统计机器学习
Statistical Machine Learning

魏莱

上海海事大学信息工程学院

2019 年 3 月 11 日

参考书籍

- 教材：周志华，《机器学习》，清华大学出版社
- 参考教材：
  - 李航，《统计学习方法》，清华大学出版社
  - Tom Mitchell, 《机器学习》，机械工业出版社
  - Andrew Ng，CS229:Machine learning, http://cs229.stanford.edu/
  - Christopher Bishop，《Pattern Recognition and Machine Learning》, Springer
  - Trevor Hastie etc.，《The Elements of Statistical Learning》，Springer

第一章：绪论

1. 基本术语

- **样本 (sample)/示例 (instance)**: 对事物的抽象化描述
  - 李明:(学号 = “201610311”, 专业 = “计算机”, 年龄 = “23”)
  - $\mathbf{x}_i : (x_{i1}, x_{i2}, x_{i3})$
- **属性 (attribute)/特征 (feature)**: 事物某方面的描述
  - 学号 x_{i1}, 专业 x_{i2}, 年龄 x_{i3}, $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})^t$
  - 属性值: 属性上的取值。x_{i2} = 计算机, $x_{i3} = 23$
- **特征向量 (feature vector)**: 样本对应于属性空间/样本空间的某一个点
  - $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})^t$
- **维数 (dimensionality)**: 特征向量中分量的个数
  - $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id})^t$,$d$ 即为维数

1. 基本术语

- **数据集 (Database/Data matrix)**: 所有数据样本/特征向量组成的集合
  - $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$,$n$ 为数据样本个数
- **标签 (label)**: 对某个样本的某种标记
  - $\mathbf{x}_i \to y_i, \quad y_i = \{-1, +1\}/\{0, 1\}$
  - $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n) \to \mathbf{y} = (y_1, y_2, \cdots, y_n)$
  - $\mathbf{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\}$

1. 基本术语

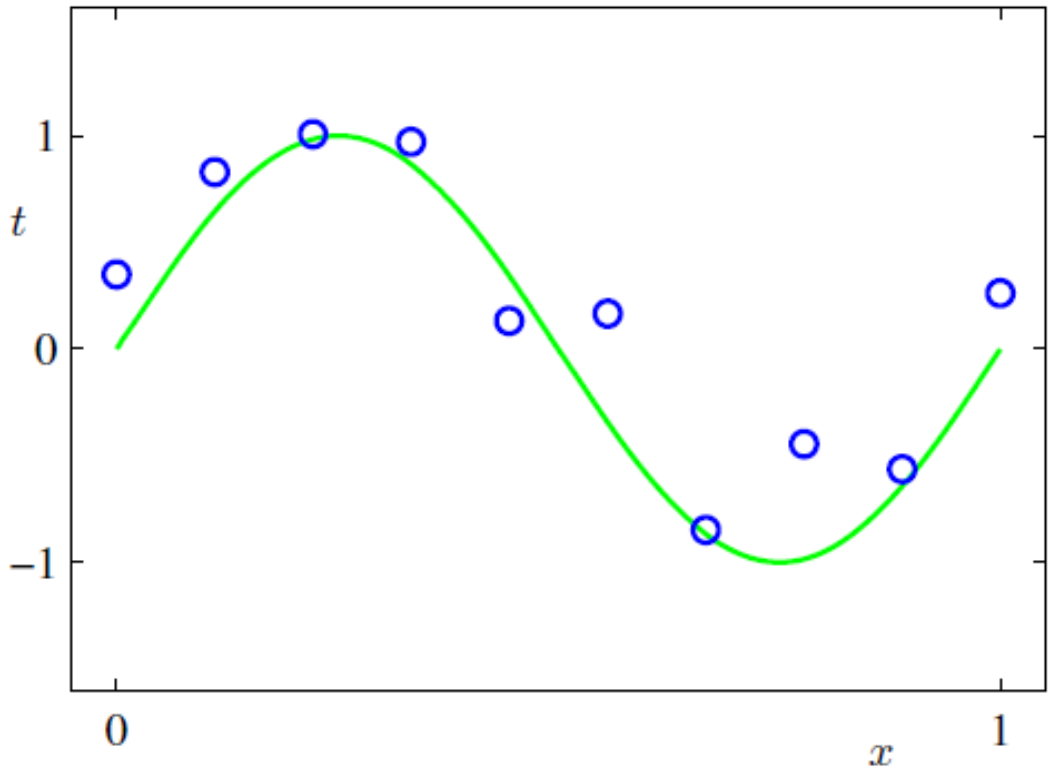
- **(机器) 学习 (machine learning)**: 从给定的数据集中学习出某种模型的算法
- **训练集 (training set)**，$\mathbf{X}_t$: 数据集中用来学习模型的部分数据
- **测试集 (testing set)**，$\mathbf{X}_s$: 数据集中用来测试学习到模型性能的部分数据，因此我们有：
  - $\mathbf{X} = \mathbf{X}_t + \mathbf{X}_s$
  - $\mathbf{X} = \mathbf{X}_t + \mathbf{X}_v + \mathbf{X}_s$，$\mathbf{X}_v$ 为验证集, 常常用来调试学得的模型中某些参数

