華東師範大學
EAST CHINA NORMAL UNIVERSITY

第1.2节 机器学习基本概念

Basic concepts of machine learning

第1.2节 机器学习基本概念

一、 机器学习准备

二、 基本概念

三、 欠拟合和过拟合

四、 机器学习的分类

一、 机器学习准备

1. 数据预处理

- 数据清洗
- 数据集成
- 数据采样

数据清洗

对各种脏数据进行对应方式的处理，得到标准、干净、连续的数据，提供给数据统计、数据挖掘等使用。

- 数据的完整性
  - 例如人的属性中缺少性别、籍贯、年龄等；
  - 解决方法：信息补全；剔除；
- 数据的合法性
  - 例如获取数据与常识不符，年龄大于150岁；
  - 解决方法：设置字段内容；类型的合法规则
- 数据的一致性
  - 例如不同来源的不同指标，实际内涵一样，或是同一指标内涵不一致；
  - 解决方法：建立数据体系，包含但不限于指标体系、维度、单位等

- 数据的唯一性
  - 例如不同来源的数据出现重复的情况等；
  - 解决方法：按主键去重 / 按规则去重
- 数据的权威性
  - 例如出现多个来源的数据，且数值不一样；
  - 解决方法：为不同渠道设置权威级别。

数据集成

数据集成是要将互相关联的分布式异构数据源集成到一起，使用户能够以透明的方式访问这些数据源。

- 集成是指维护数据源整体上的数据一致性、提高信息共享利用的效率;
- 透明的方式是指用户无需关心如何实现对异构数据源数据的访问,只关心以何种方式访问何种数据。

数据集成方法：

- 模式集成方法
  - 在构建集成系统时将各数据源的数据视图集成为全局模式，使用户能够按照全局模式透明地访问各数据源的数据
  - 联邦数据库以及中间件集成方法

- 数据复制方法
  - 将各个数据源的数据复制到与其相关的其它数据源上，并维护数据源整体上的数据一致性、提高信息共享利用的效率.
  - 数据异构性问题
- 综合性集成方法: 将模式集成方法和数据复制方法结合在一起
  - 想办法提高基于中间件系统的性能，该方法仍有虚拟的数据模式视图供用户使用，同时能够对数据源间常用的数据进行复制
  - 对于用户简单的访问请求，综合方法总是尽力通过数据复制方式. 在本地数据源或单一数据源上实现用户的访间需求;
  - 对那些复杂的用户请求，无法通过数据复制方式实现时，才使用虚拟视图方法。

数据采样

- 数据不平衡(imbalance): 数据集的类别分布不均
  - 例如：一个二分类问题，100个训练样本
    - 理想情况：正类、负类样本的数量相差不多
    - 类不平衡：正类样本有99个、负类样本仅1个
  - 过采样(Over-Sampling)
    - 随机复制少数类增加其中的实例数量，从而可增加样本中少数类的代表性
  - 欠采样(Under-Sampling)
    - 随机消除占多数的类的样本平衡类分布，直到多数类和少数类的实例实现平衡
  - SMOTE(Synthetic Minority Over-sampling Technique)算法
    - 合成新的少数类样本，而不是简单地复制样本

一、 机器学习准备

- 数据集拆分
  - 机器学习中将数据划分为3份
    - 训练集：训练模型
    - 验证集：评估模型，根据效果调整模型。
    - 测试集：确认模型的效果

2. 特征工程

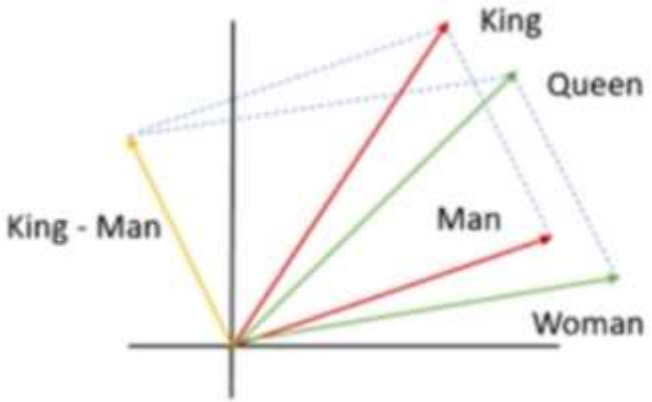
- 特征编码
- 特征选择
- 特征降维
- 规范化

特征编码

- 将数据转化为数值形式进行编码

	Direction	District	Elevator	Floor	Garden	id	Layout	Price	Region	Renovation	Size	Year
0	东西	灯市口	NaN	6	锡拉胡同21号院	101102647043	3室1厅	780.0	东城	精装	75.0	1988
1	南北	东单	无电梯	6	东华门大街	101102650978	2室1厅	705.0	东城	精装	60.0	1988
2	南西	崇文门	有电梯	16	新世界中心	101102672743	3室1厅	1400.0	东城	其他	210.0	1996
3	南	崇文门	NaN	7	兴隆都市馨园	101102577410	1室1厅	420.0	东城	精装	39.0	2004

- one-hot编码: 使用N位向量来对N个状态进行编码，每个状态都有它独立的位置，并且在任意时候，其中只有一位有效。
- 语义编码: 将数据根据每位特征的语义处理成向量。



特征选择

例1：两类问题：男性、女性

数据特征：

- 身高、体重、音频、头发长短
- 出生日期、家庭住址、籍贯、专业

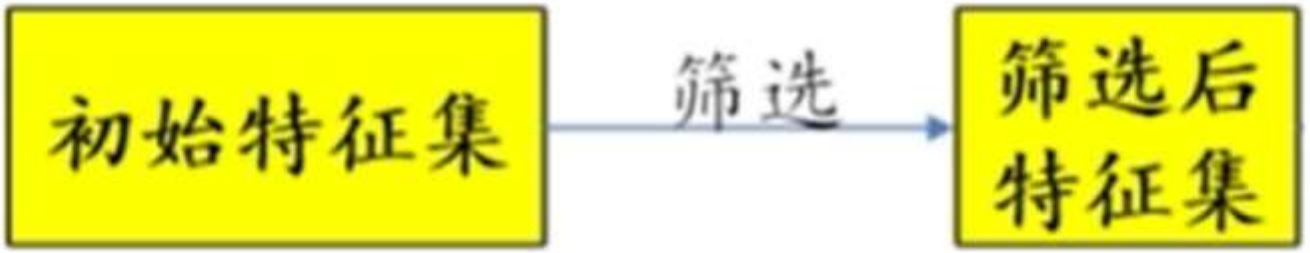
例2：高维数据的稀疏情况

- 数据1：$1, 0, 0, 0, 0, 0, 1, 3, 0, 0, 0, 0, 0, 0$
- 数据2：$1, 0, 0, 0, 0, 0, 3, 3, 0, 0, 0, 0, 0, 0$
- 数据3：$1, 0, 1, 0, 0, 0, 5, 3, 0, 2, 0, 0, 0, 0$

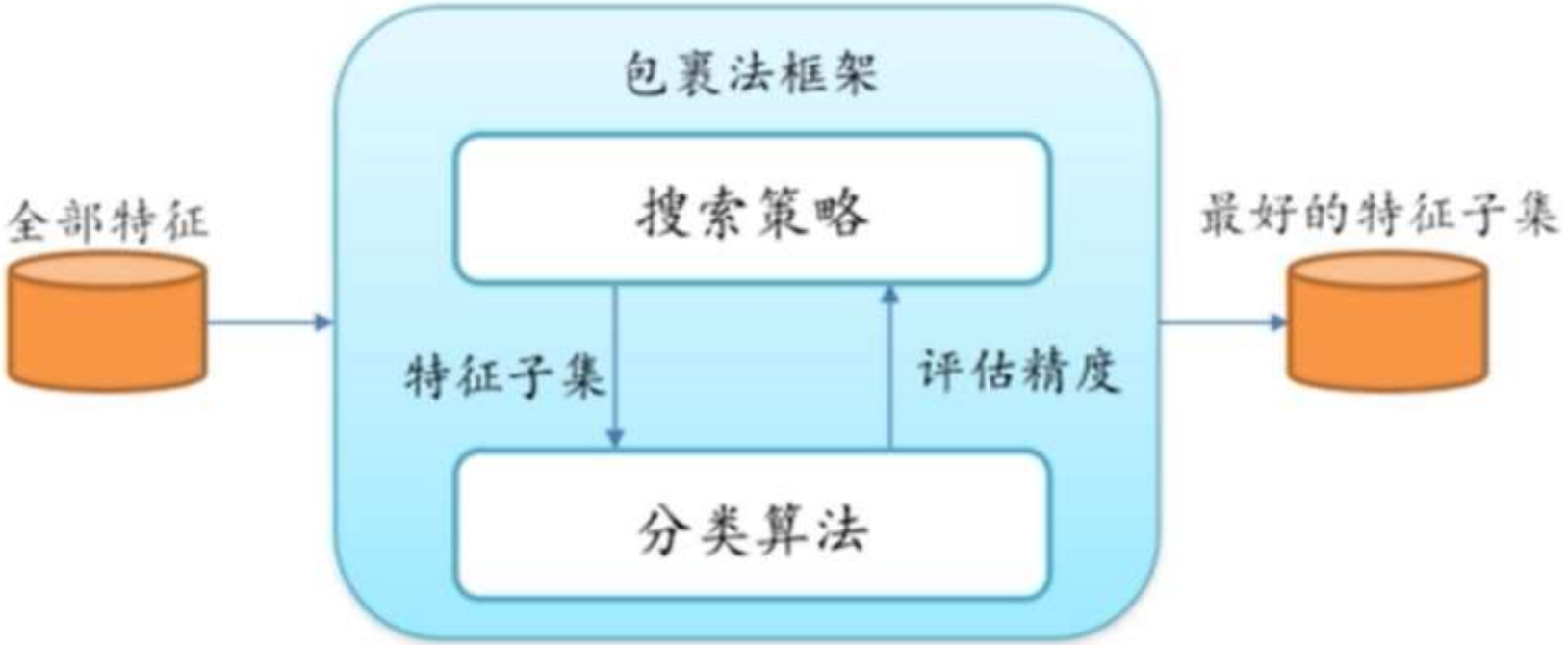
特征选择的主要方法

- 过滤法
  - 按照发散性或相关性对各特征进行评分，设定阈值完成特征选择
  - 互信息：两个随机变量之间的关联程度

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$



- 包裹法
  - 从初始特征集合中不断的选择特征子集，训练学习器，根据学习器的性能来对子集进行评价，直到选择出最佳的子集



- 嵌入式(Embeded)
  - 利用正则化的思想，将部分特征属性的权重调整到0，则这个特性相当于就是被舍弃了。

特征降维

- 主成分分析(PCA)
  - 将原始特征空间映射到彼此正交的特征向量空间，在非满秩的情况下使用奇异值分解(SVD)来构建特征向量。
- 线性判别分析(LDA)
  - 给出一个标注了类别的数据集，投影到了一条直线之后，能够使得点尽量的按类别区分开。

