机器学习

李成龙

安徽大学人工智能学院
“多模态认知计算”安徽省重点实验室
合肥综合性国家科学中心人工智能研究院

内容安排

- 什么是机器学习
- 机器如何学习
- 如何让机器学习的更好
- 为什么机器能学习

内容安排

- 如何让机器学习的更好
  - 数据问题
  - 评测问题
  - 模型问题
  - 算法问题

本节目录

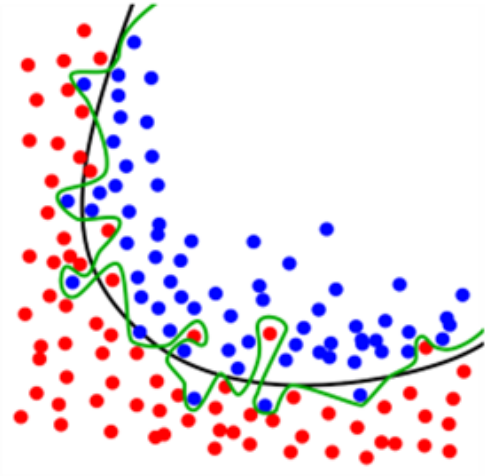
- 数据问题
- 评测问题
- 模型问题
- 算法问题

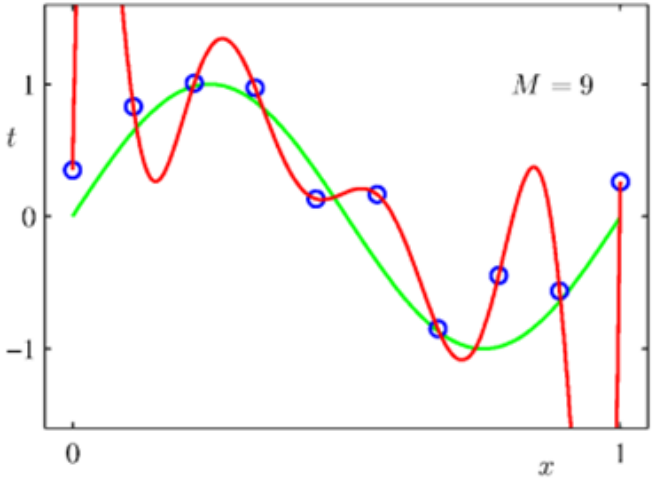
本节目录

- 数据问题
- 评测问题
- 模型问题
- 算法问题

数据问题

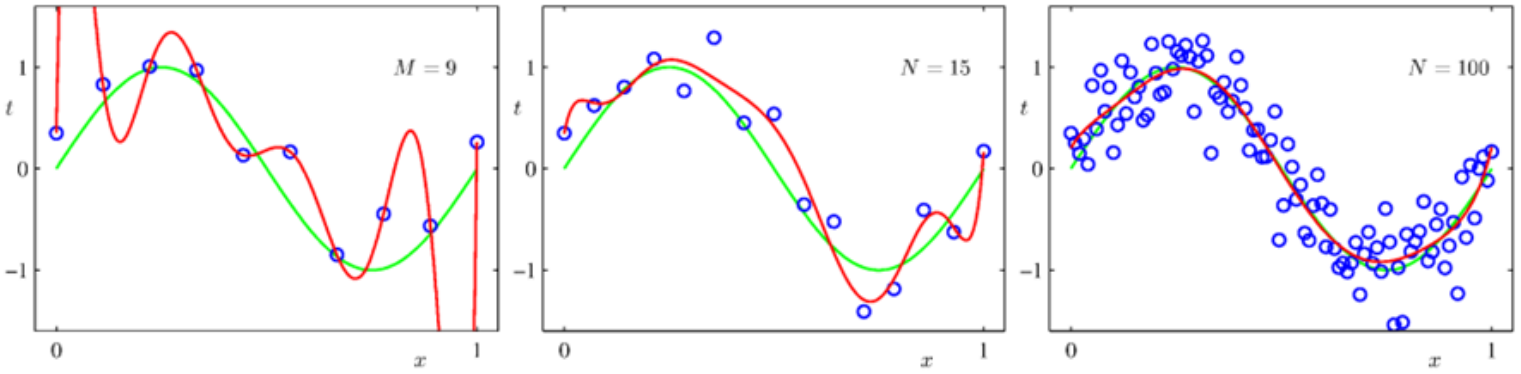
- **数据增广**
  - 完全拟合训练数据不是我们真正关心的
    - 模型有可能会拟合数据中的噪声
    - 这样的模型在新数据预测方面往往表现很差