2. 学习的分类

- **无监督学习 (unsupervised learning)**: 学习过程中没有用到样本标签，即训练数据集 $\mathbf{X}_t$ 不带标签
  - 聚类 (clustering)
- **有监督学习 (supervised learning)**：学习过程中用到样本标签
  - 分类 (classification)、回归 (regression)
- **半监督学习 (semi-supervised learning)**：训练数据部分带标签，部分不带标签

3. 模型的选择

例 1: 考虑这样一个问题：现在有 10 个样本点组成训练集合，每一样本点 (x, t) 为二维平面上一个坐标点。这些数据点是由函数 $t = \sin(2\pi x)$ 加入一些随机噪声生成。我们的目的是**在不知道绿色曲线的情况下，学习出生成曲线。**



考虑多项式拟问题：

$$y = f(x, \mathbf{w}) = w_0 + w_1 x^1 + w_2 x^2 + \cdots + w_M x^M = \sum_{j=1}^{M} w_j x^j \quad (1)$$

其中 M 为多项式阶数。

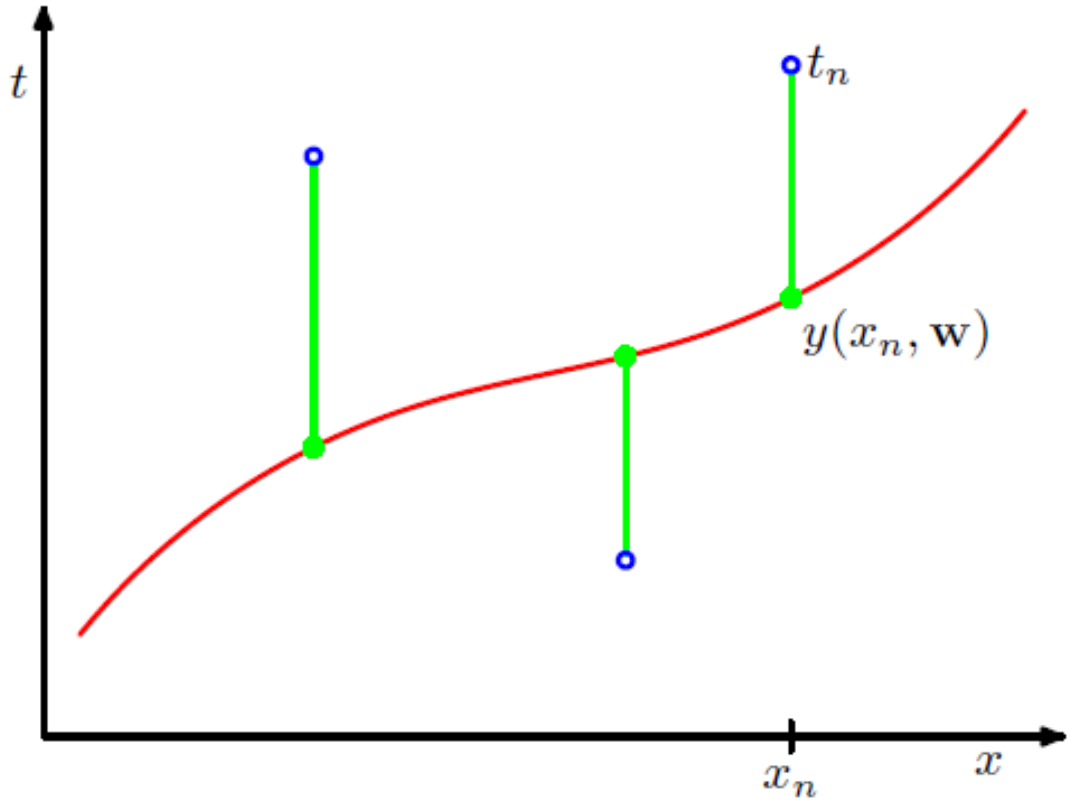
假设 $\mathbf{w} = (w_0, w_1, \cdots, w_M)^t$ 已知，那么有一个 x_i 可以通过上式求得 y_i。现在 $\mathbf{w}$ 未知，如果上式能够很好的拟合曲线，那么对 $\forall i$，有：