规范化

不同属性具有不同量级时会导致：

- 数量级的差异将导致量级较大的属性占据主导地位；
- 数量级的差异将导致迭代收敛速
- 依赖于样本距离的算法对于数据的数量级非常敏感

标准化：通过减去均值然后除以方差（或标准差），将数据按比例缩放，使之落入一个小的特定区间。

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

数据标准化为了不同特征之间具备可比性，经过标准化变换之后的特征数据分布没有发生改变。

就是当数据特征取值范围或单位差异较大时，最好是做一下标准化处理。

区间缩放(最大-最小规范化)：将属性缩放到一个指定的最大和最小值（通常是1-0）之间。

$$x = \frac{x - \min}{\max - \min}$$

数据归一化的目的是使得各特征对目标变量的影响一致，会将特征数据进行伸缩变化，所以数据归一化是会改变特征数据分布的。

需要做数据归一化/标准化

- 线性模型，如基于距离度量的模型包括KNN(K近邻)、K-means聚类、是需要做数据归一化/标准化处理的。

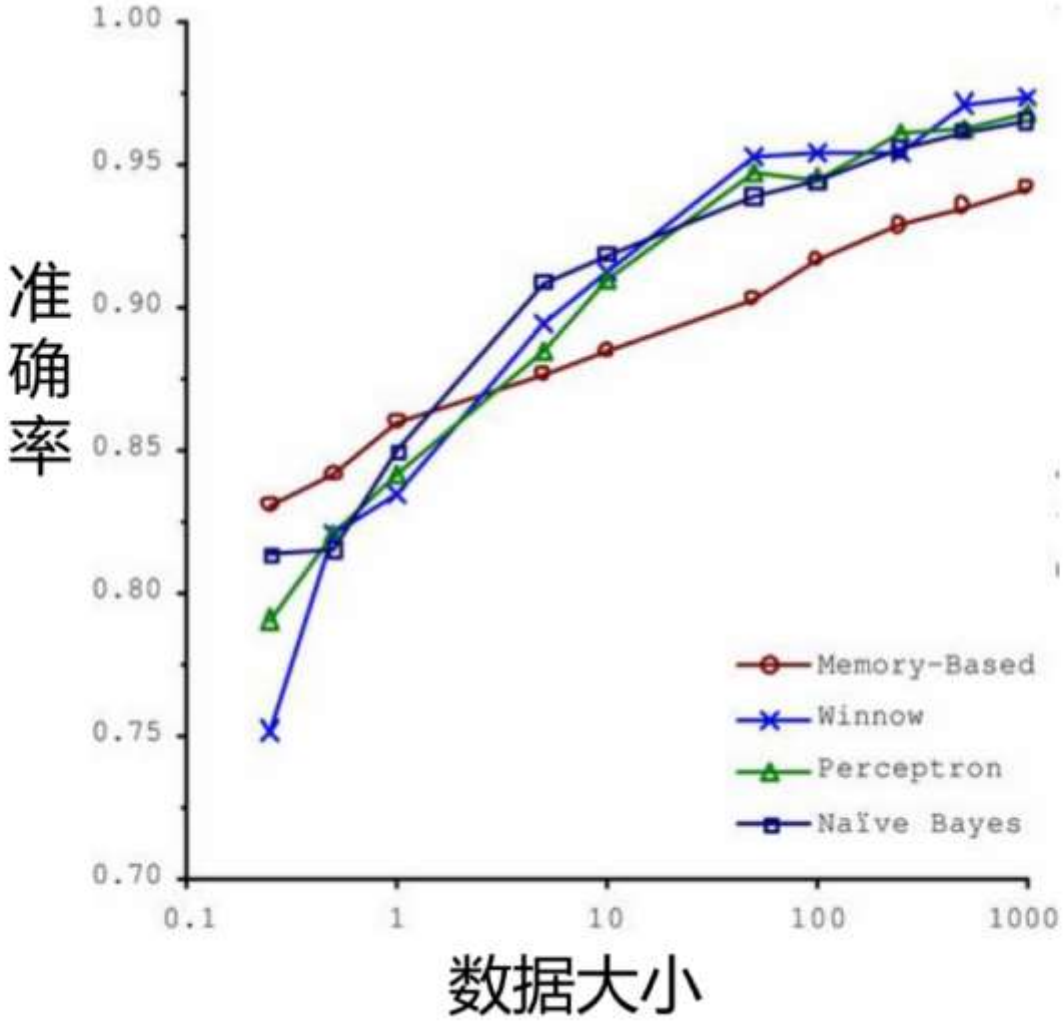
不需要做数据归一化/标准化

- 决策树、基于决策树的Boosting和Bagging等集成学习模型对于特征取
  值大小并不敏感，如随机森林、XGBoost、LightGBM等树模型，以及朴素贝叶斯，以上这些模型一般不需要做数据归一化/标准化处理。

优点：

- 提升模型精度
  - 不同维度之间的特征在数值上有一定比较性，可以大大提高分类器的准确性。
- 提升收敛速度
  - 最优解的寻优过程明显会变得平缓，更容易正确的收敛到最优解。
- 标准化更好保持了样本间距
- 标准化更符合统计学假设

成功的机器学习应用不是拥有最好的算法，而是拥有最多的数据！



以监督学习为例

1. 基本术语

- 输入: $\mathbf{X} \in \mathcal{X}$
- 输出: $Y \in \mathcal{Y}$
- 输入实例: $\mathbf{x} = (x_1, x_2, \ldots, x_d)$或者$\mathbf{x} = (x^1, x^2, \ldots, x^d)$
- 输出实例: $y \in \mathcal{Y}$
- 数据集: $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$
- 目标函数: $Y = f(\mathbf{X})$;目标分布:$P(Y|\mathbf{X})$
- 对具体的输入时: $y = f(\mathbf{x})$或$P(y|\mathbf{x})$

2. 机器学习三要素: 模型

模型：决策函数$Y = f(\mathbf{X})$或者条件概率分布$P(Y|\mathbf{X})$

假设空间：决策函数或者条件概率分布的集合

决策函数集合：$\mathcal{H} = \{f \mid Y = f(\mathbf{X}; \theta), \theta \in \mathbb{R}^n\}$

条件概率的集合：$\mathcal{H} = \{P \mid P(Y|\mathbf{X}; \theta), \theta \in \mathbb{R}^n\}$

3. 机器学习三要素: 策略

策略：从假设空间中选取最优模型**损失函数**:

0-1损失函数:

$$L(Y, f(\mathbf{X})) = \mathbb{I}(f(\mathbf{X}) \neq Y) = \begin{cases} 1, & f(\mathbf{X}) \neq Y \\ 0, & f(\mathbf{X}) = Y \end{cases}$$

平方损失函数:

$$L(Y, f(\mathbf{X})) = (Y - f(\mathbf{X}))^2$$

绝对损失函数:

$$L(Y, f(\mathbf{X})) = |Y - f(\mathbf{X})|$$

对数损失函数:

$$L(Y, P(Y|\mathbf{X})) = -\log P(Y|\mathbf{X})$$

例1. 在分类数为M的分类问题中，设$p_i(\mathbf{X})$为分类器将$\mathbf{X}$预测为类别i的概率，则其对数损失函数为?

$$L(Y, P(Y|\mathbf{x})) = -\log P(Y|\mathbf{x}) = -\sum_{i=1}^{M} \mathbb{I}(Y=i) \log p_i(\mathbf{X})$$

更进一步，考虑数据集容量为N，则该数据集的平均损失函数（代价函数）为:

$$L(Y, P(Y|\mathbf{X})) = -\log P(Y|\mathbf{X}) = -\frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{M} \mathbb{I}(Y=i) \log p_i(\mathbf{X}_j)$$

风险函数:

期望风险(泛化误差): $R_{exp}(f) = \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(\mathbf{x}))P(\mathbf{x}, y)d\mathbf{x}dy$

由于$P(\mathbf{x}, y)$是不可知的，所以监督学习是一个病态(ill-formed)问题。

给定训练集: $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$

经验风险: $R_{emp}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(\mathbf{x}_i))$

结构风险: $R_{srm}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(\mathbf{x}_i)) + \lambda J(f)$

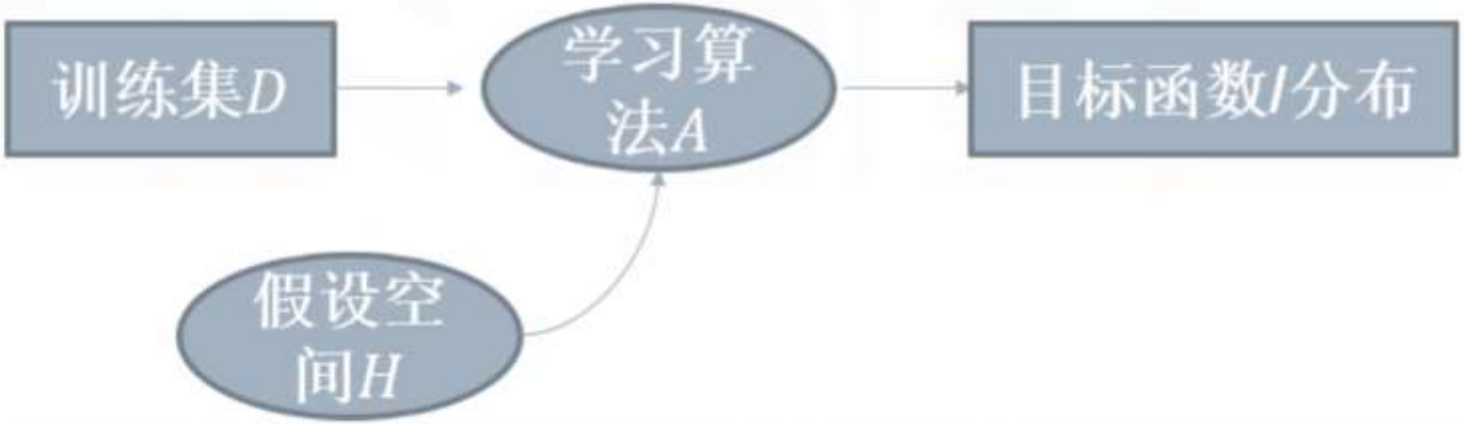
策略: $\min_{f\in\mathcal{H}} R_{emp}(f)$ 或者 $\min_{f\in\mathcal{H}} R_{srm}(f)$