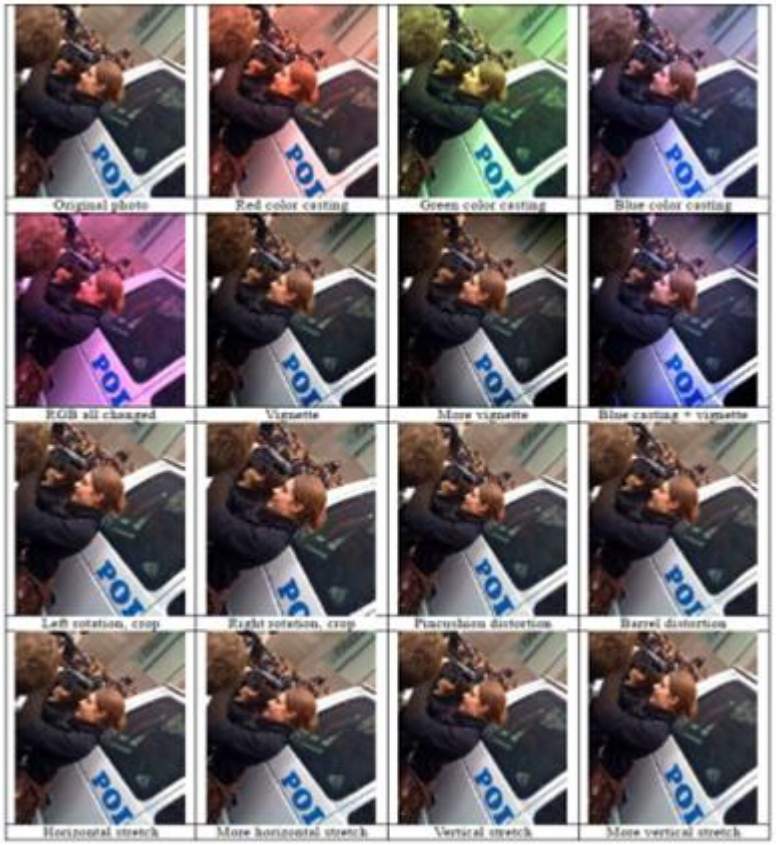
数据问题

- **数据增广**
  - 使用更多的数据可以降低噪声影响
    - 获取新数据往往难度较大
    - 标注代价一般也非常高



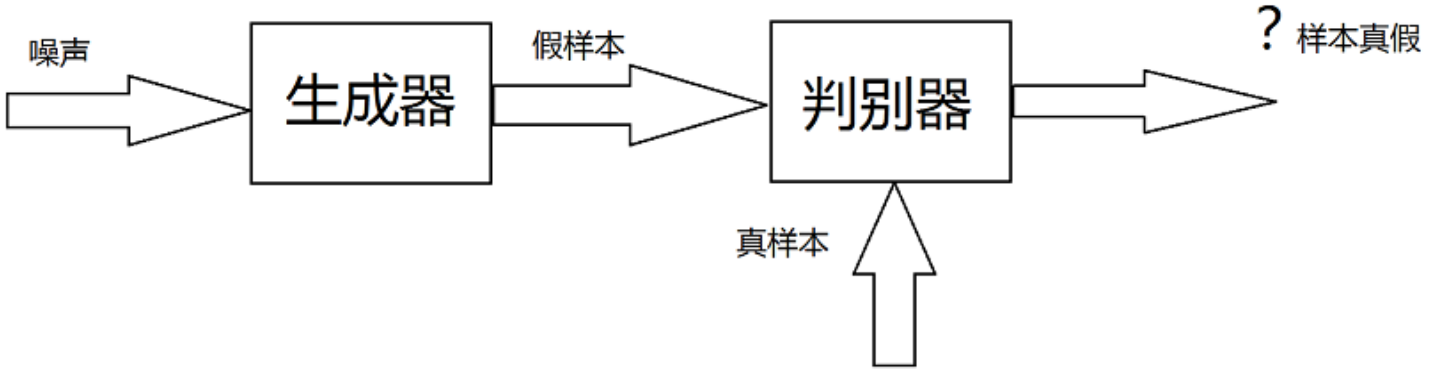
数据问题

- **数据增广**
  - 使用更多的数据可以降低噪声影响



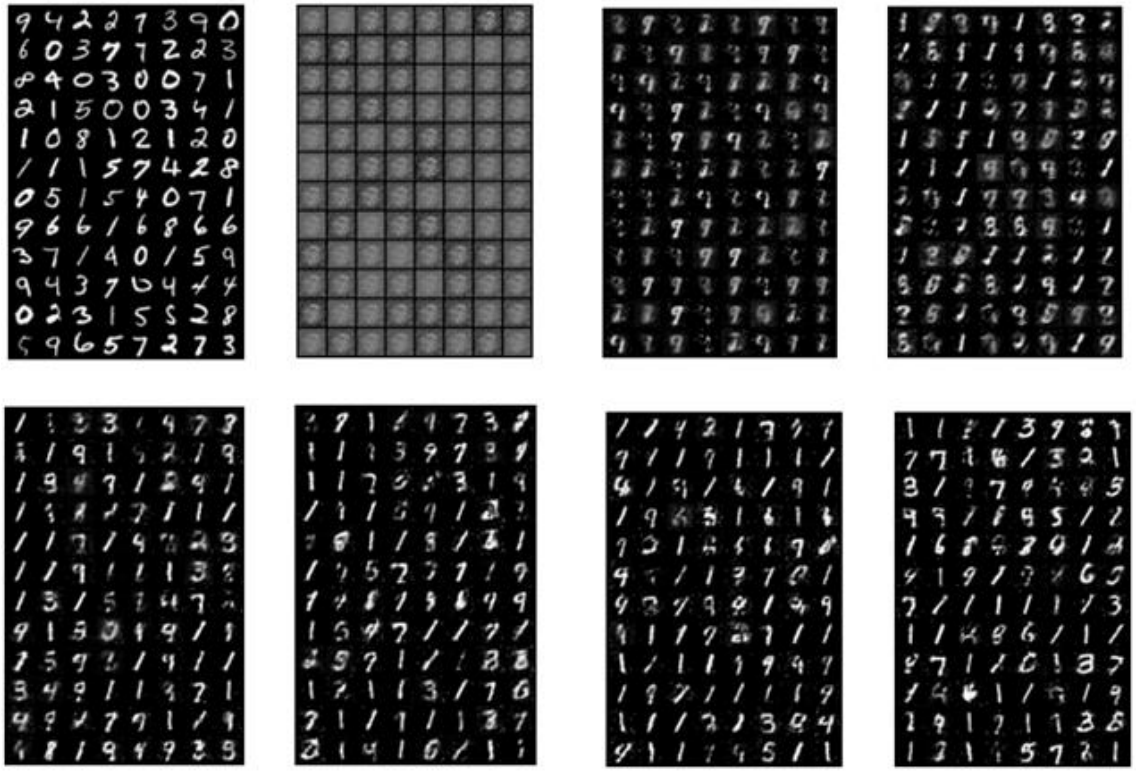
数据问题

- **数据生成**
  - 生成对抗网络基本思想：生成器生成新的样本，将生成的虚拟样本和训练集中实际图像样本随机输入判别器中，通过GAN模型的判别器判别输入样本是否为虚拟样本



数据问题

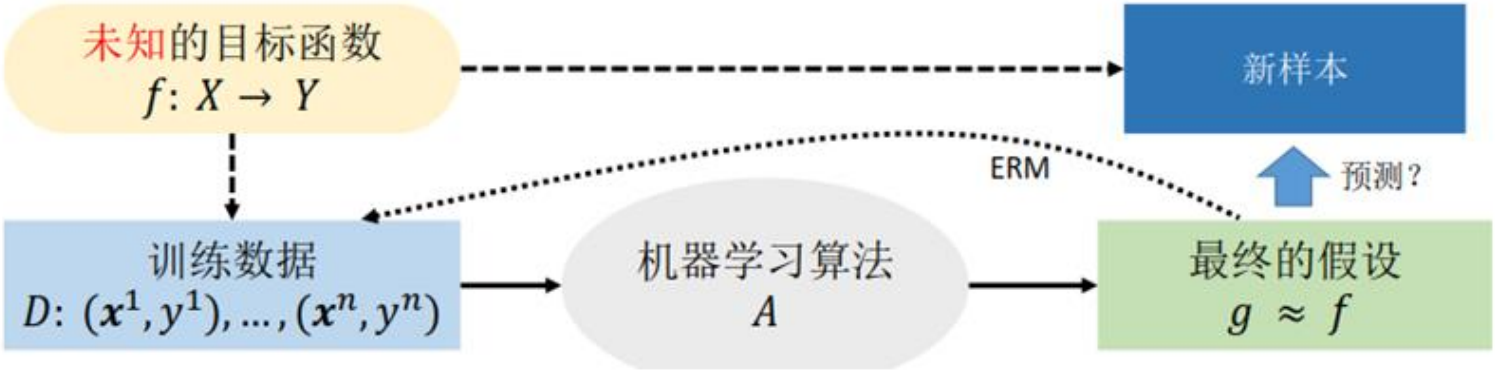