$$y_i \approx t_i \Leftrightarrow (y_i - t_i)^2 \approx 0$$

定义误差函数:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(f(x_i, \mathbf{w}) - t_i)^2 \tag{2}$$

上式称为平方误差函数。



现在有样本点集合 $\{(x_1, t_1), \cdots, (x_n, t_n)\}$，通过最小化式(1)，即

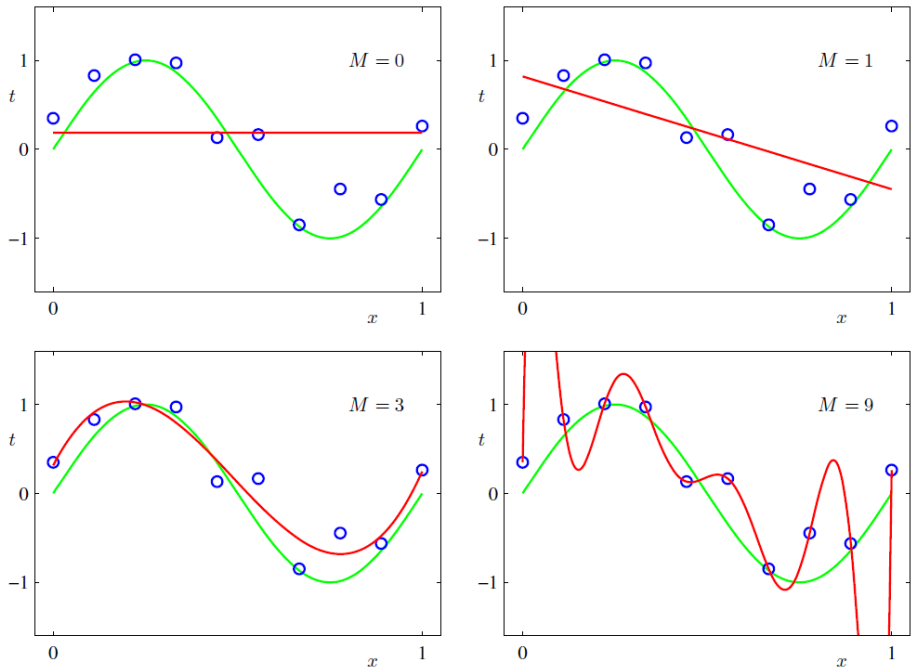
$$\min E(\mathbf{w}) = \min \frac{1}{2} \sum_{i=1}^{n} (f(x_i, \mathbf{w}) - t_i)^2$$

可否求得 $\mathbf{w}$?