4. 机器学习三要素: 算法

算法: 学习模型的具体算法, 选取最优模型

- 最优化问题: $\min\limits_{f\in\mathcal{H}} R_{emp}(f)$ 或者 $\min\limits_{f\in\mathcal{H}} R_{srm}(f)$
  - 极值问题
  - 梯度下降
  - 牛顿法和拟牛顿法
  - 约束优化问题——拉格朗日乘数法

5. 机器学习的一般流程



步骤：

(1) 得到一个有限的训练数据集合

(2) 确定包含所有可能的模型的假设空间，即学习模型的集合

(3) 确定模型选择的准则，即学习的策略

(4) 实现求解最优模型的算法，即学习的算法

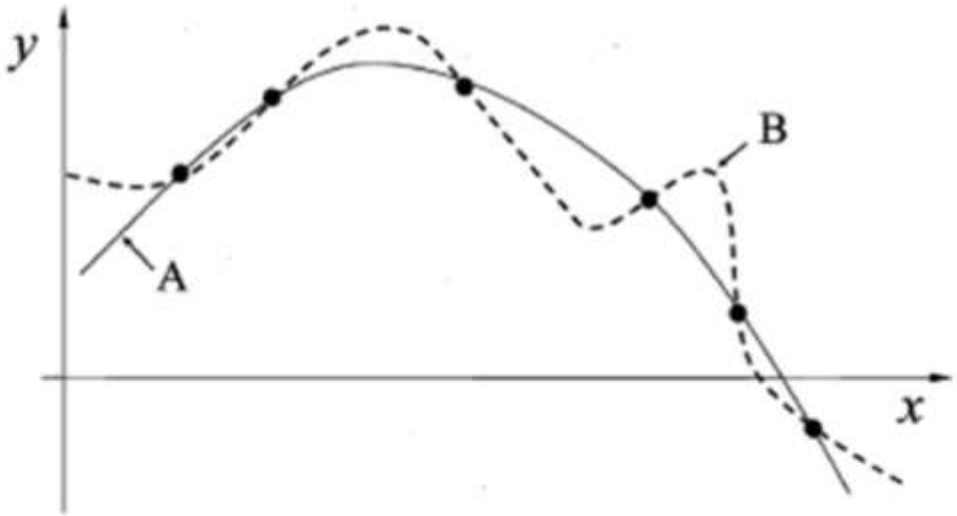
(5) 通过学习方法选择最优模型

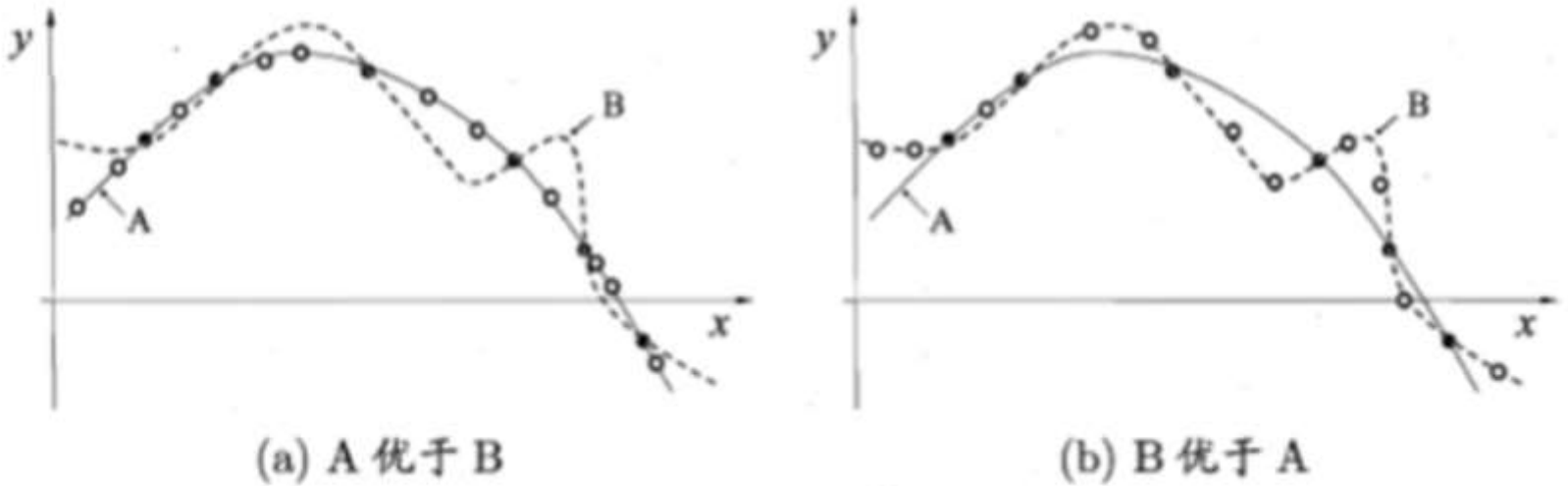
(6) 利用学习的最优模型对新数据进行预测和分析。

6. 归纳偏好

机器学习算法在学习过程中对某种类型假设的偏好. 假设偏好可看做学习算法自身在一个可能很庞大的假设空间中对假设进行选择的启发式或价值观.

奥卡姆剃刀(Occam's razor): 若有多个假设与观察一致，则选最简单的那个。





(a) A 优于 B　　(b) B 优于 A

定理1 (没有免费的午餐定理 (no free lunch theorem)).

对任意两个学习算法A和B, 若在某些问题上A比B好, 则必然存在另外一些问题B比A好.

证明: 为简单起见, 假设样本空间$\mathcal{X}$和假设空间$\mathbf{H}$都是离散的. 对某个特定的学习算法, 令$p(h|D,\mathcal{L}_a)$为基于训练数据D产生假设h的概率, 再令f代表我们希望学习的真实目标函数, 则该算法在训练集之外的所有样本上的期望误差为:

$$\begin{aligned} & E_{ote}(\mathcal{L}_a|D,f) \\ = \quad & E_h[E_{\mathbf{x}\in\mathcal{X}-D}[\mathbb{I}(f(\mathbf{x})\neq h(\mathbf{x}))]] \\ = \quad & \sum_{h\in\mathcal{H}}\sum_{\mathbf{x}\in\mathcal{X}-D}\mathbb{I}(f(\mathbf{x})\neq h(\mathbf{x}))p(\mathbf{x})p(h|D,\mathcal{L}_a) \end{aligned}$$

考虑二分类问题，且真实目标函数可以是任何函数

$$f:\mathcal{X}\to\{0,1\}$$

对所有可能的函数按照均匀分布对误差求和，有

$$
\begin{aligned}
& \textstyle\sum_f E_{ote}(\mathcal{L}_a|D,f) \\
= \; & \textstyle\sum_f \sum_h \sum_{\mathbf{x}\in\mathcal{X}-D} p(\mathbf{x})\mathbb{I}(f(\mathbf{x})\neq h(\mathbf{x}))p(h|D,\mathcal{L}_a) \\
= \; & \textstyle\sum_{\mathbf{x}\in\mathcal{X}-D} p(\mathbf{x}) \sum_h p(h|X,\mathcal{L}_a) \sum_f [\![h(\mathbf{x})\neq f(\mathbf{x})]\!] \\
= \; & \textstyle\sum_{\mathbf{x}\in\mathcal{X}-D} p(\mathbf{x}) \sum_h p(h|X,\mathcal{L}_a) \frac{1}{2} 2^{|\mathcal{X}|} \\
= \; & \textstyle\frac{1}{2} 2^{|\mathcal{X}|} \sum_{\mathbf{x}\in\mathcal{X}-D} p(\mathbf{x}) \sum_h p(h|X,\mathcal{L}_a) \\
= \; & \textstyle 2^{|\mathcal{X}|-1} \sum_{\mathbf{x}\in\mathcal{X}-D} P(\mathbf{x})\cdot 1
\end{aligned}
$$

也就是$\sum_f E_{ote}(\mathcal{L}_a|X,f) = \sum_f E_{ote}(\mathcal{L}_b|X,f)$

没有一种机器学习算法是适用于所有情况的。但事实上，f并不是均匀分布，学习算法自身的归纳偏好与问题是否匹配，往往会起到决定性的作用. 脱离具体问题，空泛地谈论“什么学习算法更好”毫无意义。

1. 机器学习的目标

机器学习的目的就是为了使得学习得到的映射f_θ逼近真相f.