- **数据生成**
  - 生成对抗网络基本思想：生成器生成新的样本，将生成的虚拟样本和训练集中实际图像样本随机输入判别器中，通过GAN模型的判别器判别输入样本是否为虚拟样本



本节目录

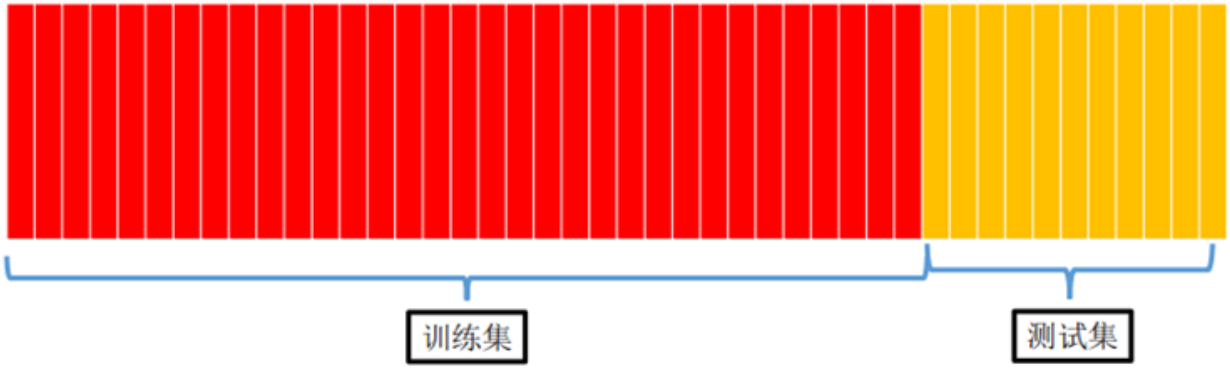
评测问题

- **学习目标**
  - 泛化能力：针对从真实函数（分布）中产生的新数据，做出良好预测
  - 如果模型在拟合当前数据方面表现较好，那么如何相信该模型对其他新样本做出良好预测呢



评测问题

- **拆分数据集**
  - 确保测试集满足以下两个条件
    - 规模足够大，可产生统计意义的结果
    - 能代表整个数据集，即挑选的测试集的特征应该与训练集的特征相似
    - 典型陷阱：测试误差低得令人惊讶，可能意味着对测试数据进行了训练