3. 模型的选择

$\mathbf{w} = (w_0, w_1, \cdots, w_M)^t$, 其中 M 为多项式阶数。不同的 M 导致多项式

$f(x, \mathbf{w}) = w_0 + w_1 x^1 + w_2 x^2 + \cdots + w_M x^M = \sum_{j=1}^{M} w_j x^j$ 不同。



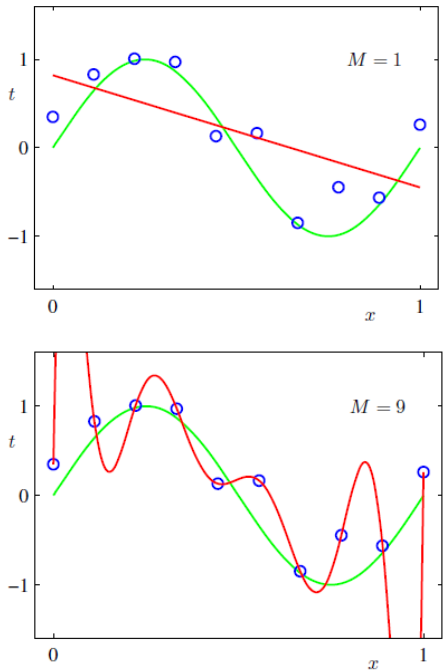
很显然，当 $M = 3$ 时，学得的多项式能够最好的拟合数据集。

3. 模型的选择

- 拟合问题

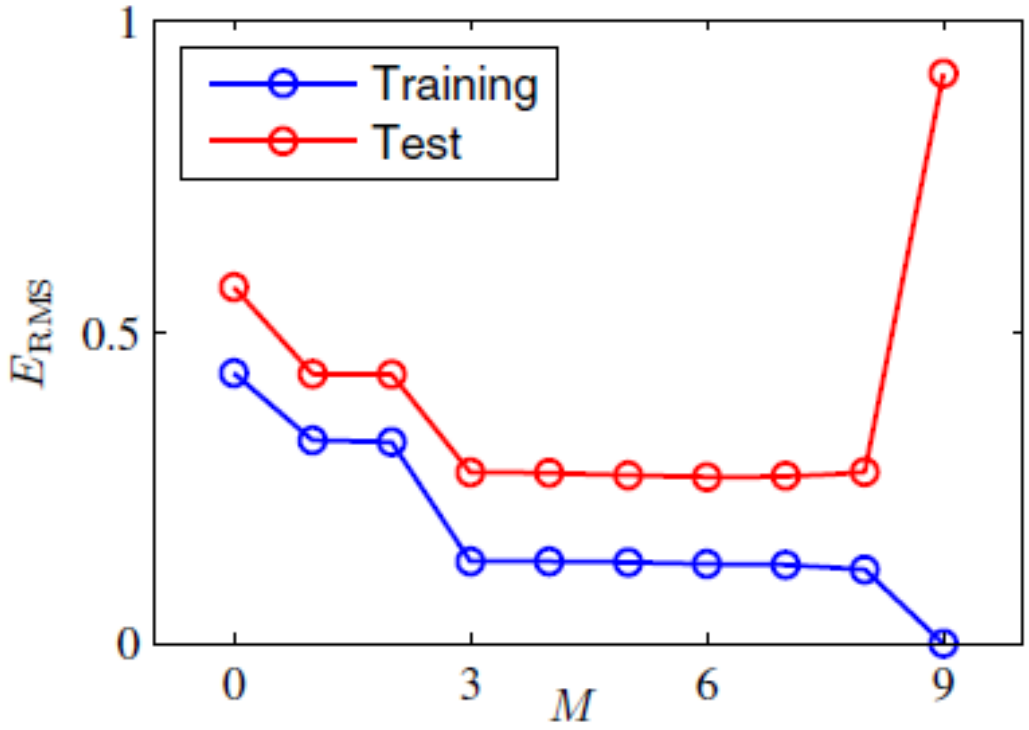
> **欠拟合 (under-fitting)**: 学得的模型不足以描述数据分布。训练数据误差较大。模型过于简单。