- 不仅仅是训练误差小，更重要的是让泛化误差小，也就是提高学得模型f_θ适用于未见示例的能力

机器学习分为训练(Training)和测试(Testing)两个阶段：

- 训练：给定一个包含N个样本的训练集

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$$

调整模型f_θ的参数θ，使得预测结果$f_\theta(\mathbf{x}_n)$和y_n标注尽可能接近。

- 测试：将函数f_θ应用在一个独立的测试集

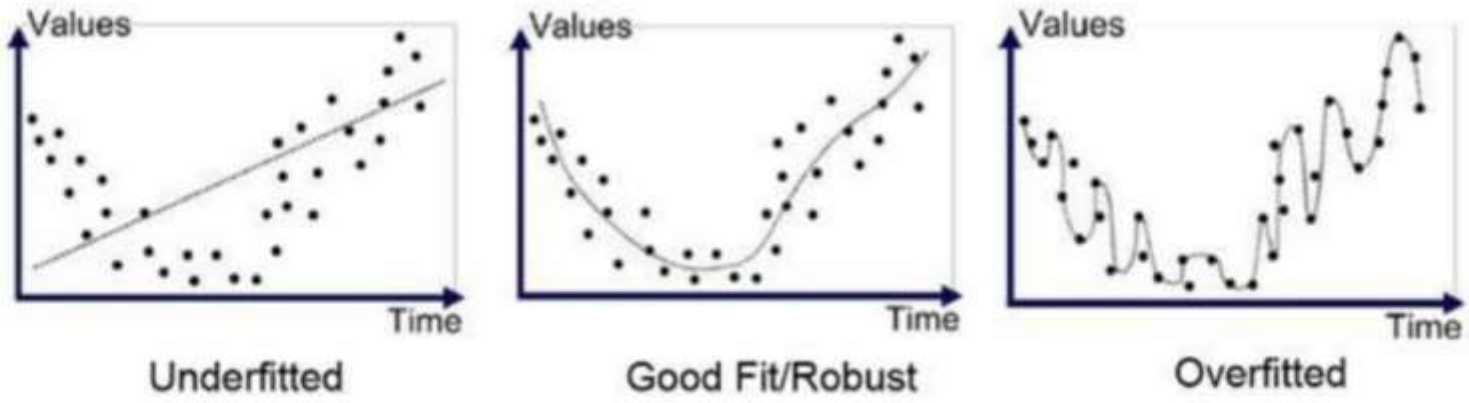
$$T = \{(\tilde{\mathbf{x}}_1, \tilde{y}_1), (\tilde{\mathbf{x}}_2, \tilde{y}_2), \ldots, (\tilde{\mathbf{x}}_{N'}, \tilde{y}_{N'})\}$$

验证所学到的函数能否在该数据集上做出准确预测，即$f_\theta(\tilde{x}_n)$是否与$\tilde{y}_n$接近

2. 欠拟合和过拟合

训练集上的表现	测试集上的表现	结论
好	好	Good fitting
不好	不好	欠拟合
好	不好	过拟合
不好	好	不太可能

图形化展示如下：



三、 欠拟合和过拟合

欠拟合(underfitting)). 学习器对训练样本的一般性质尚未学好，训练误差高，导致泛化能力低.

过拟合(overfitting). 学习器把训练样本学得“太好了”，把训练样本自身的一些特点当作了所有潜在样本都会有的一般性质，导致泛化能力下降

- **参数过拟合**: 模型训练过程中对参数调节得过于细致，导致对训练数据学习过度
  - 数据量过小，训练次数过多
- **结构过拟合**: 选择的模型过于复杂，以致对训练数据描述的过于精细
  - 模型学习能力过于强大，参数较多
  - 数据中噪声过大

三、 欠拟合和过拟合

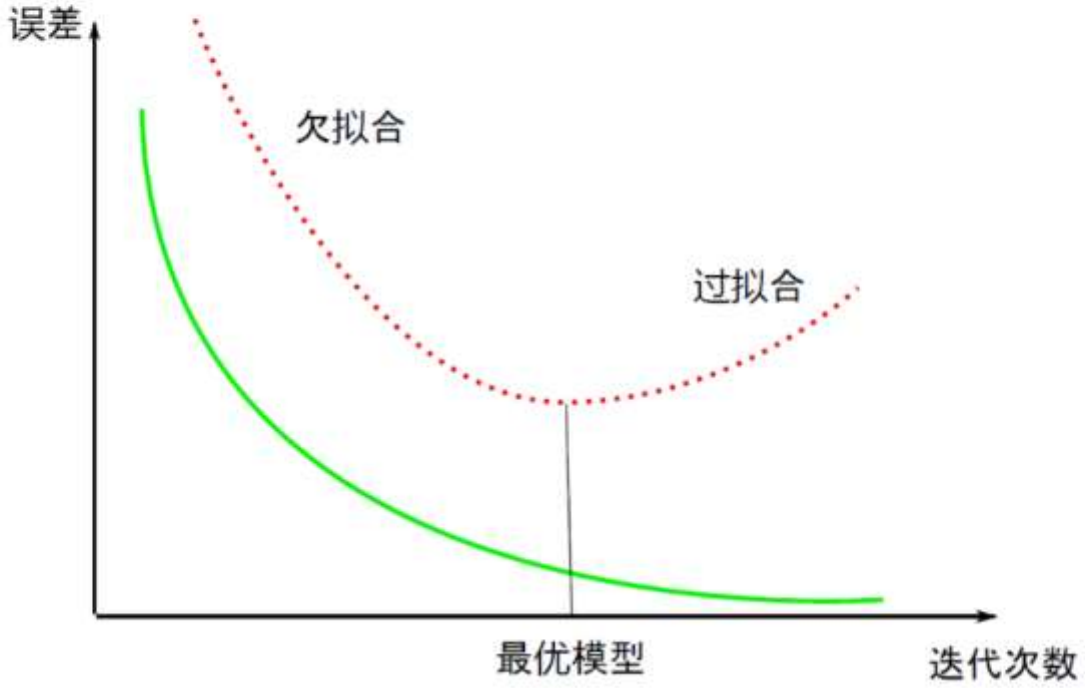
欠拟合的处理:

- 添加新特征
  - 当特征不足或者现有特征与样本标签的相关性不强时，模型容易出现欠拟合。通过挖掘组合特征等新的特征，往往能够取得更好的效果。
- 增加模型复杂度
  - 简单模型的学习能力较差，通过增加模型的复杂度可以使模型拥有更强的拟合能力。例如，在线性模型中添加高次项，在神经网络模型中增加网络层数或神经元个数等。
- 减小正则化系数
  - 正则化是用来防止过拟合的，但当模型出现欠拟合现象时，则需要有针对性地减小正则化系数。

三、 欠拟合和过拟合

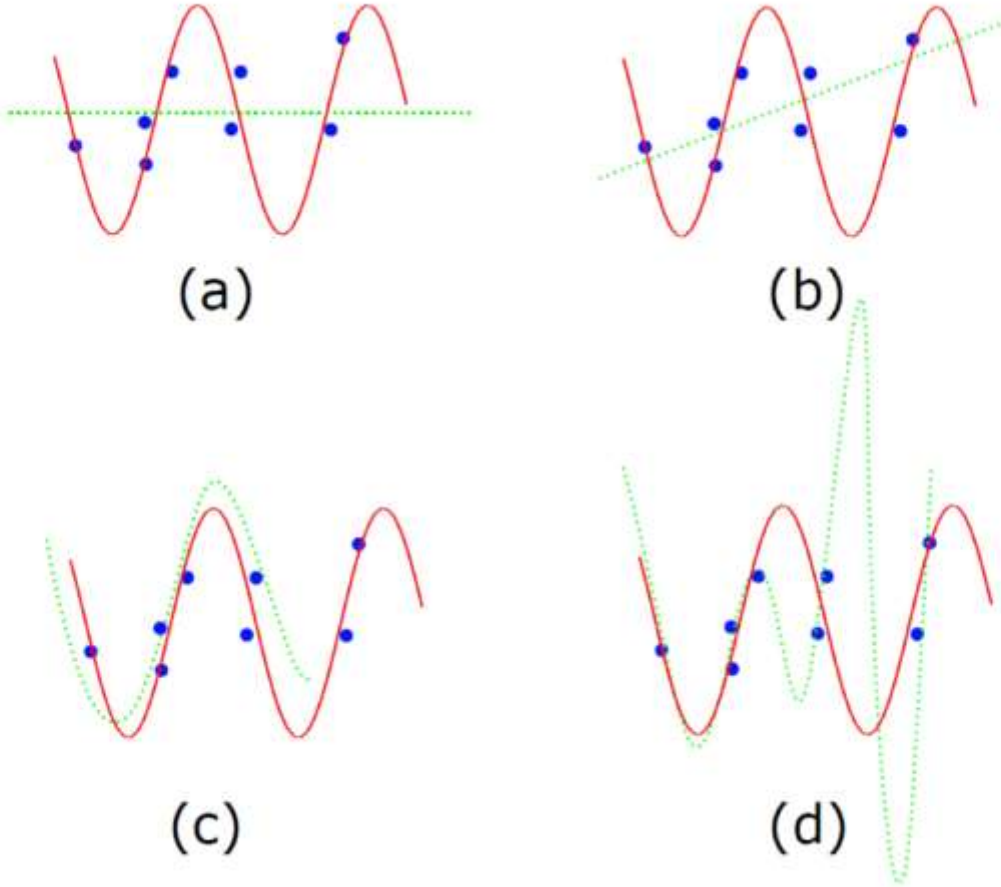
参数过拟合

- 在训练初期，模型参数还没有足够优化，因此在训练和测试数据上的性能都比较差，模型处于欠拟合状态；
- 当训练继续进行，在测试集上达到最优时，模型处于Good fitting状态
- 随着训练持续进行，模型对训练数据的学习越来越精细，模型进入过拟合状态



