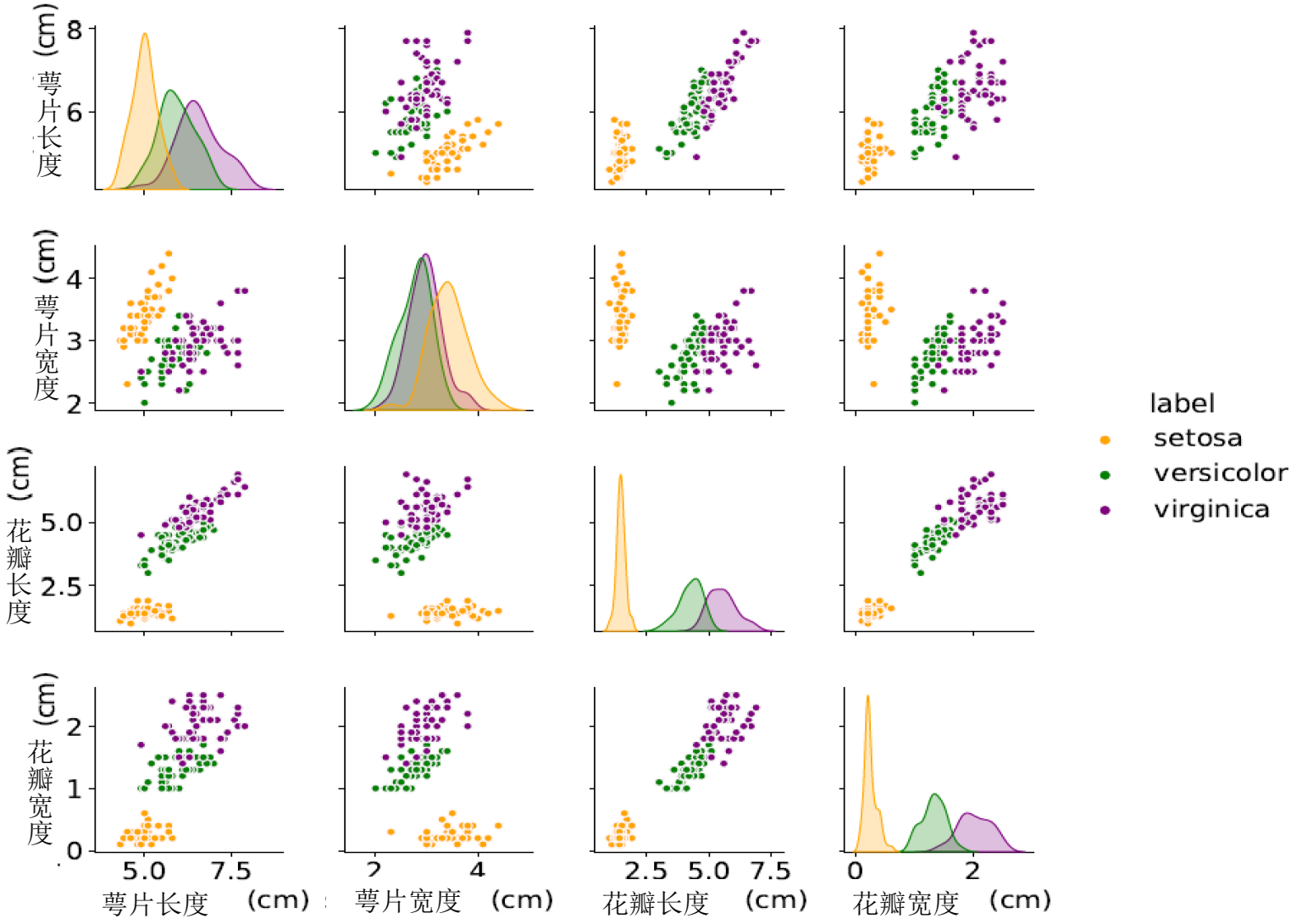
一个图像分类的例子

- 图像集合X构成一个高维空间：
  - 彩色图像有3个通道，Channels = 3 (e.g., RGB)
  - 每个通道有像数：D1 × D2 个
  - 一张图像有像数：D = C × D1 × D2
  - 图像集合X就构成空间：$X = R^D$
  - 像数在 {0,1,..,255}范围内取值
- 给定一个图像集，已知每张图片的类别：
  - 猫、狗、帽子、杯子
- 任意给定一个图像，将它分为 4 类中的某类：
  - 猫、狗、帽子、杯子



探索性数据分析

- 首先对数据进行分析是有益的
  - 低维数据可用二维数据图
- 例如，鸢尾花数据
  - 二维散点图：{p(x,y)}
  - 对角线为每类特征的边缘分布
- 对于高维数据
  - 通常首先做降维

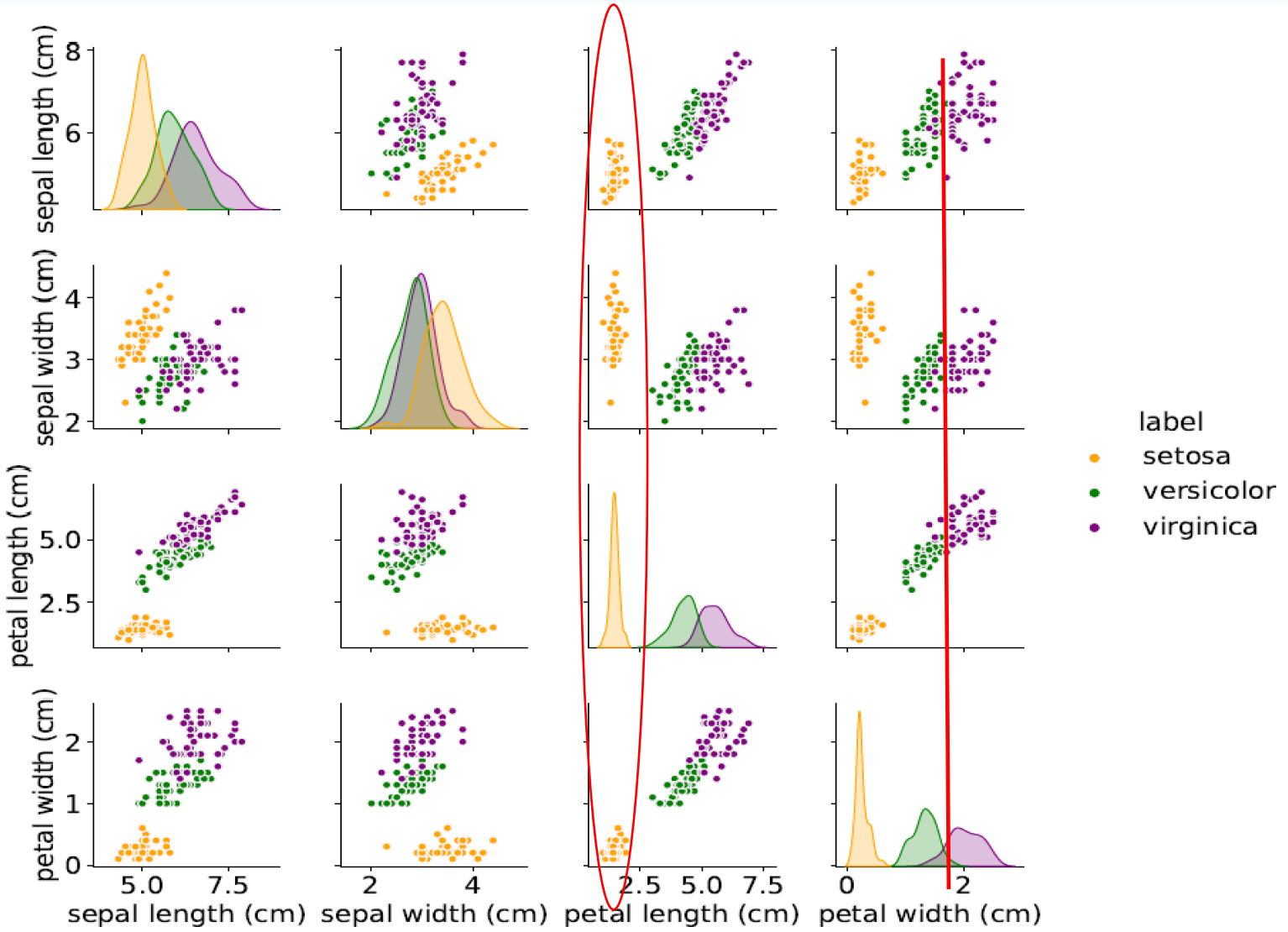


学习一个分类器

- 从图中容易看出
  - Setosa花与其他两类花区别明显
- 构造一个简单的分类器:

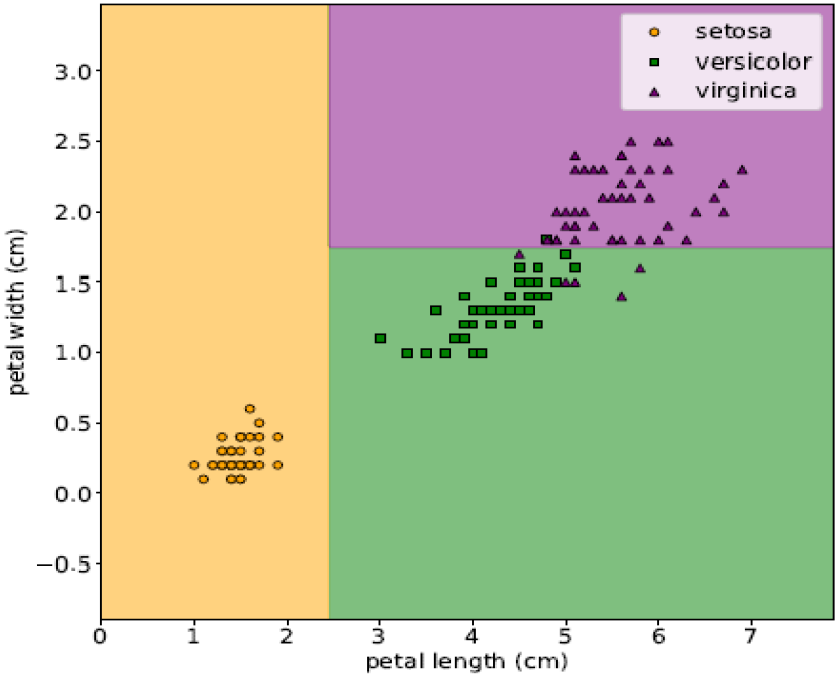
$$f(x,\theta)=\begin{cases}\text{Setosa} & \text{若花瓣长度}<2.45\\ \text{Versicolor 或 Virginica} & \text{否则}\end{cases}$$

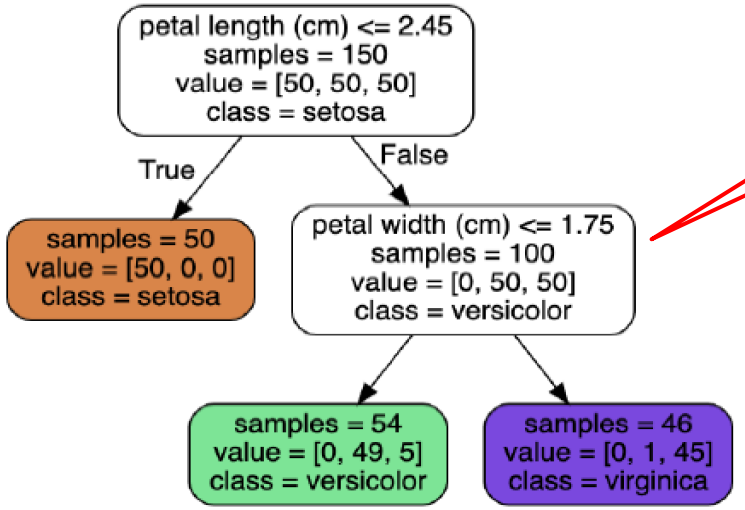
- 决策边界
  - 花瓣长度 $=2.45$



进一步区分另两类花

- 基于深度为2的决策树
  - 只使用花瓣长度与花瓣宽度两个特征





错分了6个样本

错分率

- 错分率是度量分类模型效果的常用方法
  - 错分率定义：

$$L(\theta) = \frac{1}{N}\sum_{n=1}^{N} I(y_n \neq f(x_n, \theta))$$

- 指示函数

$$I(e) = \begin{cases} 1 & if\ e\ is\ true \\ 0 & if\ e\ is\ false \end{cases}$$

经验风险

- 定义
  - 经验风险是：分类器在训练集上的平均损失

$$L(\theta) = \frac{1}{N}\sum_{n=1}^{N} l(y_n \neq f(x_n, \theta))$$

- 一种损失函数
  - 0-1 损失：

$$l_{01}(y, \hat{y}) = I(y \neq \hat{y})$$

鸢尾花分类的损失矩阵

		Estimate		
		Setosa	Versicolor	Virginica
Truth	Setosa	0	1	1
	Versicolor	1	0	1
	Virginica	10	10	0

经验风险最小化

- 模型拟合或训练模型问题的一种定义方法
  - 找到一组参数，使训练集上的经验风险最小化

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} L(\theta) = \underset{\theta}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} l(y_n \neq f(x_n, \theta))$$

- 但我们的真正目标：模型能够泛化，不仅在训练集上表现好
  - 极小化模型在未知数据上的预期损失。

不确定性

- 许多情况下，无法完全预测给定输入的确切输出
  - 由于缺乏输入输出映射的知识（认知不确定性或模型不确定性）
  - 由于映射中的内在随机性（随机不确定性或数据不确定性）
- 可以使用条件概率分布来捕捉不确定性：

$$p(y = c|\boldsymbol{x}; \boldsymbol{\theta}) = f_c(\boldsymbol{x}; \boldsymbol{\theta})$$

  - 其中，$f : \boldsymbol{X} \rightarrow [0, 1]^C$ 将输入映射到C个输出标签的概率分布.
  - 对于每个c，有 $0 \leq f_c \leq 1$, 并且 $\sum_{c=1}^{C} f_c = 1$

softmax 函数

- 如果希望把向量$\mathbf{a}=\{a_1,\ldots,a_C\}$看成是一个概率分布，可以使用softmax函数对它做归一化

$$S(\boldsymbol{a}) = \left[\frac{e^{a_1}}{\sum_{c=1}^{C} e^{a_1}}, \cdots, \frac{e^{a_C}}{\sum_{c=1}^{C} e^{a_C}}\right]$$

  - 其中，$S(\mathbf{a}) : \mathbf{R}^C \rightarrow [0, 1]^C$ 将输入映射成满足概率约束的向量

$$0 \leq S(a)_c \leq 1 ,$$

$$\sum_{c=1}^{C} S(a)_c = 1$$

- 对于我们构造的概率分布函数 f(x; θ)，为了让它确实概率分布约束，可以定义为：

$$p(y = c|\boldsymbol{x}; \boldsymbol{\theta}) = S_c(f(\boldsymbol{x}; \boldsymbol{\theta}))$$

最大似然估计

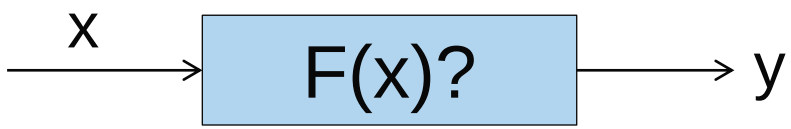
- 似然函数：给定了输入样本，输出值的概率就成了参数的函数 $p(y_n = c|\boldsymbol{x_n}; \boldsymbol{\theta}) = f_c(\boldsymbol{x_n}; \boldsymbol{\theta})$
- 一种常用的拟合概率模型的损失函数
  - **基于负log概率的损失函数：** $l(y, f(\boldsymbol{x}; \boldsymbol{\theta})) = -\log p(y, f(\boldsymbol{x}; \boldsymbol{\theta}))$
- **负log 似然**
  - 训练集的平均负log概率 (经验风险):

$$\mathrm{NLL}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N} \log p(y_n, f(\boldsymbol{x_n}; \boldsymbol{\theta}))$$

- 最小化这个NLL，等价于最大似然估计:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, NLL(\theta) = \underset{\theta}{\operatorname{argmax}} \frac{1}{N}\sum_{n=1}^{N} logp(y_n, f(x_n, \theta))$$