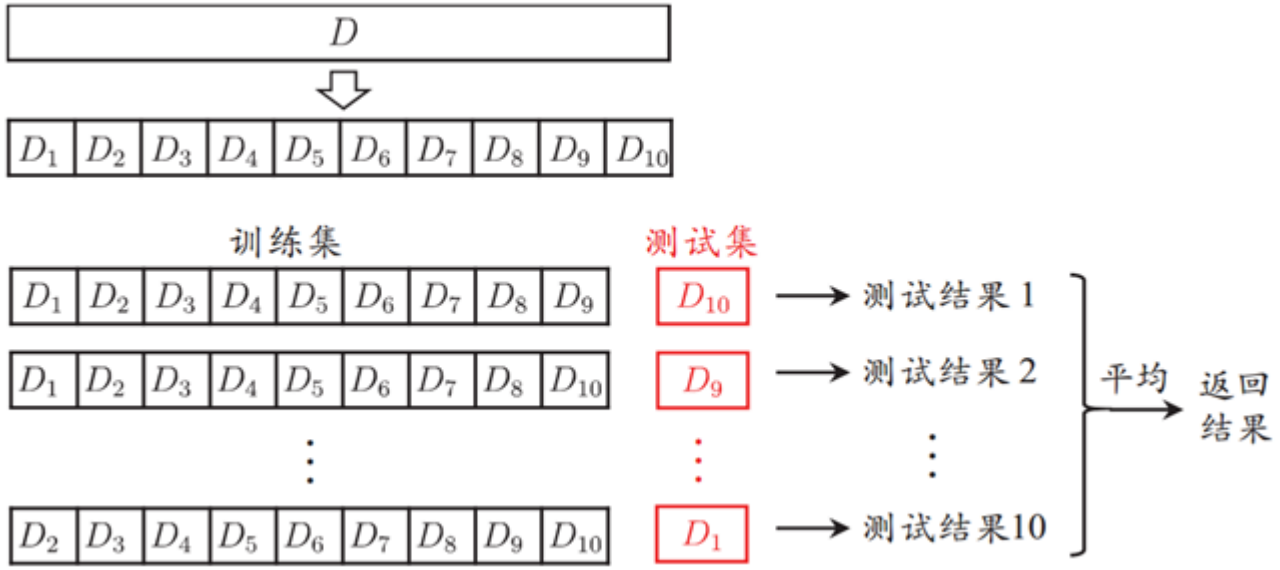


评测问题

- **拆分数据集**
  - 留出法
    - 直接将数据集划分为两个互斥集合
    - 训练/测试集划分要尽可能保持数据分布的一致性
    - 一般若干次随机划分、重复实验取平均值
    - 训练/测试样本比例通常为2:1~4:1

评测问题

- **拆分数据集**
  - 交叉验证法
    - 将数据集分层采样划分为k个大小相似的互斥子集
    - 每次用k-1个子集的并集作为训练集，余下的子集作为测试集，最终返回k个测试结果的均值
    - k最常用的取值是10



评测问题

- **拆分数据集**
  - 交叉验证法
    - 与留出法类似，将数据集D划分为k个子集同样存在多种划分方式，为了减小因样本划分不同而引入的差别，k折交叉验证通常随机使用不同的划分重复p次，最终的评估结果是这p次k折交叉验证结果的均值，例如常见的“10次10折交叉验证”
    - 假设数据集D包含m个样本，若令k=m，则得到留一法
      - 不受随机样本划分方式的影响
      - 结果往往比较准确
      - 当数据集比较大时，计算开销难以忍受

评测问题

- **拆分数据集**
  - 自助法
    - 以自助采样法为基础，对数据集D有放回采样m次得到训练集，其余作为测试集
    - 实际模型与预期模型都使用m个训练样本
    - 约有1/3的样本没在训练集中出现
    - 从初始数据集中产生多个不同的训练集，对集成学习有很大的好处
    - 自助法在数据集较小、难以有效划分训练/测试集时很有用；由于改变了数据集分布可能引入估计偏差，在数据量足够时，留出法和交叉验证法更常用

本节目录

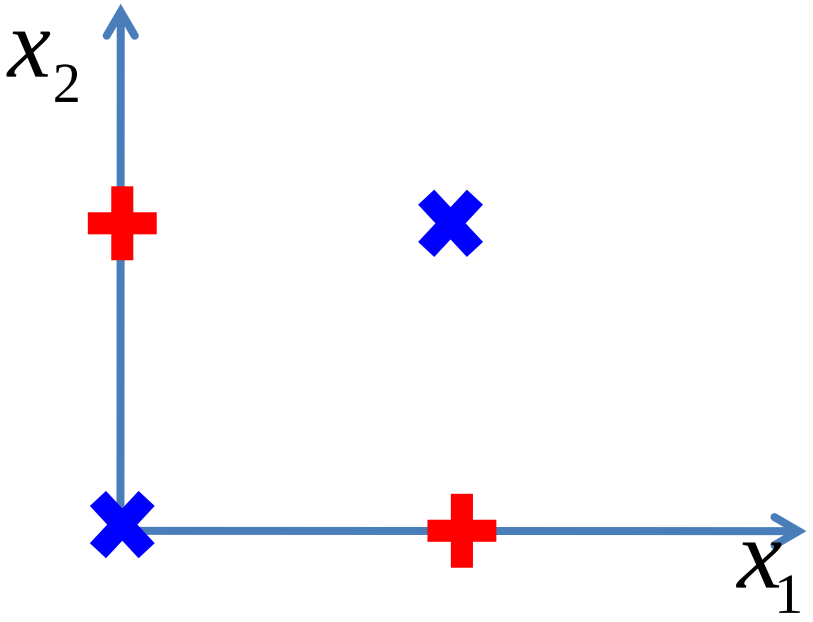
- 数据问题
- 评测问题
- **模型问题**
- 算法问题

模型问题

- **第一次AI低谷**
  - 1969年，美国著名人工智能学者马文·明斯基和西蒙·派柏特共同出版了《感知器：计算几何简介》一书，论证了感知器模型的两个关键问题
    - 单层的神经网络无法解决线性不可分问题，如异或门
    - 当时计算机的能力不足，无法满足计算量的需求
  - 由于这些问题当时无法得到解决，感知器的发展几乎停滞，以神经网络为基础的人工智能研究开始陷入低潮，相关项目长期无法得到政府经费支持，这段时间称为AI的“第一次低谷”