结构过拟合

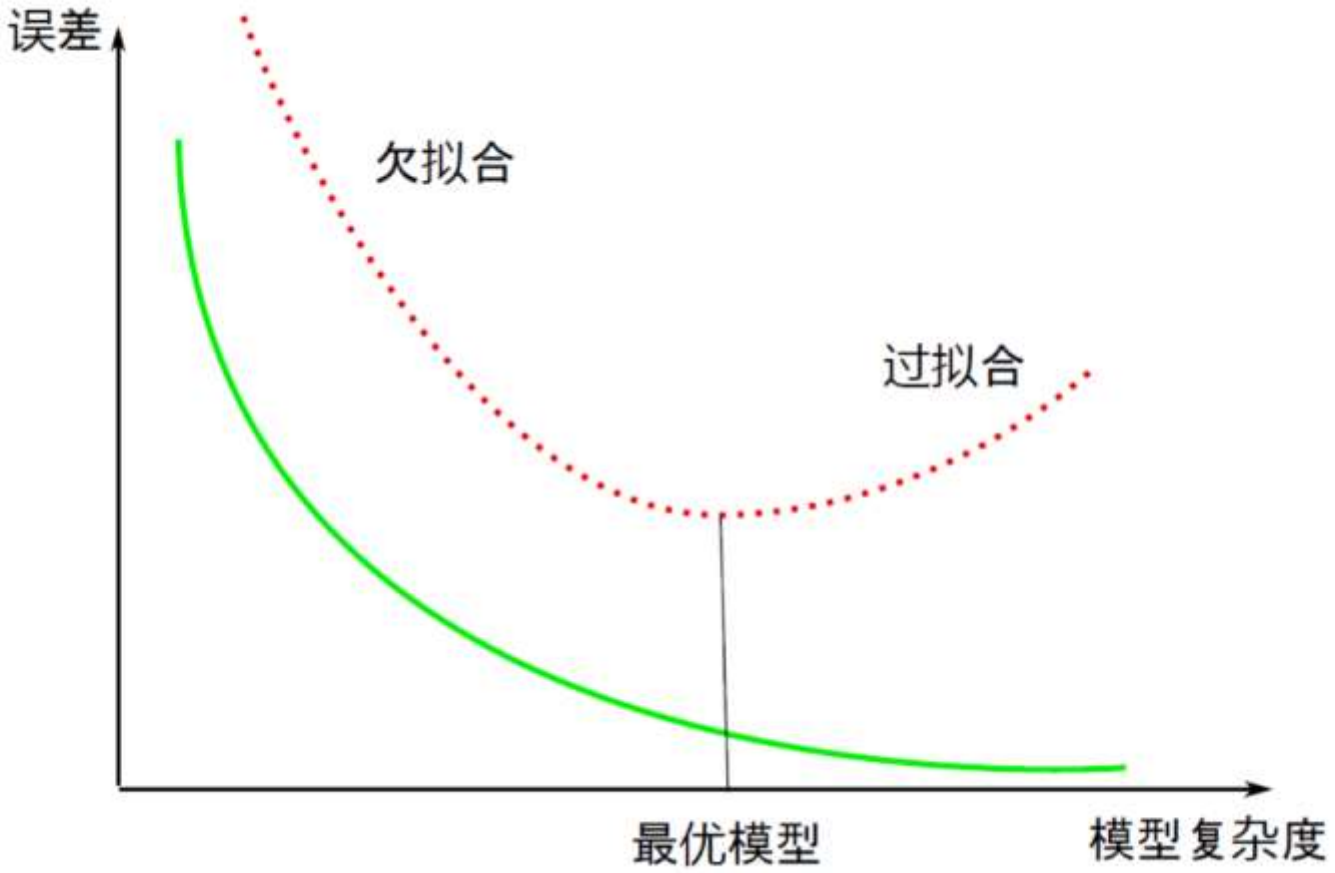
不同复杂度的模型对数据的描述能力不同，越复杂的模型对数据的描述能力越强，但产生过度学习的风险也越大，导致结构过拟合。



三、 欠拟合和过拟合

模型复杂度：模型复杂度较低容易产生过拟合，复杂度过高容易产生过拟合。

模型在训练集(绿色)和测试集(红色)上的表现



3. 泛化误差

定理: 偏差-方差-噪声分解(bias-variance decomposition)

设测试数据$\mathbf{x}$的真实目标值为$h(\mathbf{x})$, 观察到的目标值为t, 模型的预测值$y(\mathbf{x})$, 并记$\mathbf{x}$和t的联合分布为$p(\mathbf{x},t)$, 则目标值t与预测值$y(\mathbf{x})$之间的误差为

$$
\begin{aligned}
& \iint (y(\mathbf{x})-t)^2 p(\mathbf{x},t)d\mathbf{x}dt \\
= & \iint (y(\mathbf{x})-h(\mathbf{x})+h(\mathbf{x})-t)^2 p(\mathbf{x},t)d\mathbf{x}dt \\
= & \iint (y(\mathbf{x})-h(\mathbf{x}))^2 p(\mathbf{x},t)d\mathbf{x}dt + \iint (h(\mathbf{x})-t)^2 p(\mathbf{x},t)d\mathbf{x}dt \\
= & \text{预测误差} + \text{噪声}
\end{aligned}
$$

三、 欠拟合和过拟合

预测函数$y(\mathbf{x})$是通过某一数据集D训练出来的，将$y(\mathbf{x})$明确写作$y(\mathbf{x};D)$，模型预测的期望值为$\mathbb{E}_D[y(\mathbf{x};D)]$

$$
\begin{aligned}
& (y(\mathbf{x};D)-h(\mathbf{x}))^2 \\
= \quad & \{y(\mathbf{x};D)-\mathbb{E}_D[y(\mathbf{x};D)]\}^2+\{\mathbb{E}_D[y(\mathbf{x};D)]-h(\mathbf{x}))\}^2 \\
& +2(y(\mathbf{x};D)-\mathbb{E}_D[y(\mathbf{x};D)])(\mathbb{E}_D[y(\mathbf{x};D)]-h(\mathbf{x}))
\end{aligned}
$$

对数据集D取期望，有：

$$
\begin{aligned}
& \mathbb{E}_D\{y(\mathbf{x};D)-h(\mathbf{x})\}^2 \\
= \quad & \mathbb{E}_D\{y(\mathbf{x};D)-\mathbb{E}_D[y(\mathbf{x};D)]\}^2+\{\mathbb{E}_D[y(\mathbf{x};D)]-h(\mathbf{x}))\}^2 \\
= \quad & \text{方差}+\text{偏差}
\end{aligned}
$$

即：

$$\text{TotalError}=\text{Bias}+\text{Variance}+\text{Noise}$$

三、 欠拟合和过拟合

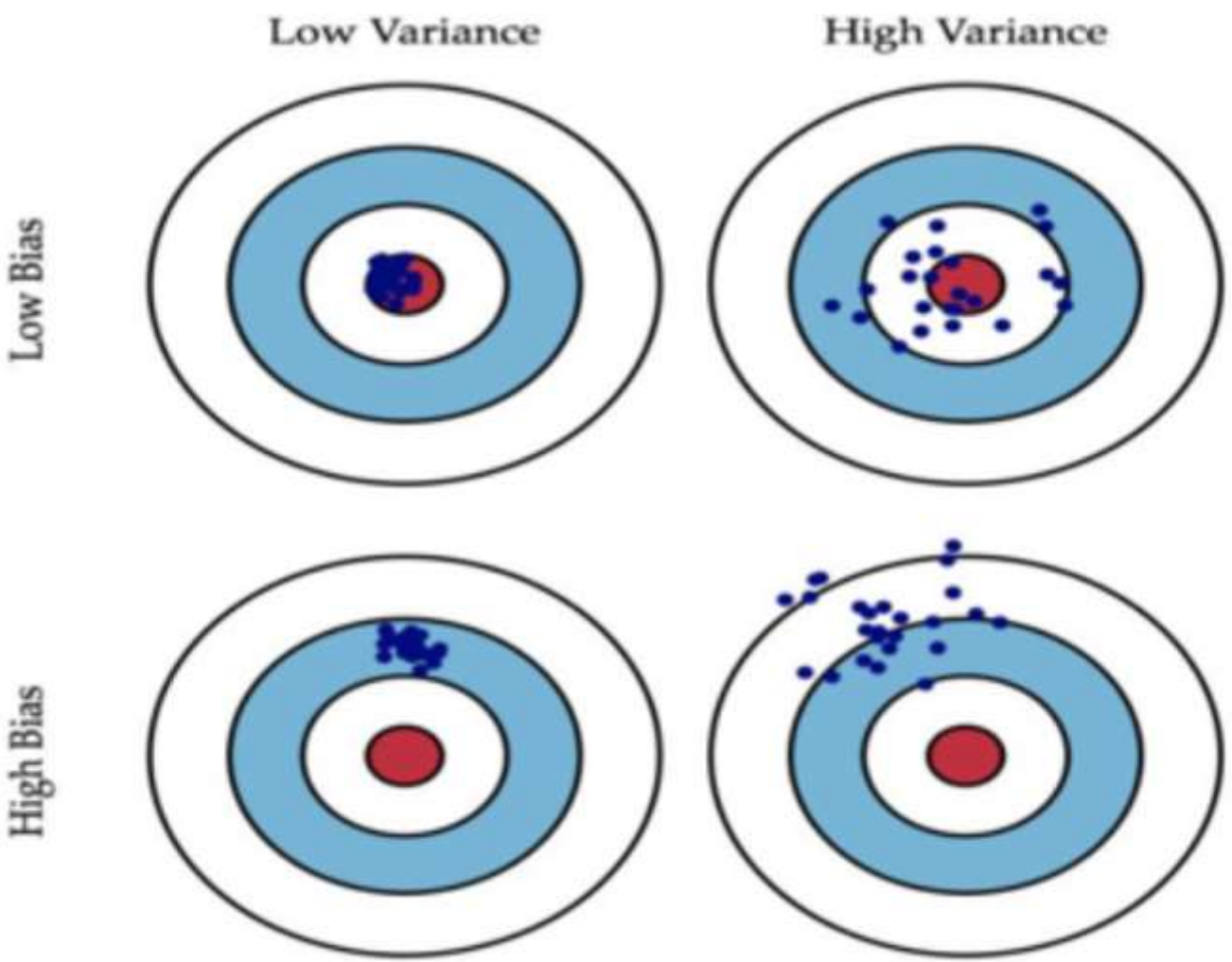
偏差-方差-噪声分解各项的含义.

- 偏差$\{\mathbb{E}_D[y(\mathbf{x};D)]-h(\mathbf{x})\}^2$: 模型的表达能力
  - 偏差度量了学习算法的期望预测$\mathbb{E}_D[y(\mathbf{x};D)]$和真实标记$h(\mathbf{x})$的偏离程度, 即刻画了学习算法本身的拟合能力.
- 方差$\mathbb{E}_D\{y(\mathbf{x};D)-\mathbb{E}_D[y(\mathbf{x};D)]\}^2$: 模型的泛化能力
  - 方差度量了同样大小的训练集的变动所导致的学习性能的变化, 即刻画了数据扰动所造成的影响，也即模型在某一数据集上训练出的模型在其他数据集上的有效性。
- 噪声$\mathbb{E}_D\{h(\mathbf{x})-t\}^2$: 学习任务本身的难度
  - 噪声表达了在当前任务上任何学习算法所能达到的期望泛化误差的下界, 即刻画了学习问题本身的难度.

在数据集给定的情况下，学习算法的泛化性能是由模型的表达能力、模型的泛化能力以及学习任务本身的难度共同决定的.