分类方法



1. **判别函数模型：** 学习 F(x)？

2. **概率产生式模型：** 学习 $p(x,y)$，通过贝叶斯公式得到 $p(y|x)=\frac{p(x|y)p(y)}{p(x)}$

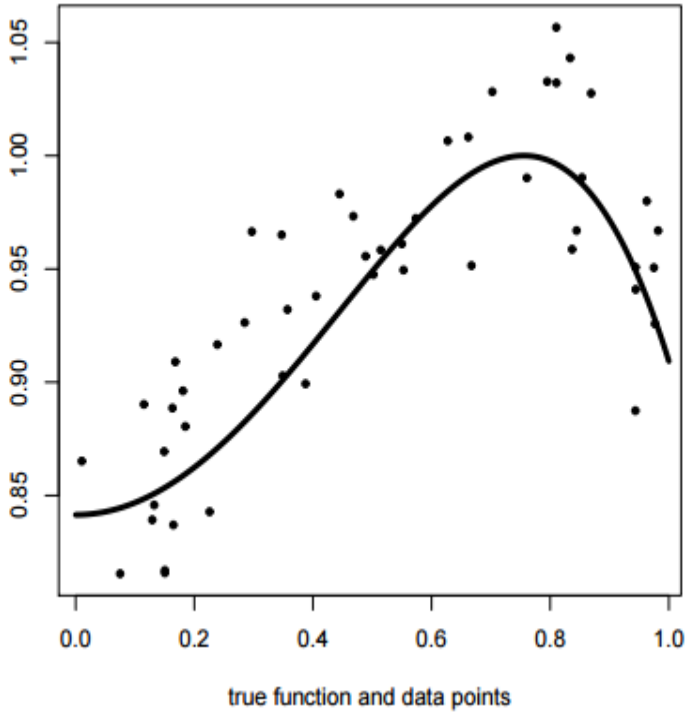
3. **概率判别式模型：** 直接学习 p(y / x)=f(y，x)?

回归问题

Regression

回归的思路

- 回归是指给定输入后，预测一个实数输出量 y∈R
  - 构造一个回归函数：$\hat{y} = f(x_n, \theta)$
  - 回归很类似于分类，但它的输出是实数
    - ⑩ 我们必须使用不同的损失函数



- 常用的损失函数：二次损失函数，或者 l_2 损失函数:

$$l_2(y, \hat{y}) = (y - \hat{y})^2$$

- 经验风险等于均方误差:

$$\mathrm{MSE}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y})^2$$

- 求均方误差最小，则可求出回归函数的各个参数

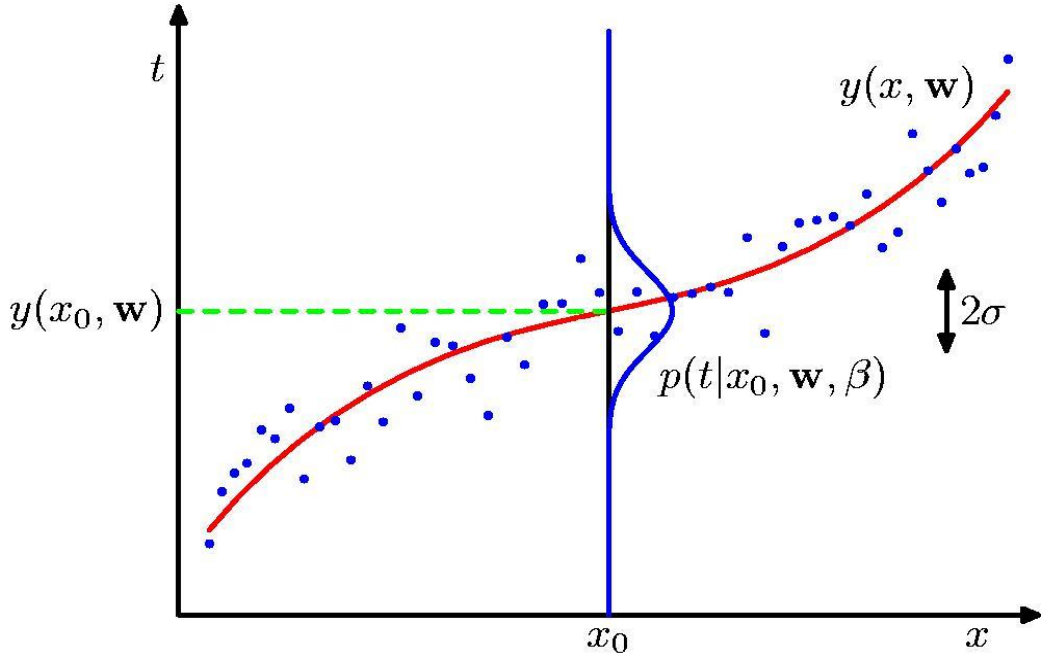
概率视角的均方误差

- 为了描述模型的不确定性，增加一个噪声变量w

$$y = f(\boldsymbol{x};\boldsymbol{\theta}) + \omega;\ \ \omega \sim N(\omega|0,\sigma^2)$$

- 因此输出y也变成了随机变量

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = N(y|\mu,\sigma^2) = N(y|f(\boldsymbol{x};\boldsymbol{\theta}),\sigma^2)$$



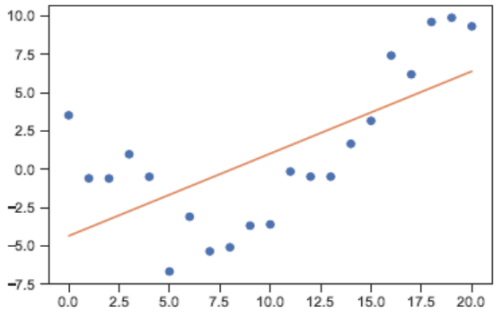
- 负 log 似然变成为

$$\mathrm{NLL}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N}\log\left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} exp\left(-\frac{1}{2\sigma^2}\left(y_n - f(x_n,\theta)\right)^2\right)\right] = \frac{1}{2\sigma^2}MSE(\theta) + \mathrm{const}$$

线性回归

- 线性回归函数形如：
  - $\emptyset(x)$: 基函数(特征提取器)

$$f(x, \theta) = \mathbf{w}^T \phi(x) = \sum_{1 \le i \le D} w_i \varphi(x_i)$$

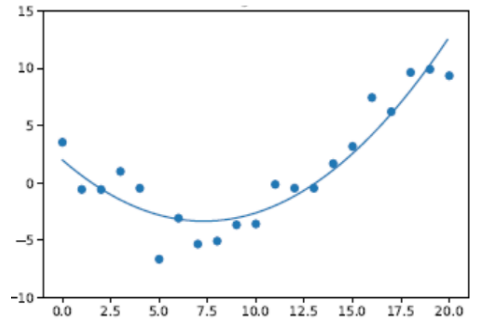


- 简单线性回归：

$$f(\mathbf{x}, \theta) = \mathbf{w}^T \mathbf{x} = \sum_{1 \le i \le D} w_i x_i$$