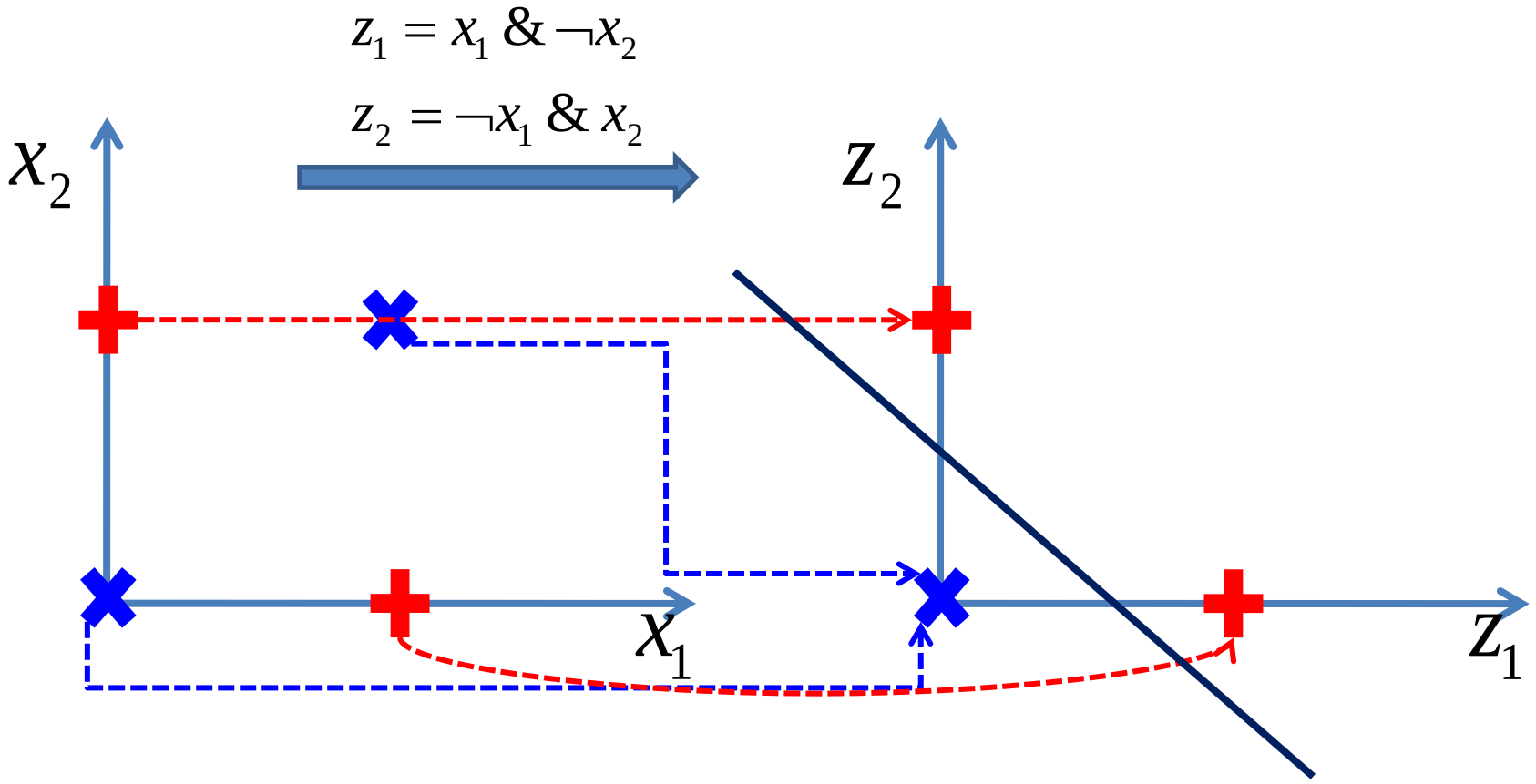
模型问题

- 例：异或门



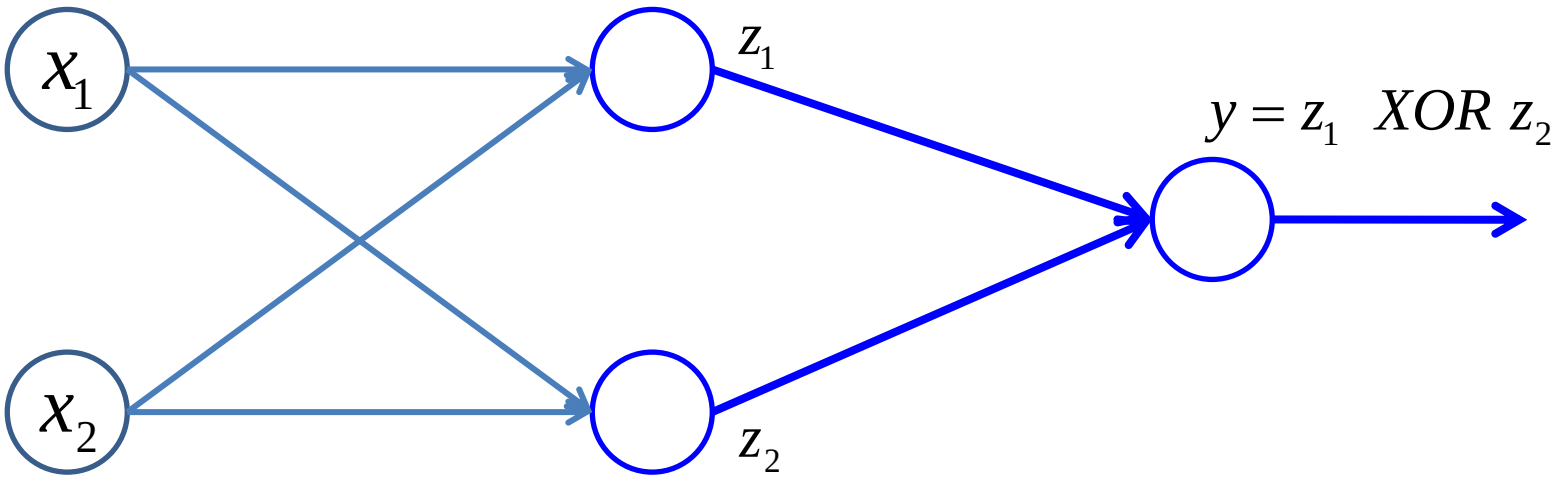
模型问题

- 例：异或门



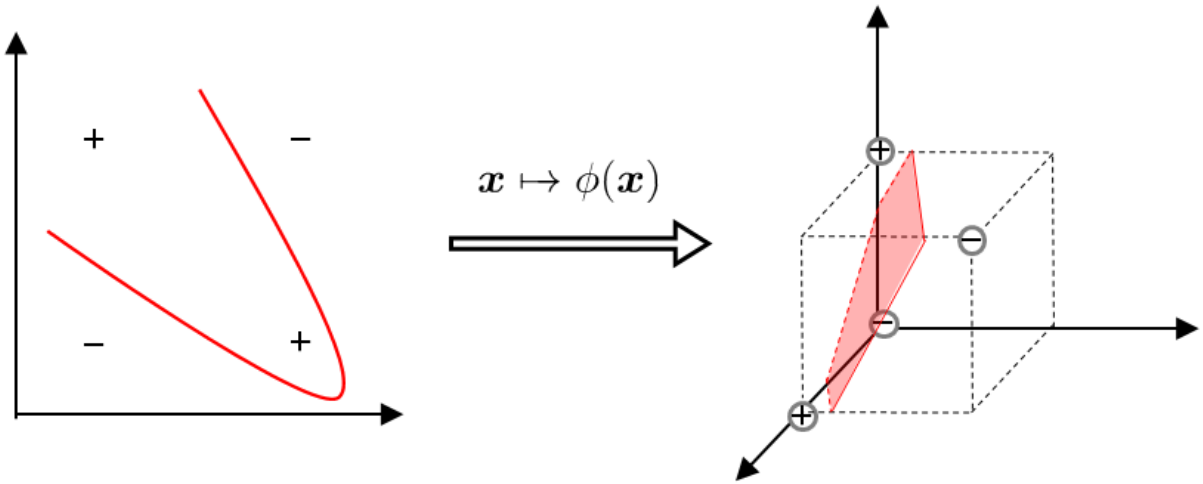
模型问题

- 例：异或门