> **过拟合 (over-fitting)**: 学得的模型对训练数据过分匹配精确，导致无法适应新的数据，测试数据误差较大。模型过于复杂。



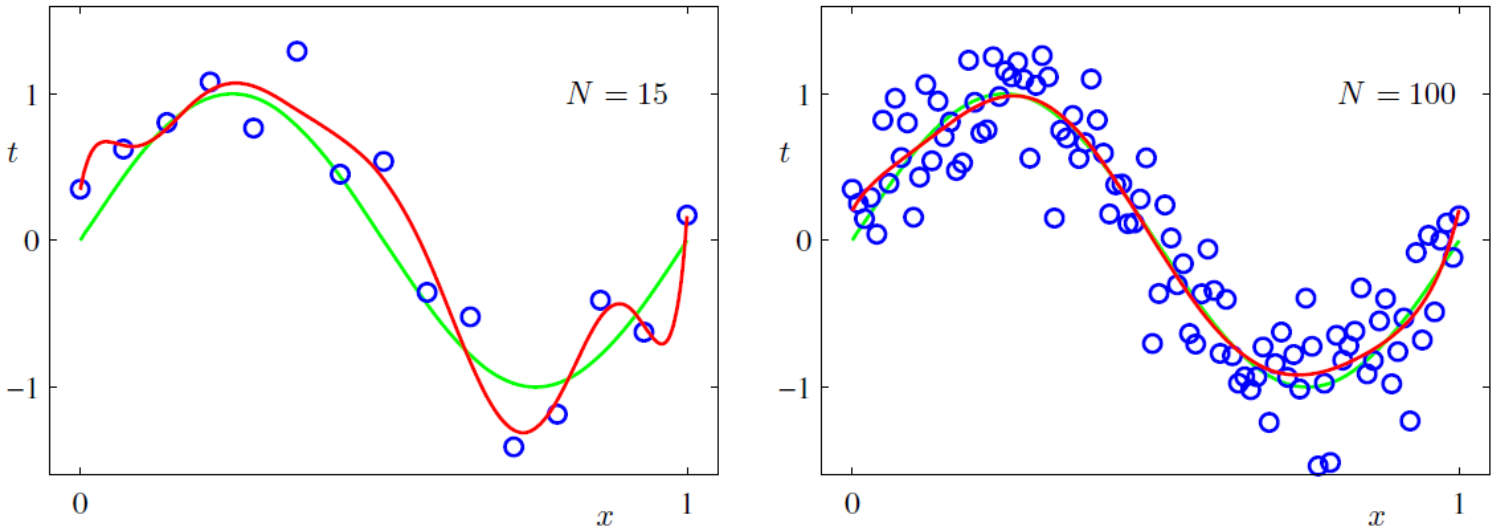
3. 模型的选择

假设现在我们除了上述的 10 个数据样本，又由 $\sin(2\pi x)$ 产生 90 个新的数据样本作为测试集，则可以根据不同的 M 值，来分别计算训练集误差及测试集误差，如下图所示：