- 多项式回归：

$$f(x, \theta) = w_0 + w_1 x + w_2 x^2 + \cdots + w_d x^d$$



基于深度神经网络的回归

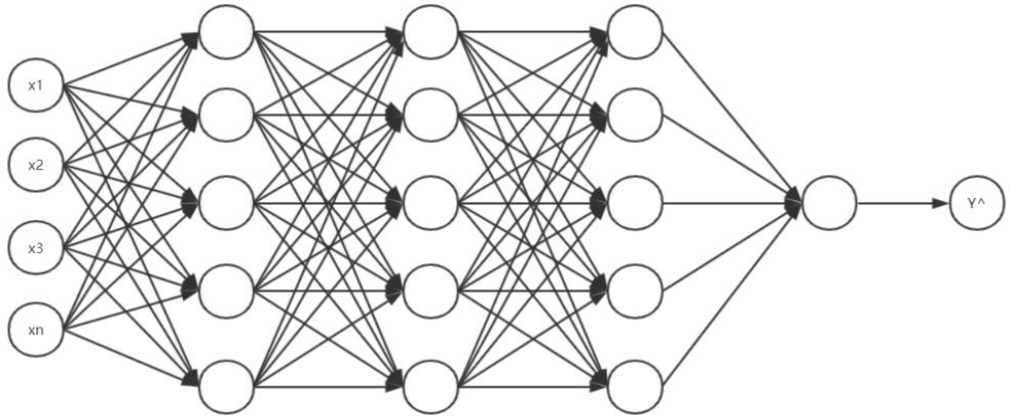
- 假设特征提取器 ∅(x) 有自己的一套参数：

$$f(x, \omega, V) = \omega^T \phi(x, V)$$

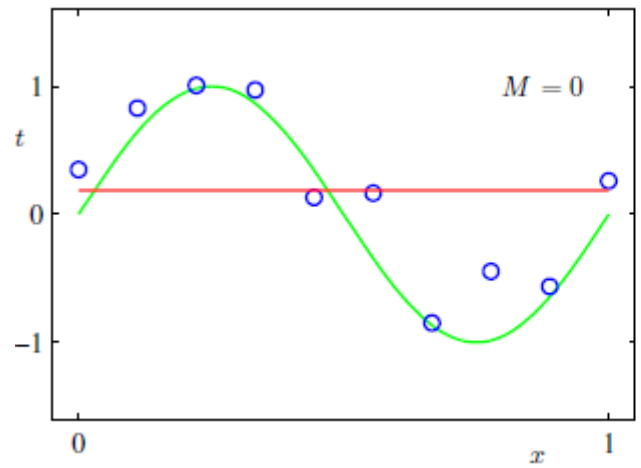
- 深度神经网络的关键思路
  - 递归地分解Φ(x;V) 为更简单的函数的组合

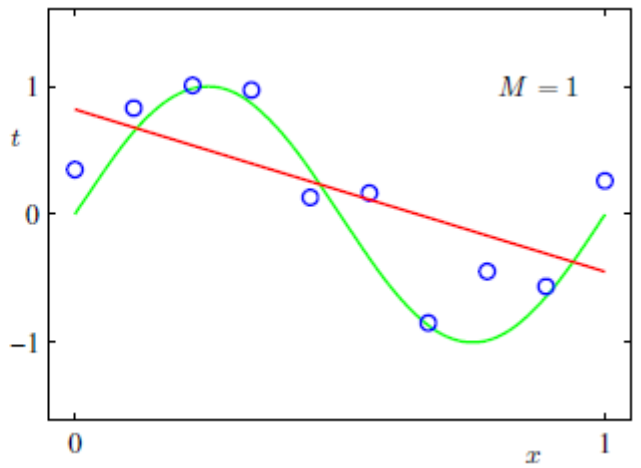
❿模型变成了有L层嵌套的函数栈：

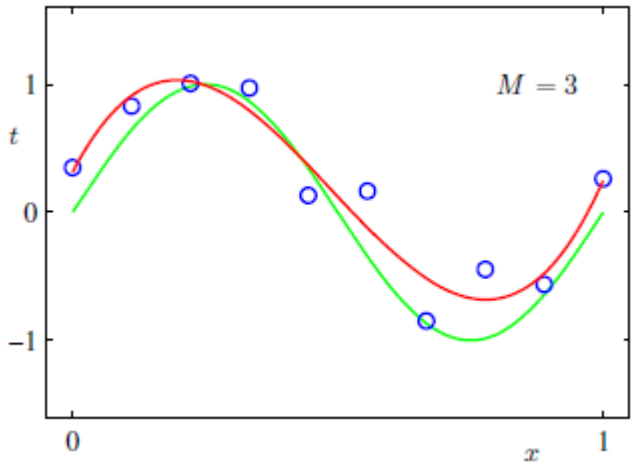
$$f(\mathbf{x}, \theta) = f_L(f_{L-1}(\ldots(f_1(x))\ldots))$$

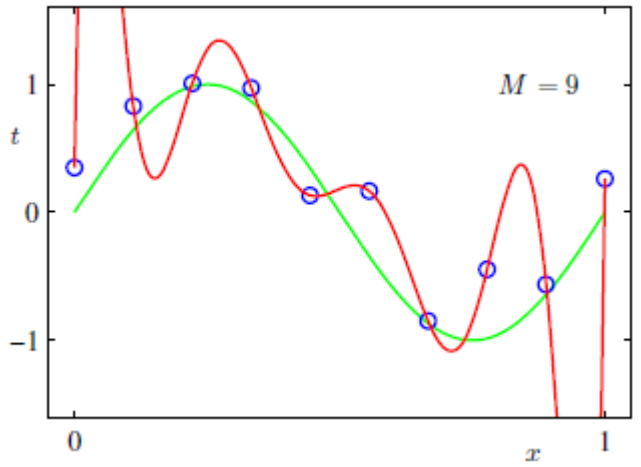


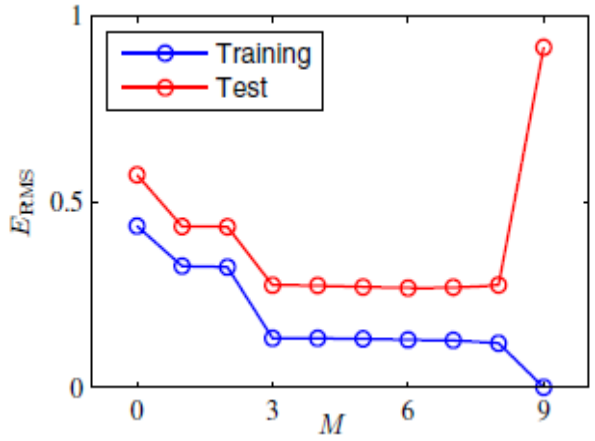
欠拟合、过拟合与泛化


M = 0
t
x
1
0
−1


M = 1
t
x
1
0
−1


M = 3
t
x
1
0
−1


M = 9
t
x
1
0
−1


Training
Test
E_{RMS}
M
1
0.5
0
0
3
6
9

如何选择合适的模型

- 所有模型都是有错误的，但有些模型是有用的
- 我们都想知道哪一款模型最好。
  - 不幸的是，没有哪个单一模型能对所有问题都是最优
- 某个领域效果很好的模型在另一个领域可能就不行。
- 那么，如何选择模型呢？
  - 基于领域知识
  - 反复试验
  - 贝叶斯方法

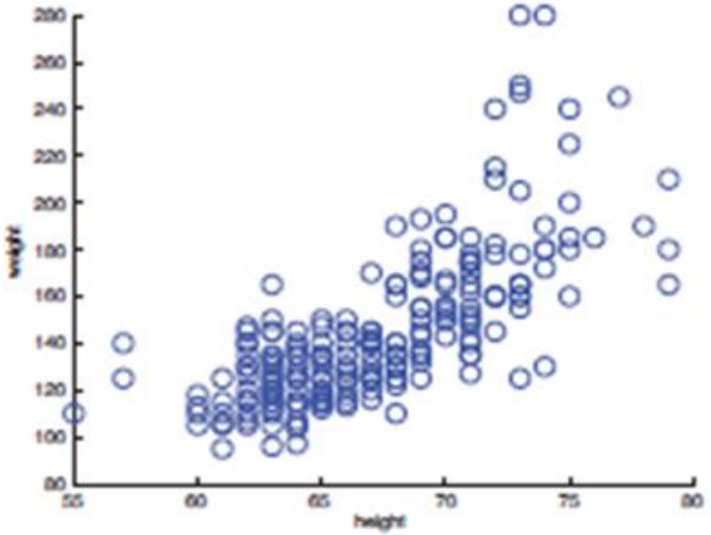
无监督学习

Unsupervised leanring

……

无监督学习的目的

- 一件非常有意义的事情是试图“理解”数据
  - 我们只观测到“输入” $D = \{x_n : n = 1 : N\}$，没有任何“输出” y_n。
- 从概率的角度来看，无监督学习的任务：
  - 拟合形如p(x)的模型，
  - 该模型可以生成新的数据x