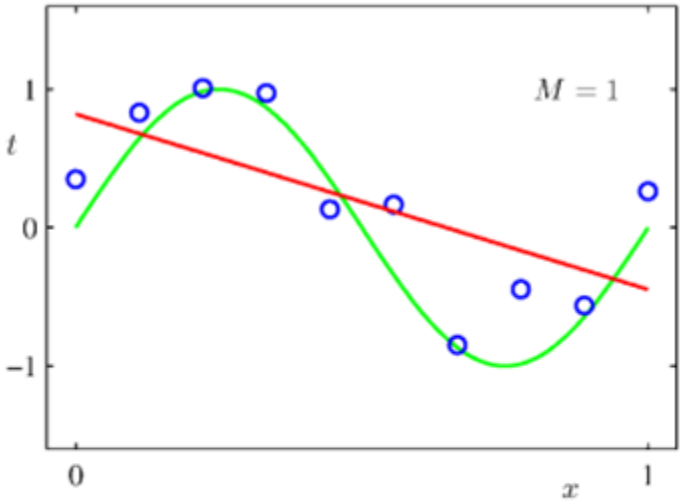
模型问题

- **多项式线性回归**
  - 线性不可分问题
    - 任何一个超平面都不能将不同类别的样本分开
  - 可以创建一个新的特征
    - 可以通过将x_1和x_2组合创新新的特征x_3
    - $x_3 = x_1 * x_2$
  - 使用扩展的线性模型时辅以特征组合一直都是训练大规模训练集的有效方法