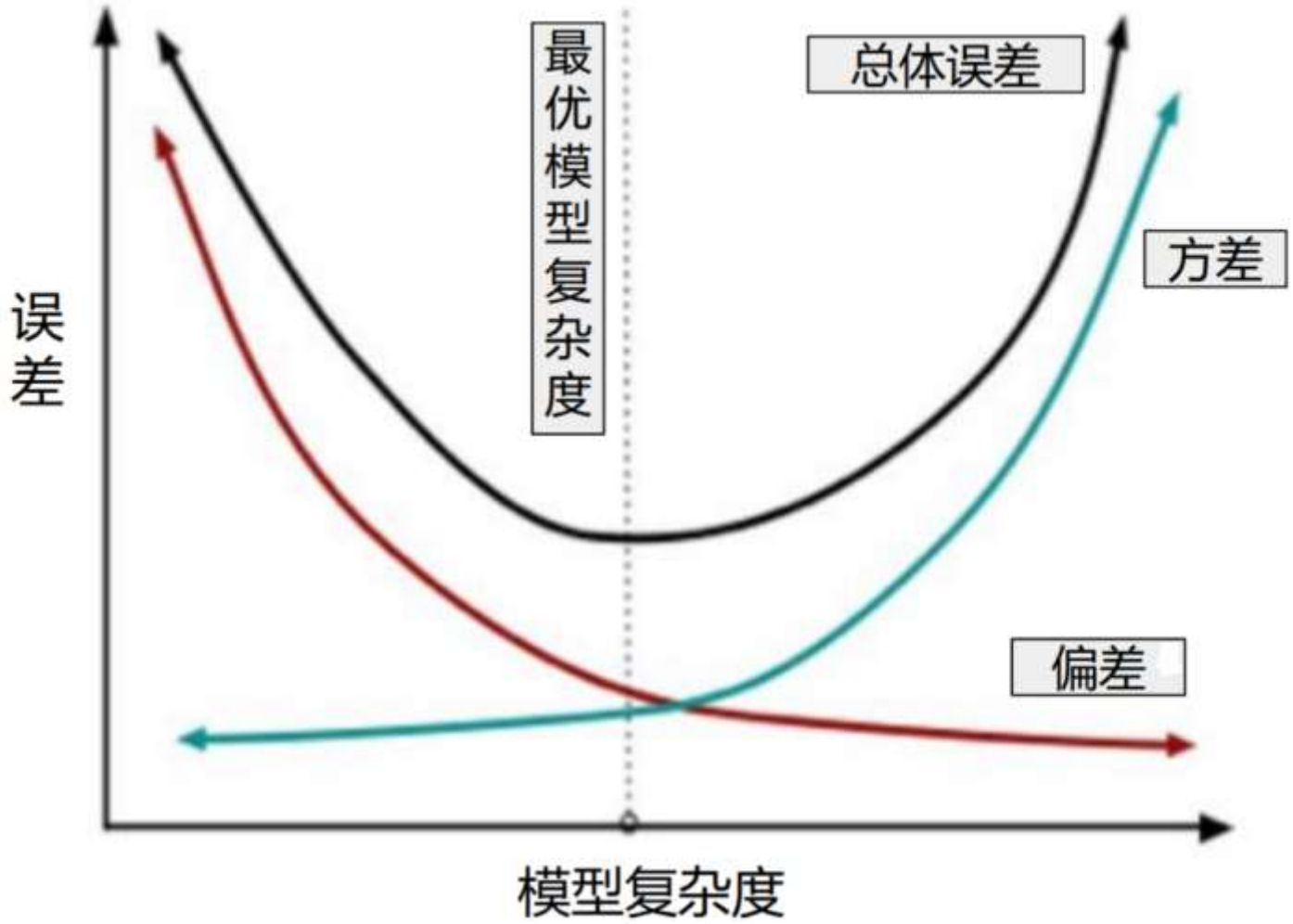
三、 欠拟合和过拟合

偏差方差图示：



三、 欠拟合和过拟合

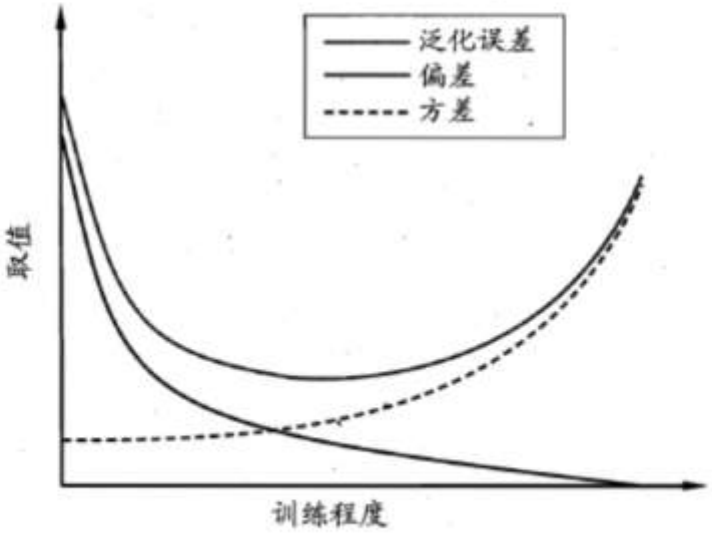
模型复杂度偏差-方差关系



三、 欠拟合和过拟合

偏差-方差窘境(bias-variance dilemma).

- 给定学习任务，在训练不足时，学习器的拟合能力不够强，训练数据的扰动不足以使学习器产生显著变化,此时偏差主导了泛化错误率。
- 随着训练程度的加深，学习器的拟合能力逐渐增强，训练数据发生的扰动渐渐能被学习器学到，方差逐渐主导了泛化错误率。
- 在训练程度充足后，学习器的拟合能力已非常强，训练数据发生的轻微扰动都会导致学习器发生显著变化，若训练数据自身的、非全局的特征被学习器学到了，则将发生过拟合。



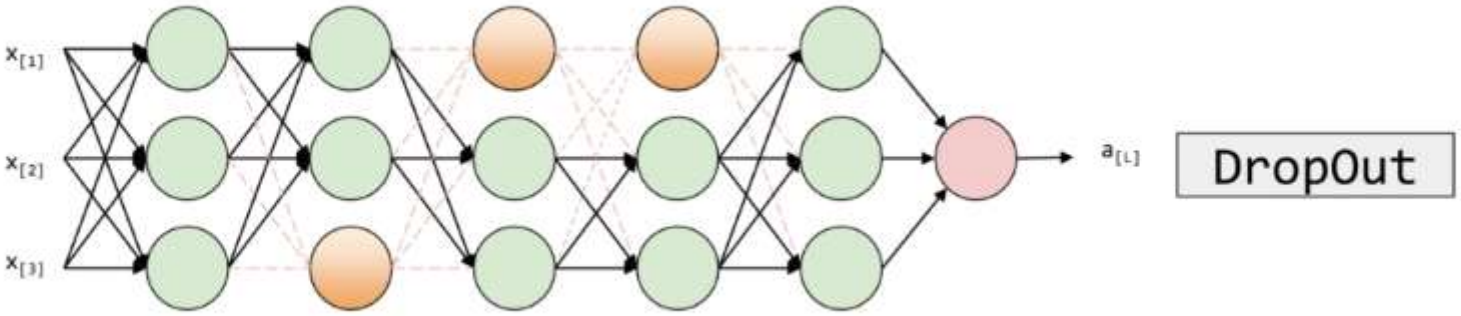
三、 欠拟合和过拟合

4. 缓解过拟合

- 从简单模型开始(奥卡姆剃刀原则)
- 正则化(可以降低模型有效的复杂度、减小测试误差)
  - 保留所有的特征，但是减少参数的大小
- 数据清洗和预处理(效果不一定)
- 数据扩充
  - 更多的样本能够让模型学习到更多更有效的特征，减小噪声的影响
  - 图像的数据增强
    - 随意翻转、镜像。
    - 随意裁剪。
    - 扭曲变形图片。
    - 颜色转换，然后给R、G和B三个通道上加上不同的失真值。产生大量的样本，进行数据增强。
- 验证集(用于对泛化误差进行估计，通常用于模型选择)

三、 欠拟合和过拟合

- 集成学习方法
  - 集成学习是把多个模型集成在一起，来降低单一模型的过拟合风险
- 早停(Early stopping)
  - 提早停止训练过程
  - 不能独立处理减少代价函数问题和过拟合问题，因为提早停止梯度下降，也就是停止了优化代价函数。
  - early stopping优点：只运行一次梯度下降，你可以找出w的较小值，中间值和较大值，而无需尝试正则化超级参数λ的很多值。
- 失活(dropout)



三、 欠拟合和过拟合

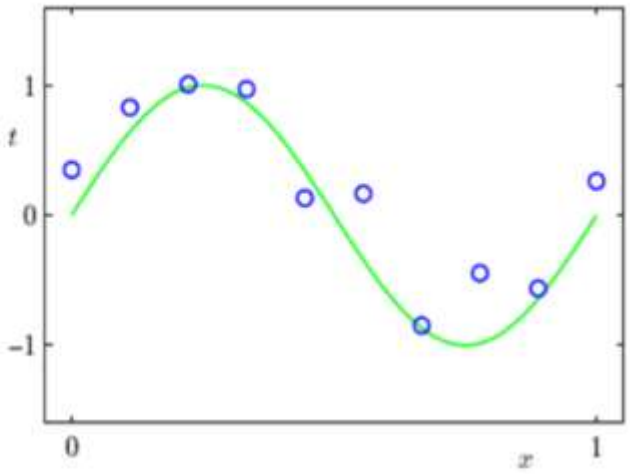
dropout为该网络每一层的神经元设定一个失活（drop）概率，（保留概率 keep-prob），在神经网络训练过程中，我们会丢弃一些神经元节点，在网络图上则表示为该神经元节点的进出连线被删除。

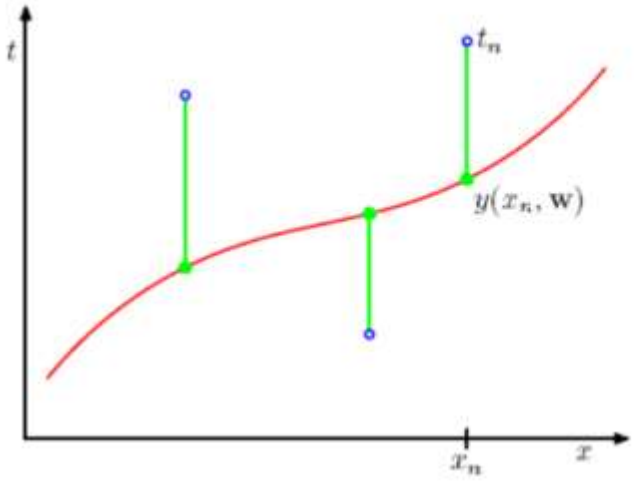
- keep-prob=1(没有dropout)
- keep-prob=0.5(常用取值，保留一半神经元)
- 在训练阶段使用，在测试阶段不使用!
- 对于不同神经元个数的神经网络层，我们可以设置不同的失活或者保留概率
- 对于单元数多、容易过拟合的层、含有较多权值的层，可以选择设置较大的失活概率（即较小的保留概率keep-prob）
- 对于不容易过拟合可以用大一点的keep-prob
- Dropout的功能类似于$L2$正则化，有收缩权重的效果

三、 欠拟合和过拟合

例2. 多项式曲线拟合问题

- 给定一个训练集，输入为$(x_1, x_2, \cdots, x_N)^T$，输出为$(y_1, y_2, \cdots, y_N)^T$，$N = 10$.