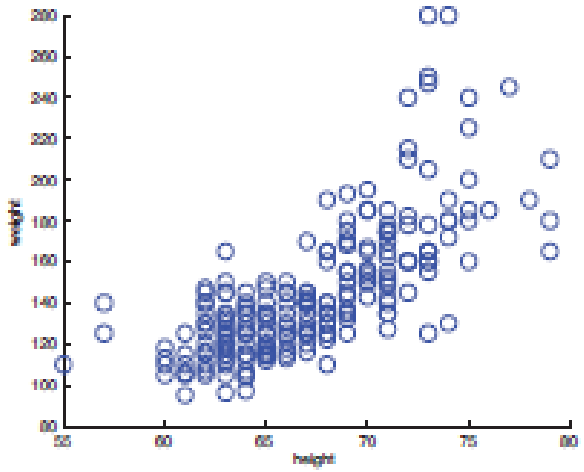
无监督学习的特点

- 不需要收集大型标记过的训练集
- 不需要学习如何将问题划分为若干种类别。
- 它要求模型
  - 要“解释”高维输入数据，而不关注低维输出。
- 这使我们能够学习到更好的关于“问题本质”的模型。

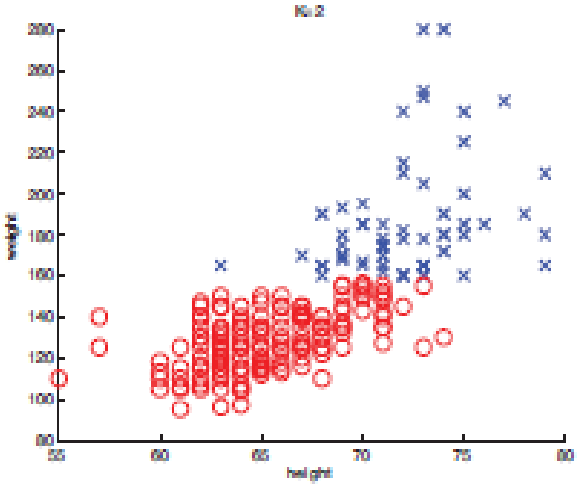
聚类（Clustering）

- 无监督学习的一个任务就是：在数据中进行聚类
  - 目标是将输入数据划分为包含“相似”点的多个区域。
- 这可能有聚类不正确的情况
  - 因此，我们需要考虑模型复杂性和数据聚类正确性之间的权衡

- 比如：图中表示
  - 一个人群身高/体重的数据集
  - 数据点不带如何类标签



(a)



(b)

发现潜在因素

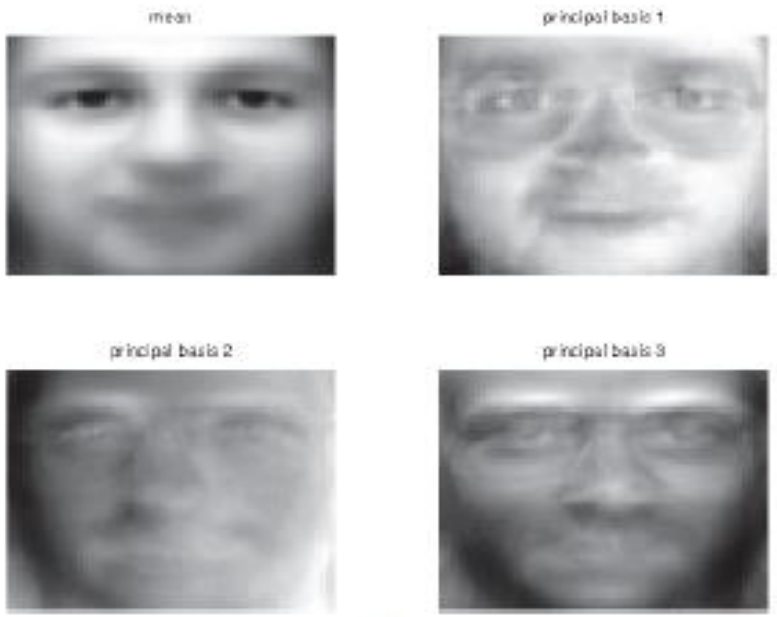
- 降低数据的维度通常是有用的
  - 它能捕捉到数据的“本质”。
- 方法之一就是：假设
  - 高维数据 $x_n \in R^D$ 是由低维潜在因素（**latent factors**） $z_n \in R^K$ 生成的
  - 粗略地，可将模型表示成: $z_n \rightarrow x_n$
- 线性模型
  - 因素分析: $p(x_n|z_n;\boldsymbol{\theta}) = N(x_n|wz_n+\mu,\sum)$， （$\mathbf{x}=w\mathbf{z}+\mu+\epsilon$, $\epsilon \sim N(0, \sum)$）
  - 主成分分析 (PCA): $p(x_n|z_n;\boldsymbol{\theta}) = N(x_n|wz_n+\mu, \sigma^2\mathbf{I})$
- 非线性扩展: $p(x_n|z_n;\boldsymbol{\theta}) = N(x_n|f(z_n,\boldsymbol{\theta}), \sigma^2\mathbf{I})$
  - 这类模型训练比较复杂
  - 一些近似方法能较好的解决问题，比如，变分自动编码器

PCA算法生成的图像

- 图像说明，一个主要成分就能够表示数据的主要特征.

(a)

随机选择的 64 × 64 像素人脸图像

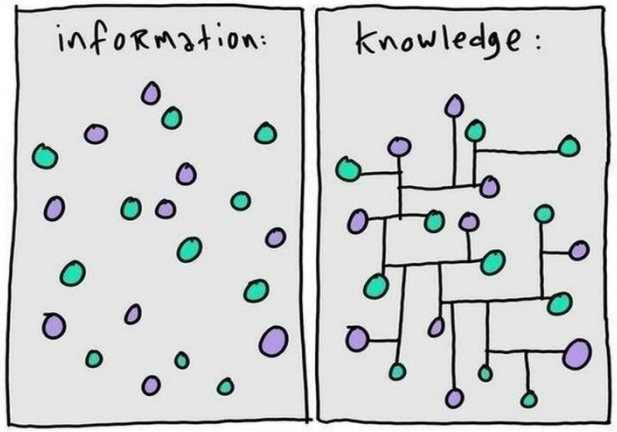


(b)

均值及前3个主成分基向量 (特征脸)

发现图结构

- 对于一组变量，我们发现哪些变量与其他变量的相关性最强。
  - 这种相关性可以用图G来表示
- 我们从数据中学习这个图结构
- 学习稀疏图有两个主要应用：
  - 为了发现新知识，
  - 为了获得更好的联合概率密度估计。
- 一个例子是预测高速公路上的交通堵塞。



图像修复

- 目标是“填补”孔洞
  - 例如，我们对图像去噪，以及估计遮挡后面的像素。
- 一种办法是
  - 构造像素的联合概率模型
  - 给定一组完整的图像，
  - 在给定已知变量（像素）的情况下推断未知变量（像素（pixels））

    。

自监督学习

- 自监督学习是无监督学习中的一种类型
  - 数据没有标签信息，利用数据本身的某些信息作为监督信息进行学习.
- 如何从数据本身找出可以利用的监督信息
  - 构造辅助任务（pretext），从大规模的无监督数据中挖掘自身的监督信息
  - 构造辅助任务的几类方法
    - ⑩基于上下文：比如，输入的语言序列，遮住其中某个词，将这个词作为监督信息
    - ⑩基于时序：比如，输入的视频序列，前后帧之间具有相似性，将这种相似性作为监督信息
    - ⑩基于对比：比如，构建正样本和负样本，对比正负样本的区别，将这种区别作为监督信息

无监督学习的评价

- 评价无监督学习方法的效果是非常困难的事情
  - 因为无监督学习根本没有真实的信息，只有样本数据
- 一种常用的评价无监督模型的方法
  - 度量模型生成的联合概率分布
  - 这种方法是将问题看成是一个密度估计问题

$$L(\boldsymbol{\theta}, \boldsymbol{D}) = -\frac{1}{|D|}\sum_{x \in D} \log p(x, \theta)$$

- 该方法的思路，一个好的模型应该具有：
  - 在数据空间中，产生了数据样本的区域应具有较大的概率，其他区域只能是小概率
  - 产生的样本的分布应该与输入样本的分布应该非常相似

另一种无监督学习评价方法

- 利用有监督学习方法，来评价无监督学习方法
  - 将无监督学习的结果作为特征输入到某个监督学习模型中
  - 如果无监督方法获得好的效果，则后面的监督方法就应该会使用更少的标记数据。

强化学习

Reinforcement leanring

.....