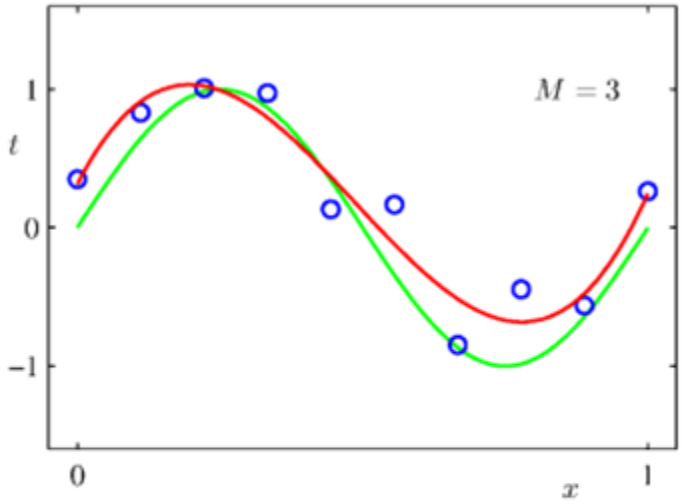


模型问题

- 多项式线性回归



$$g(\boldsymbol{x}) = w_0 + w_1 x$$



$$g(\boldsymbol{x}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

$$= w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$$

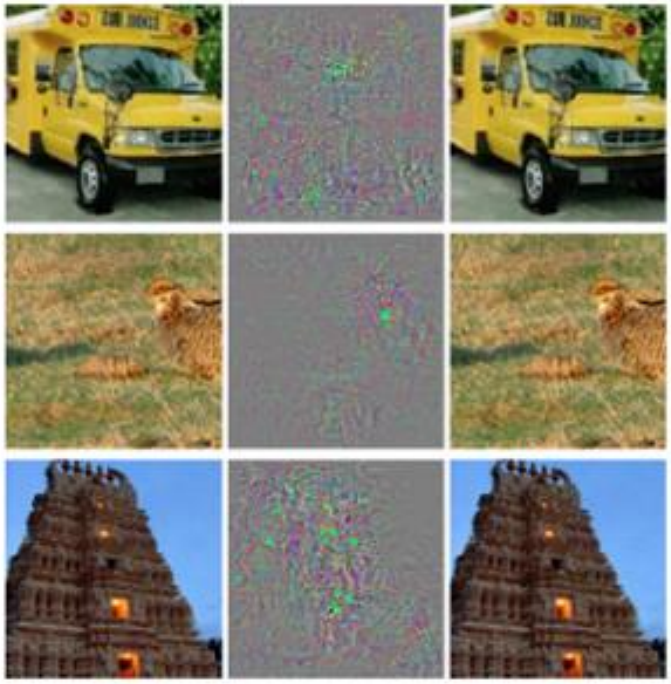
扩展的线性模型

模型问题

- **支持向量机与核函数**
  - 将原始特征向量映射到高维是解决非线性可分性的重要手段
  - 使用核函数
    - 引入核函数不需显式定义映射函数就能计算高维空间向量的内积，且计算复杂度低
    - 核函数不仅仅用于SVM，只要算法中出现了内积运算，都可以考虑核函数，从而提高处理高维数据的性能
  - 神经网络被SVM相关算法吊打了大概15年左右，直到深度学习的兴起

模型问题

- **对抗训练**
  - 有趣的现象：针对性的为样本添加微小噪声，模型对样本的判别结果很容易被篡改
    - 以下6幅原始图像添加给定噪声后均被判别为“鸵鸟”



模型问题

- **对抗训练**
  - 出现原因猜想：模型在高维空间中作为一种线性模型而引起的误差累积放大
    - 线性模型$f(X) = X\boldsymbol{\beta}$，考虑对抗样本$X' = X + \varepsilon$和真实样本$X$在该模型下输出的差异：
      $$f(X') - f(X) = \varepsilon_1\beta_1 + \varepsilon_2\beta_2 + \cdots + \varepsilon_k\beta_k$$
    - 微小的样本差异会导致模型输出差异很大
  - 对抗训练：将对抗样本作为训练样本用于模型训练过程，从而提高模型鲁棒性
  - 获取对抗样本方法
    - 简单界约束限制域拟牛顿法
    - 快速梯度符号法

模型问题

- **对抗训练**
  - 简单界约束限制域拟牛顿法
    - 基本思想：直接对优化目标添加对噪声ε的约束，使得添加了噪声ε的新样本被错误分类
    - 对于一个分类器f,定义优化目标：

$$\min||\varepsilon||_2;\ \text{s.t.}\begin{cases} f(X+\varepsilon)=l \\ \text{label}(X+\varepsilon)=\text{label}(X)\neq l \end{cases}$$

    - 寻找最小的噪声使得样本出现错分

$$f(X')-f(X)=\varepsilon_1\beta_1+\varepsilon_2\beta_2+\cdots+\varepsilon_k\beta_k$$

    - 目标函数形式可转换为

$$\min[\,c|\varepsilon|+F(X+\varepsilon,l)];\ \text{s.t. label}(X+\varepsilon)=\text{label}(X)\neq l$$

其中$c>0$，F为目标函数

模型问题

- **对抗训练**
  - 快速梯度符号法
    - 基本思想：对样本添加的噪声会通过高维线性模型累加，导致模型输出错误
    - 模型参数向量：$\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)^T$
    - 样本输入：X
    - 噪声：$\varepsilon = \gamma \text{sign}(\nabla_X F(\boldsymbol{\beta}, X, y))$
    - 对抗样本：$X' = X + \gamma \text{sign}(\nabla_X F(\boldsymbol{\beta}, X, y))$
    - $\text{sign}(\nabla_X F(\boldsymbol{\beta}, X, y))$为模型目标函数对$X$的梯度方向，在该方向上添加噪声可使的目标函数值变化最大