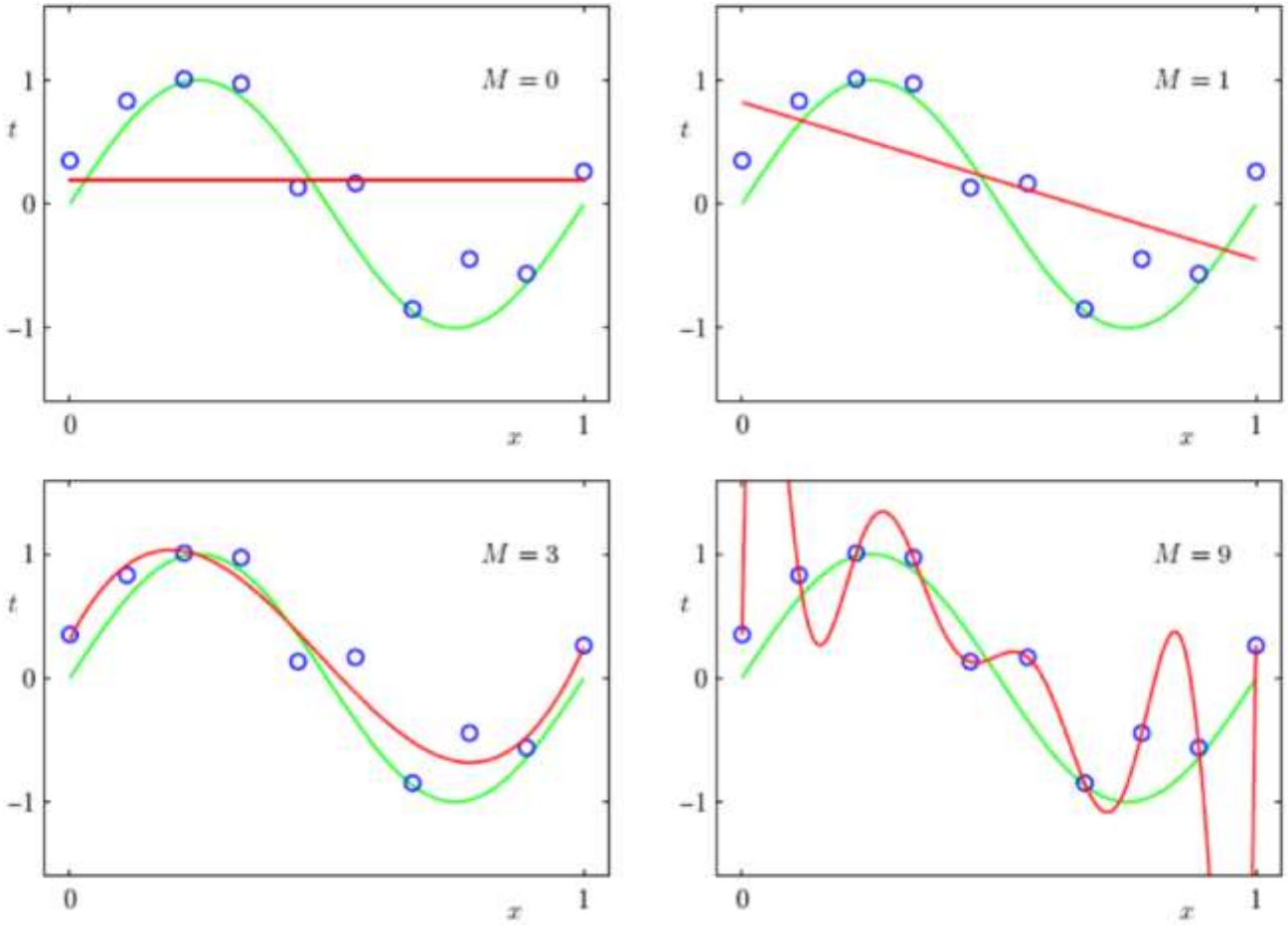




拟合多项式:

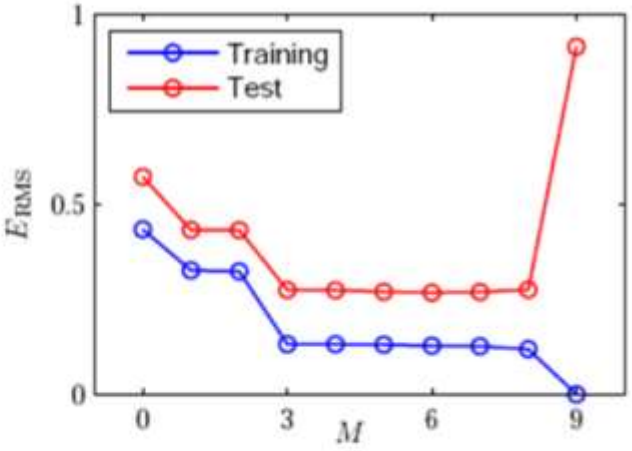
$$f(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M$$

三、 欠拟合和过拟合



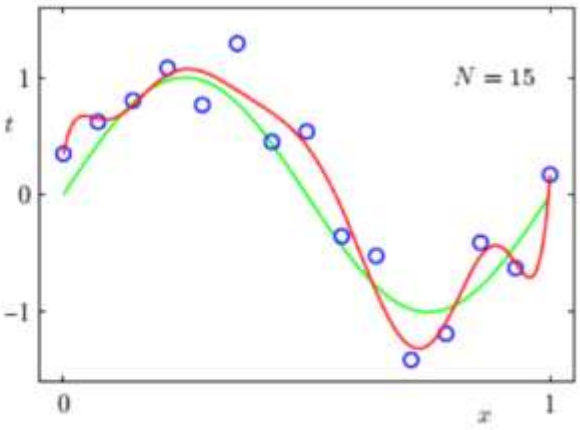
三、 欠拟合和过拟合

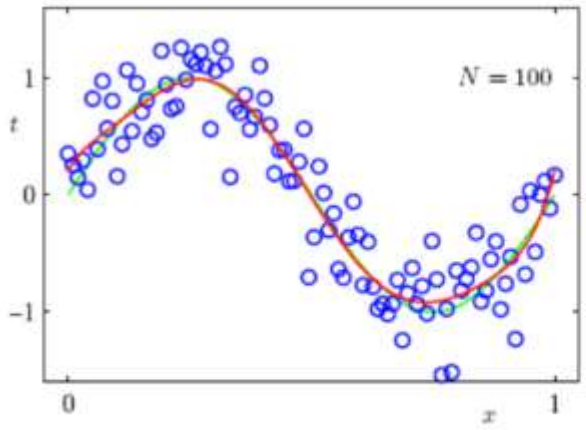
过拟合



	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

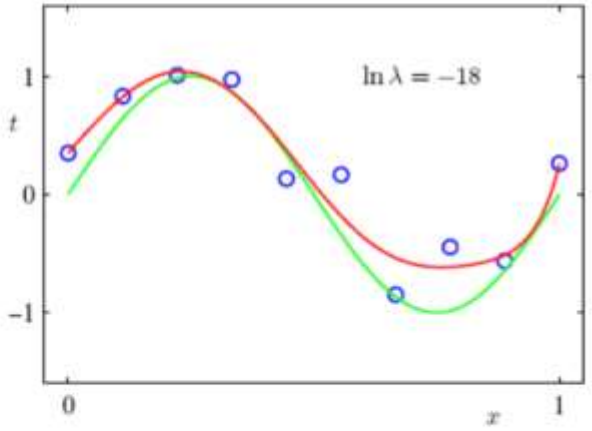
增加数据量

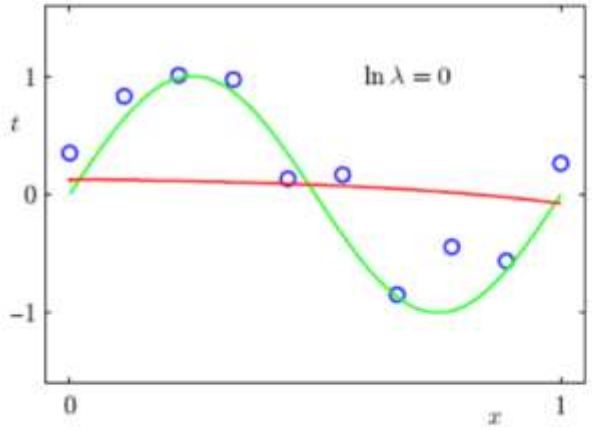




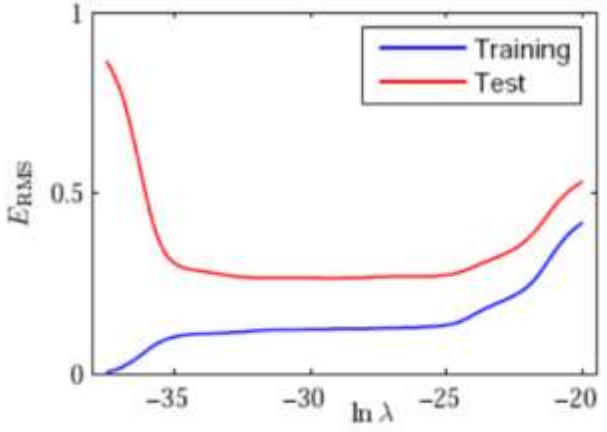
三、 欠拟合和过拟合

正则化：引入惩罚项





	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
w_0^*	0.35	0.35	0.13
w_1^*	232.37	4.74	-0.05
w_2^*	-5321.83	-0.77	-0.06
w_3^*	48568.31	-31.97	-0.05
w_4^*	-231639.30	-3.89	-0.03
w_5^*	640042.26	55.28	-0.02
w_6^*	-1061800.52	41.32	-0.01
w_7^*	1042400.18	-45.95	-0.00
w_8^*	-557682.99	-91.53	0.00
w_9^*	125201.43	72.68	0.01



5. 模型验证

训练集(Training Set):

- 帮助我们训练模型，简单的说就是通过训练集的数据让我们确定拟合曲线的参数。

测试集(testing set)

- 以测试集上的测试误差作为泛化误差的近似
- 用来测试学习器对新样本的判别能力，用于评估最终的模型
- 测试集需要和训练集互斥，假设测试样本也是从样本真实分布中独立同分布采样而得.