增强学习

- 强化学习是机器学习的方法之一，
  - 针对智能体（agent）在与环境的交互中进行学习的问题
- 强化学习的一般过程
  - 智能体可以观察环境
  - 智能体可以采取能够影响环境的行动
  - 智能体知道它的行为如何影响环境
    - ❿ 通过一个奖励函数
  - 智能体调整自己的行为，使得对环境影响不断向好

参数与非参数化模型

Parametric vs non-parametric models

……

参数化模型

- 对于形如$p(y|x)$ *或* $p(x)$的概率模型，参数化模型：
  - 模型的解析表达式已定
  - 具有固定数量的参数，
  - 例如，高斯分布：$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp(\frac{(1-\mu)^2}{2x^2})$
- 优点和缺点：
  - 其优点是使用起来往往更快，
  - 其缺点是对数据分布的性质做了过多的假设

非参数化模型

- 对于形如$p(y|x)$ *或* $p(x)$的概率模型，非参数化模型：
  - 模型的解析表达式未定
  - 模型参数的数量会随着训练集的规模增加而增加，
- 优点和缺点：
  - 其优点是更灵活，
  - 其缺点是对于大数据集，通常难以计算

一个简单的非参数化分类模型

A Simple non-parametric classifier

……

K最近邻模型(KNN)

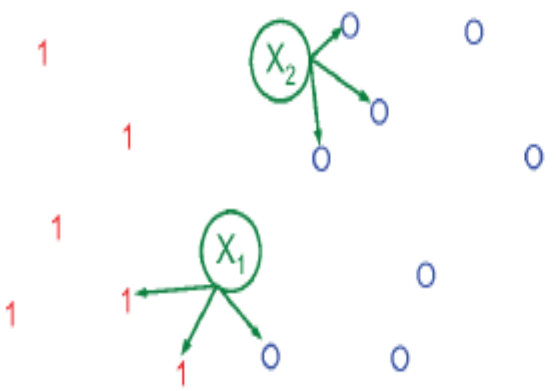
- “查看”训练集中离测试数据***x***最近的K个点，
  - 看这些邻近点都属于哪一类，来决定测试数据***x***属于哪一类
  - 通过指示函数来构造模型：

$$p(y=c\mid x,D,K)=\frac{1}{K}\sum_{i\in N_K(x)}I(y_i=c)\qquad I(e)=\begin{cases}1 & \text{if } e \text{ is true}\\ 0 & \text{if } e \text{ is false}\end{cases}$$

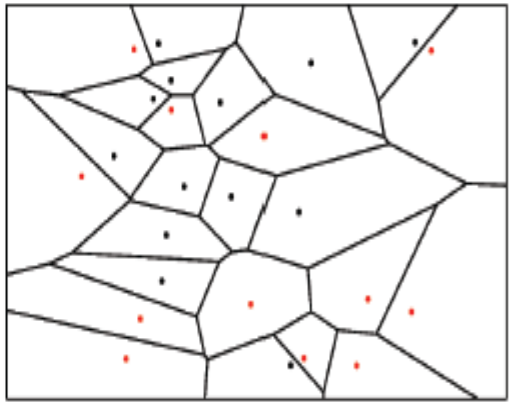
- 两大重要问题
  - 如何度量“接近度”
  - 如何选择K才是合适的？

KNN模型举例

- 图中，x1、x2是测试数据
  - 图a，K=3
  - 图b， K=1



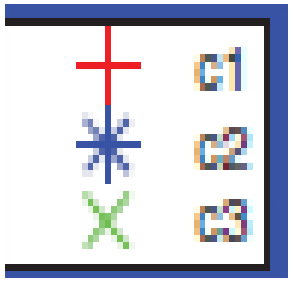
(a)

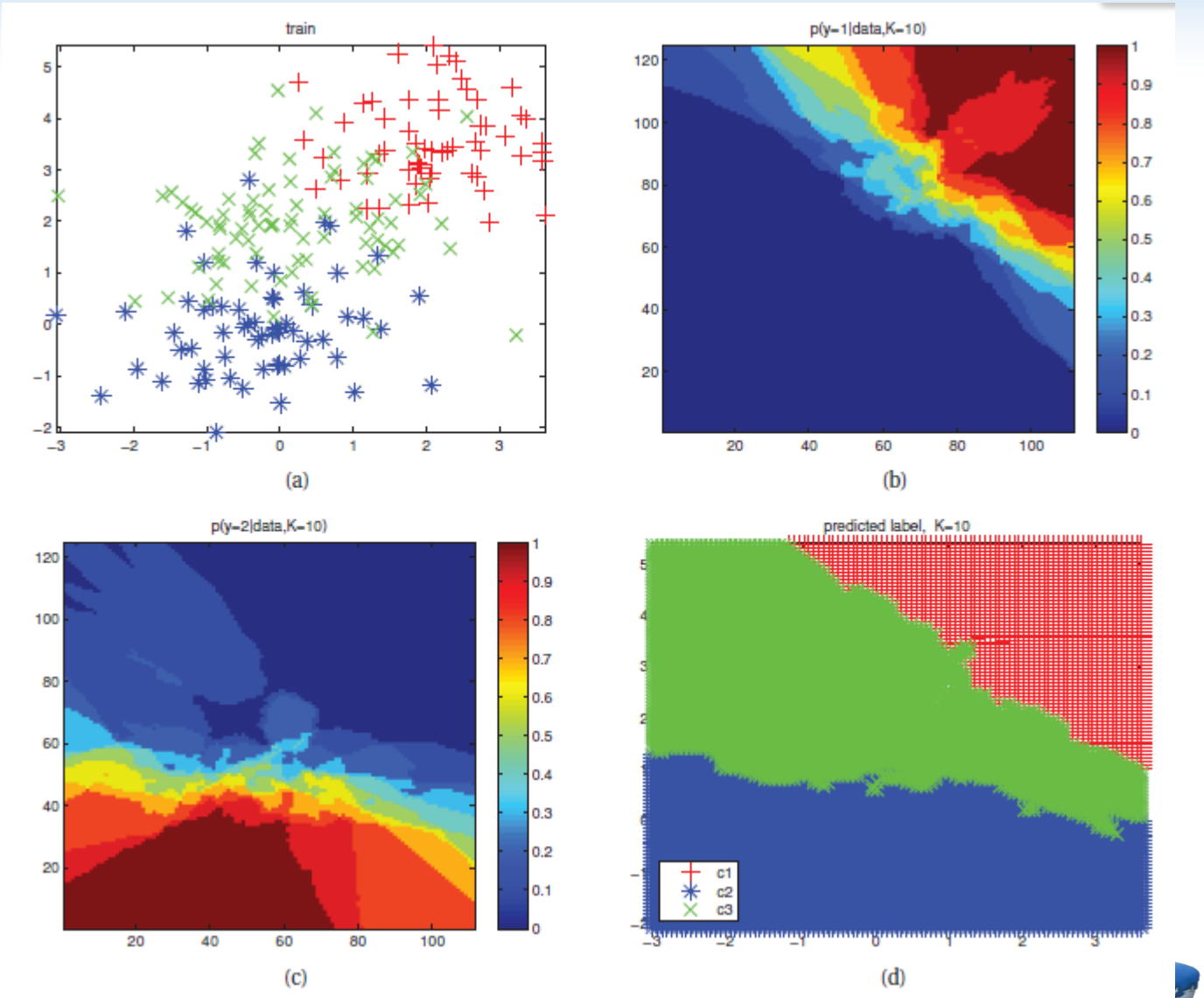

(b)

- (a) K=3
- (b) K=1

KNN模型举例

- 图a，描述了一个训练集
- 图b，描述了p(y=1|data, K=10)
- 图c，描述了p(y=2|data, K=10)
- 图d，预测了数据属于哪一类(K=10)
  - 属于哪类的概率最大，则属于哪类