本节目录

算法问题

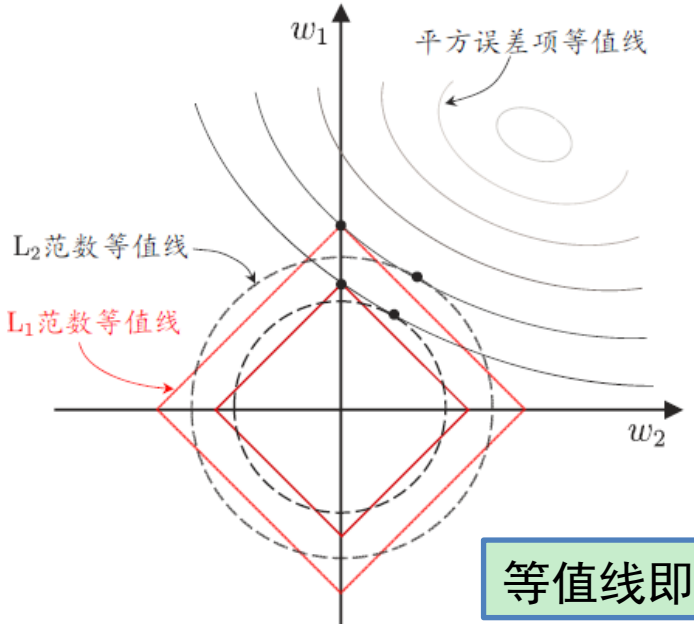
- **正则化**
  - 早停法
  - 奥卡姆剃刀定律：简单有效原理
  - 惩罚模型复杂度（降低模型容量）：以最小化损失和复杂度为优化目标
  - 结构风险最小化

$$\arg\min_{w,b} \sum_{i=1}^{n} l\left(g\left(\boldsymbol{x}^i;\boldsymbol{w}\right),y^i\right)+\lambda L(\boldsymbol{w})$$

  - 简单化： L_2范数-岭回归 $L(\boldsymbol{w}) = \|\boldsymbol{w}\|_2^2$
  - 稀疏化： L_1范数 $L(\boldsymbol{w}) = \|\boldsymbol{w}\|_1$

算法问题

- **正则化**
  - 例： L_1和L_2范数正则化
    - 假设$\boldsymbol{x}$仅有两个属性，那么$\boldsymbol{w}$有两个分量w_1和w_2. 那么目标优化的解要在平方误差项与正则化项之间折中, 即出现在图中平方误差项等值线与正则化等值线相交处
    - 从图中看出, 采用L_1范数时交点常出现在坐标轴上, 即产生w_1或者w_2为0的稀疏解



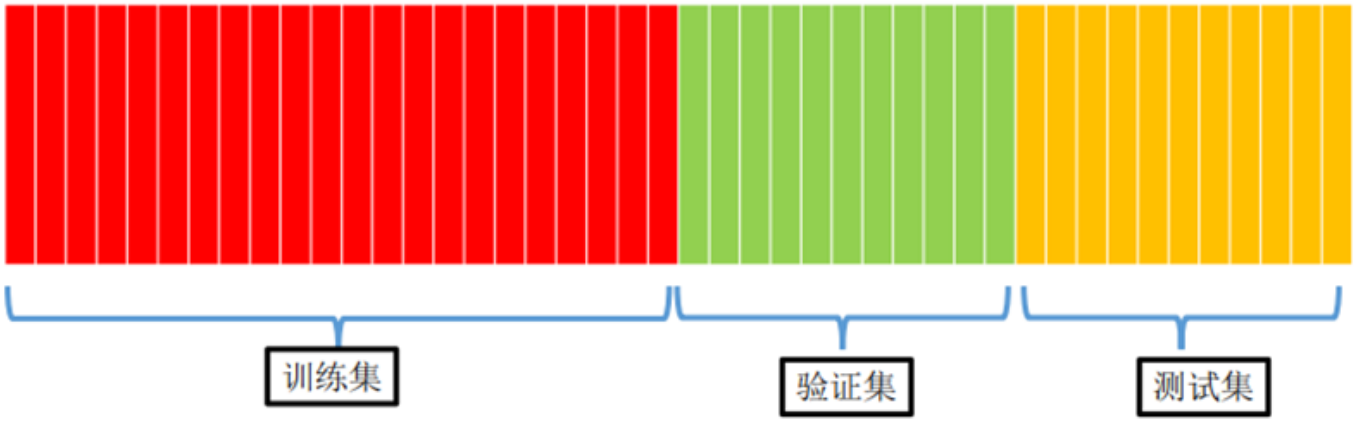
等值线即取值相同的点的连线

算法问题

- **超参数优化**
  - 超参数
    - 正则化中的 λ
    - 梯度下降法中的η
  - 超参数优化方法
    - 网格搜索法
    - 贝叶斯优化
    - 随机搜索法
    - 梯度优化法
    - 交叉验证法

算法问题

- **超参数优化**
  - 验证集划分



训练集：学习模型参数
验证集：选择模型/优化超参数
测试集：测试模型的泛化能力

总结

- **数据问题**
  - 数据增广
  - 数据生成
- **评测问题**
  - 训练集
  - 测试集
- **模型问题**
  - 特征组合
  - 核函数
  - 对抗训练
- **算法问题**
  - 正则化
  - 超参数优化

思考题

- 试分析L_1和L_2范数正则化的优缺点
- 数据增广和生成的样本数量越多越好吗

练习题

- 试通过验证集划分的方式优化模型的超参数，如梯度下降法的学习率