验证集(validation set): 训练超参数

- 把训练数据划分为训练集和验证集
- 基于验证集上的性能来进行模型选择和对超参数进行调参

三、 欠拟合和过拟合

三者划分：训练集、验证集、测试集

- 机器学习：60%，20%，20%；70%，10%，20%
- 深度学习：98%，1%，1% （假设百万条数据）

留出法(hold-out):

- 直接将训练数据划分为两个互斥的集合:训练集和验证集
- 训练/验证集的划分要尽可能保持数据分布的一致性
- 采用若干次随机划分，每次产生一对训练/验证集用于实验评估，最后报告所有结果的平均值和标准差
- 一般将大约2/3~4/5作为训练数据

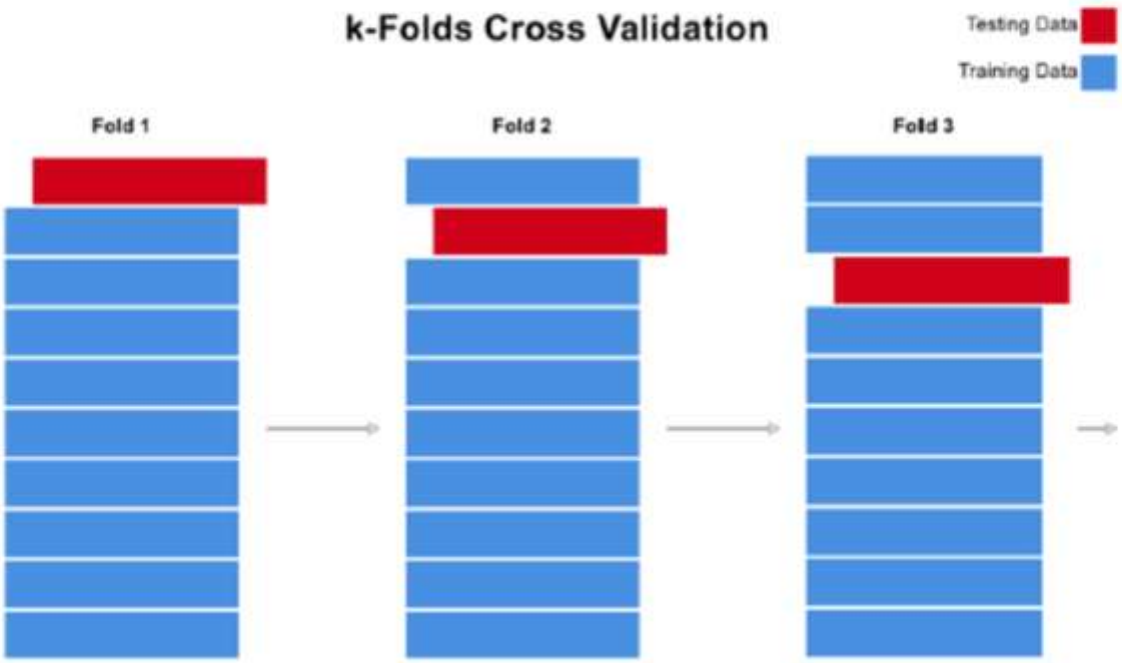
交叉验证

- 随机将训练数据划分为K个互不相交、大小相同的子集
- 利用$K-1$个子集的数据训练模型，余下的子集作为验证集；
- 将这一过程对可能的K种选择重复进行
- 最后选个K次评测中平均测试误差最小的模型

注意事项:

- K最常用取值是10, 其他常用取值是5和20
- 每个子集都尽可能保持数据分布的一致性.
- 为减小因样本划分不同而引入的差别, 一般采用若干次交叉验证, 如10次10折交叉验证.

三、 欠拟合和过拟合



留一法(leave-one-out)

- K折交叉验证在$K = m$时的特例
- 由于每个子集只有一个样本，留一法不受随机样本划分方式的影响.
- 留一法的评估结果往往被认为比较准确. 但是，在数据集比较大时，训练m个模型的计算开销是难以忍受的

三、 欠拟合和过拟合

学习	开车
过拟合	车祸
模型过于复杂，参数过多	车开太快
数据有noise	道路崎岖
数据量不够	对路况的了解程度不够
从简单模型开始	先慢慢开
数据清洗/修剪	使用更加精确的道路信息
数据提示	利用更多的道路信息
正则化	踩刹车
验证	观察仪表盘
特征转换	踩油门

三、 欠拟合和过拟合

例3. 分别给出解决高方差和高偏差的方法

解:

- 解决高方差(过拟合)
  - 获得更多的训练实例
  - 尝试减少特征的数量
  - 尝试增加正则化程度λ
- 解决高偏差(欠拟合)
  - 尝试获得更多的特征
  - 尝试增加多项式特征
  - 尝试减少正则化程度λ

四、机器学习的分类

1. 按有无标签分类

- 监督学习：垃圾邮件分类、房价预测
  - 数据集有标注
- 非监督学习：异常检测
  - 训练数据没有标注
- 半监督学习：标注语音
  - 结合（少量的）标注训练数据和(大量的)未标注数据来进行学习
  - **聚类假设**：相同聚类中的样本有较大可能拥有相同的标记
  - **流形假设**：处于一个很小的局部区域内的样本具有相似的性质
- 强化学习：Alpha GO
  - 外部环境对输出只给出评价信息而非正确答案，智能体通过强化受奖励的动作来改善自身的性能

四、机器学习的分类

其他学习方法:

归纳学习(概念学习、经验学习)：给定关于某个概念的一系列已知的正例与反例，其任务是从中归纳出一个一般的概念描述。

- 泛化Generalization用来扩展一假设的语义信息，以使其能够包含更多的正例，应用于更多的情况。
- 特化Specialization是泛化的相反的操作，用于限制概念描述的应用范围

多任务学习（Multi-task Learning）：把多个相关的任务放在一起同时学习

- 多个任务之间共享一些因素，它们可以在学习过程中，共享它们所学到的信息，相关联的多任务学习比单任务学习具备更好的泛化效果

四、机器学习的分类

2. 按输出空间分类

- 二分类：垃圾邮件分类
- 多分类：图像分类
- 回归：房价预测
- 结构化学习：机器翻译、语音识别、聊天机器人

Dog

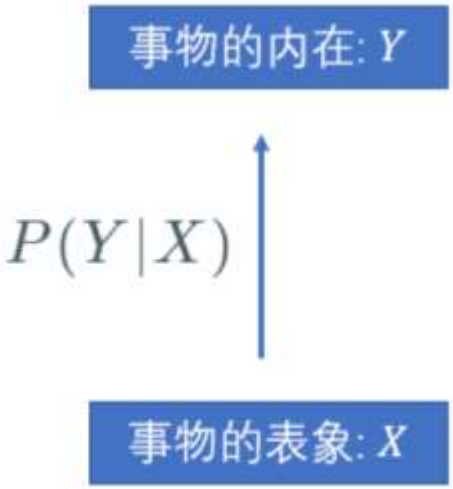
Cat

四、机器学习的分类

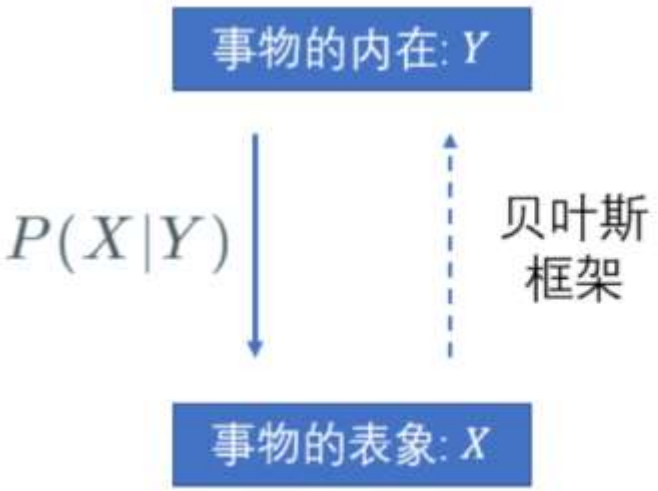
3. 按模型分类

- 生成式模型：GAN
  - 先确定$P(\mathbf{x}, y)$
  - 然后利用贝叶斯定理：$P(y|\mathbf{x}) = \frac{P(\mathbf{x},y)}{P(\mathbf{x})}$
- 判别式模型: 决策树、支持向量机
  - 直接确定$P(y|\mathbf{x})$或$f(\mathbf{x})$

判别式模型：



生成式模型：



四、机器学习的分类

4. 按算法分类

- 批量学习：一次性批量输入给学习算法，可以被形象的称为填鸭式学习
- 在线学习：按照顺序，循序的学习，不断的去修正模型，进行优化
- 主动学习：通过某种策略找到未进行类别标注的样本数据中最有价值的数据，交由专家进行人工标注后，将标注数据及其类别标签纳入到训练集中迭代优化分类模型，改进模型的处理效果