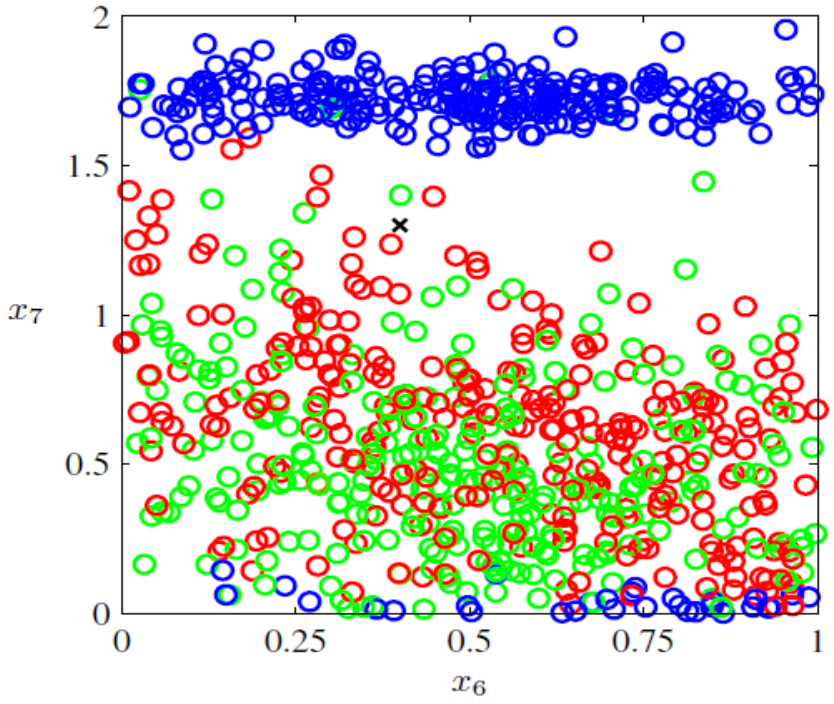
维度灾难

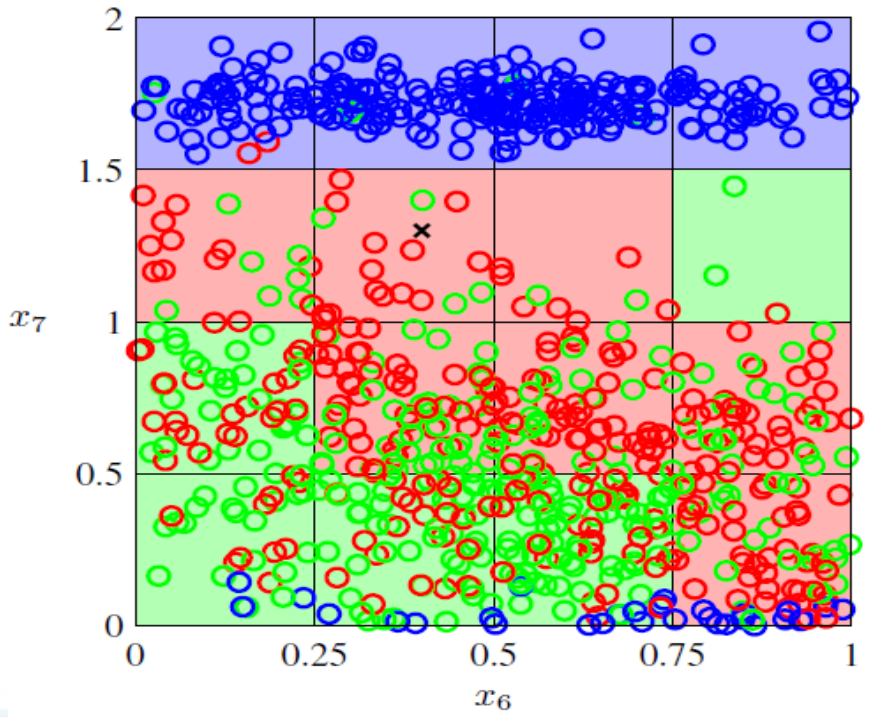
The curse of dimensionality

……

分类问题分析

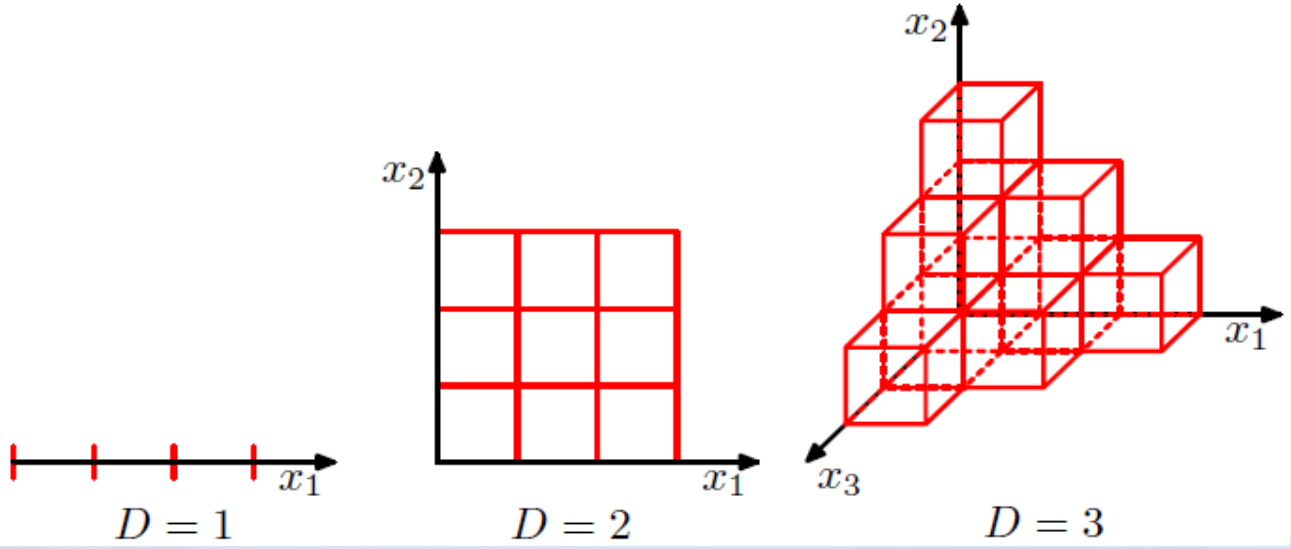
- 分类目标：用分类算法(比如，KNN)将测试数据‘×’*分到合适的类别*
- 凭直觉：
  - 确定“×”的类别，训练集中，越靠近“×”的点，应该越要起到更大的作用





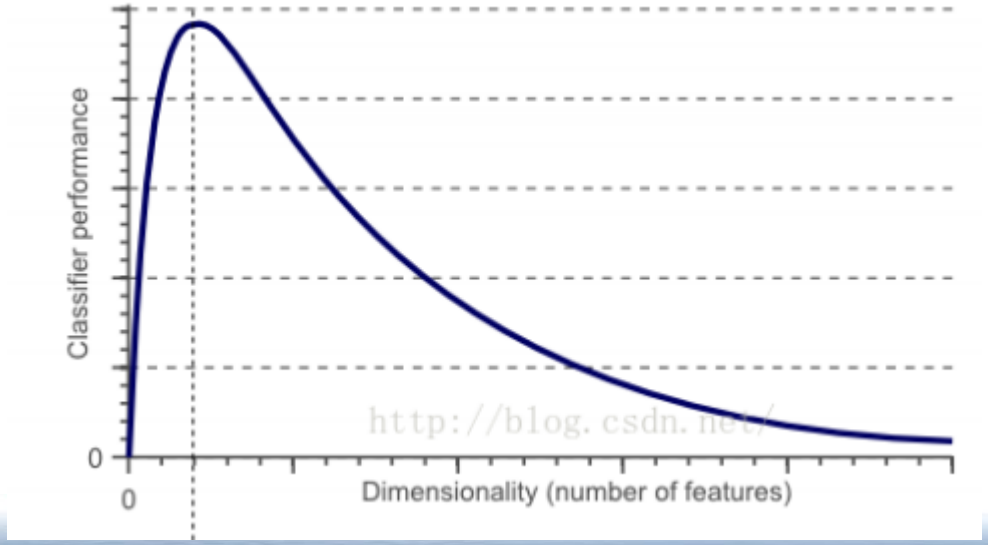
分类中一个严重的问题

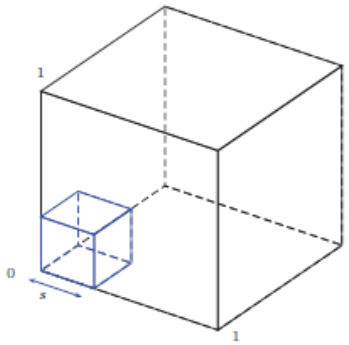
- 考虑不断提高维度的输入空间，将空间分成若干个区域。
  - 这种区域的数量呈指数级增长
  - 这就出现一个严重的问题
    - ❿ 为了正确分类任意测试数据x
    - ❿ 需要指数级增加大量的训练数据，以确保每个区域不是空的。



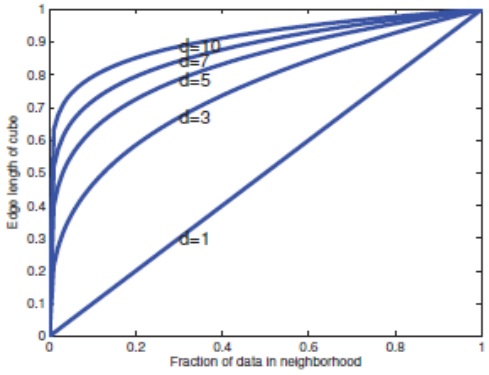
维度灾难出现了

- 当一个训练集在低维空间时，效果较好
  - 因为数据能够分布在整个空间，较好地表征整个数据空间
- 当数据空间维度升高时，还是这个训练集，效果就变差了
  - 空间被均匀划分的区域变多了，样本数据就只处于部分区域中，空间变得稀疏了
  - 要预测的x周围，小范围内只有很少数据，或者没有数据，因此，效果变差
  - 要预测的x周围，如果使用大范围，保证足够数据，意味着用关系很小的数据来预测，效果也差





(a)



(b)

防止维度灾难的方法

- 一种主要的方法
  - 对数据分布的性质一些假设（ ***p(y|x) 或 p(x)*** ）
    - ❿这些假设也称为归纳偏好
  - 这些假设通常以参数化模型的形式体现

参数化模型

Parametric classifier

……

参数化模型

- 一类参数化模型
  - 固定参数个数的统计模型
- 广泛使用的两种参数化模型
  - 线性回归
  - 逻辑回归

线性回归

Linear regression

……

线性回归的思路

- 假设输入输出之间存在线性关系（常假设就ε服从高斯分布）：

$$y(x) = W^T X + \varepsilon = \sum_{j=1}^{D} w_j x_j + \varepsilon$$

- 线性回归的目标是找到最佳的参数 W，使得模型能够最好地拟合训练数据。

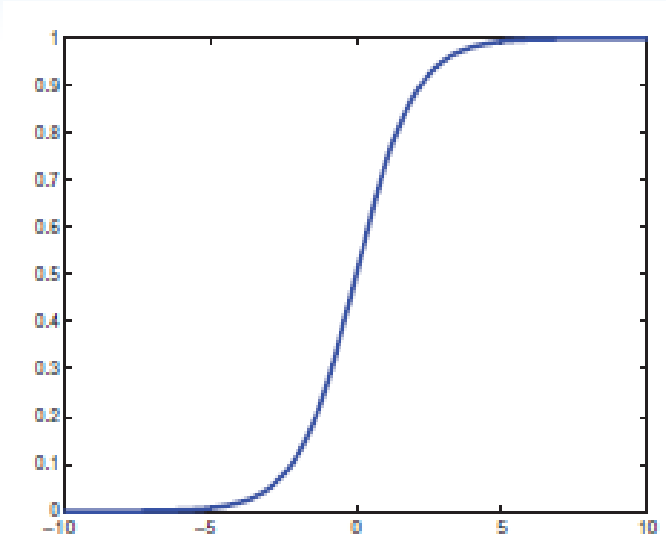
逻辑回归

Logistic regression

……

sigmoid 函数

$$sigm(x) = \frac{1}{1+\exp(-x)} = \frac{e^x}{1+e^x}$$



- Sigm() : **sigmoid** 函数 或者 逻辑函数.
- x>0.5 → y=1;
- x<-0.5 → y=0
- -0.5<x<0.5 y=0.5

逻辑回归的思路

- 希望用线性回归方法来解决分类问题
- 假设：y=sigm($w^T x$)，y服从伯努利分布

$$p(y \mid x, w) = Ber(y \mid \mathrm{sigm}(\mathrm{w}^{\mathrm{T}}\mathrm{x}))$$

其中，　$p(y = 1|x) = \mathrm{sigm}(w^T x)$

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

模型选择

Model selection

模型选择

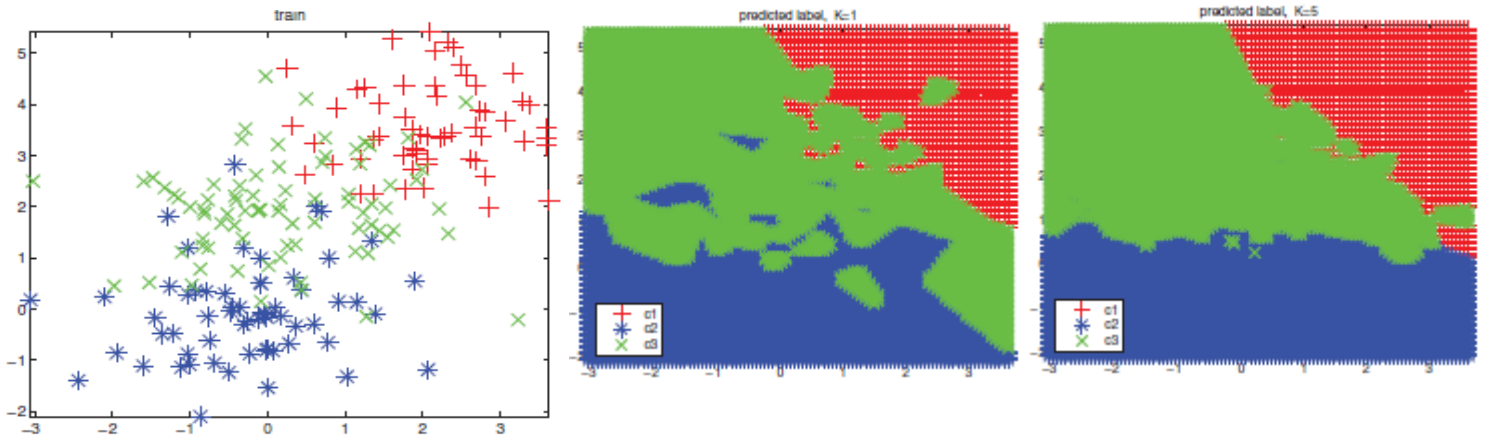
- 如果我们有各种不同复杂性的模型，例如：
  - 不同次数多项式的线性或逻辑回归模型
  - 或具有不同K值的KNN分类器
- 我们应该如何选择正确的模型？

容易想到的一种模型选择方法

- 计算每种模型在训练集上的错分率

$$err(f, D) = \frac{1}{N}\sum_{i=1}^{N} I(f(x_i) \neq y_i)$$

- 例如:



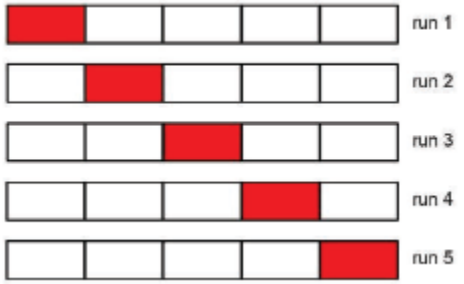
- 可以看到，对应KNN模型，增加K会增加在训练集中的误差
- 设K=1，可以在训练集上得到最小的误差
- 但我们真正关心的是：泛化误差

泛化误差

- 泛化误差：模型对未来数据的错分率
- 由于我们没有未来数据，可用下面方法近似
  - 计算模型在独立的大测试集上的错分率
- 一种选择K的方法：
  - 使模型在测试集上误差最小。
- 可是，在训练模型时，无法访问测试集

交叉检验

- 将训练集划分为两部分：
  - 一部分于训练，另一部分用于测试，称为验证集
- 交叉验证（CV）：
  - 轮流在训练集上训练，验证集上进行测试
  - 最后把几个结果进行平均，得到最终结果，评价模型




AUSTRALIA
OCEANIA
Northern Territory
Queensland
Western Australia
South Australia
NEW ZEALAND
VANUATU
TONGA
Coral Sea
Tasman Sea

Thank You !