其中 $E_{RMS} = \sqrt{2E(\mathbf{w})/n}$ 称为均方误差。

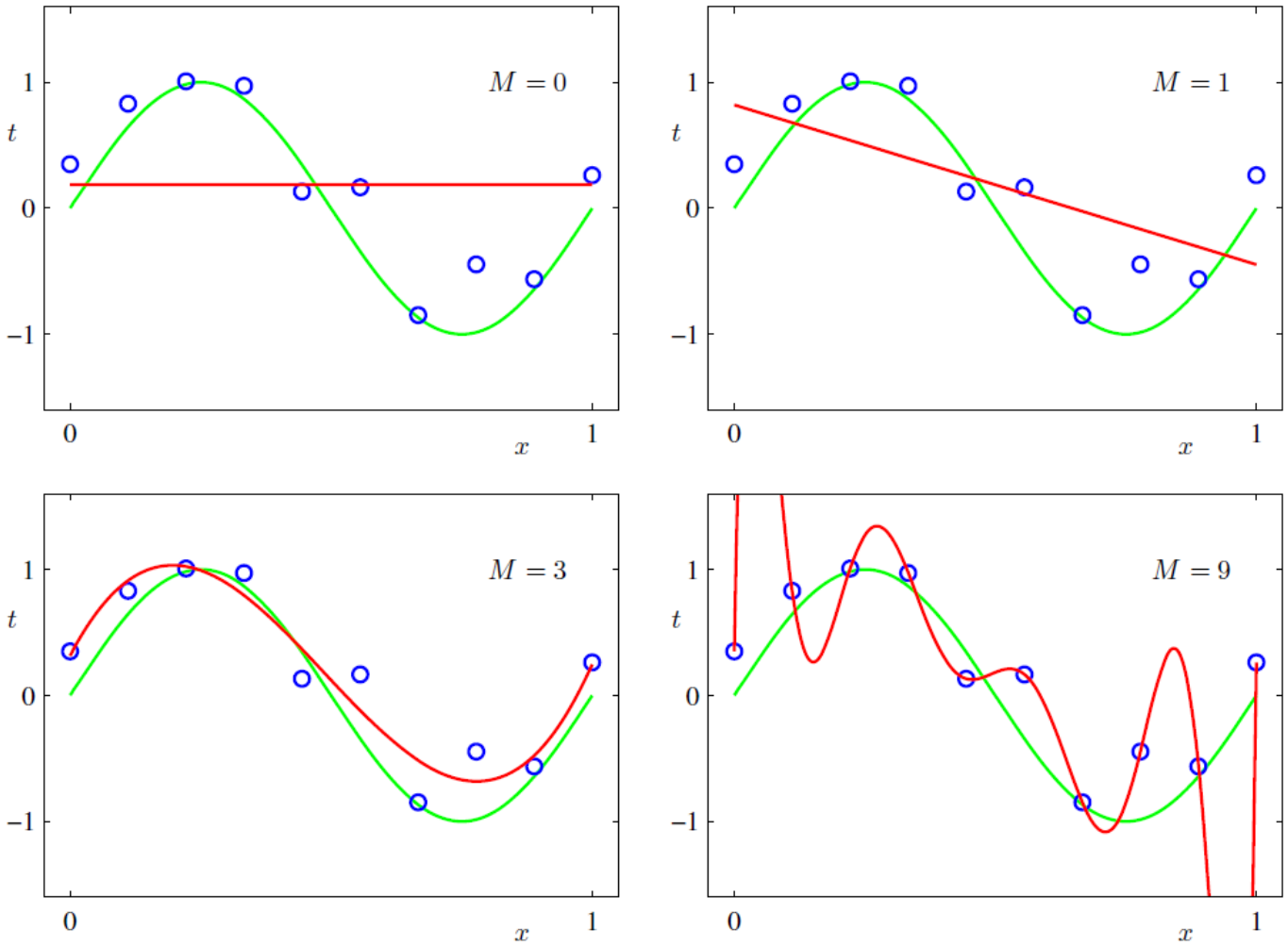
那么，当什么样的情况下 $M = 9$，是合适的呢？



结论：当训练样本较多，则模型可以相对复杂一点，当训练样本较少，则模型应该更为简单。

3. 模型的选择

那么，对于一个数据集，我们到底如何选择模型？



假设我们固定 $M = 9$, 那么对于前三张图来说其对应的 $\mathbf{w}$，分别应该为:

$$\mathbf{w}_1 = (0, 0, \cdots, 0)^t$$

$$\mathbf{w}_2 = (w_0, w_1, 0, \cdots, 0)^t$$

$$\mathbf{w}_3 = (w_0, w_1, w_2, w_3, 0, \cdots, 0)^t$$

即：**对于多项式模型来说，模型越复杂，其参数 $\mathbf{w}$ 中非零元个数越多。**

由此，我们在学习 $\mathbf{w}$ 时，我们可以在原有误差函数上增加对 $\mathbf{w}$ 的控制

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(f(x_i, \mathbf{w}) - t_i)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \tag{3}$$

这里 λ 为一参数 (parameter)，
$\|\mathbf{w}\|^2 = \mathbf{w}^t\mathbf{w} = w_0^2 + w_1^2 + \cdots + w_M^2$，称为**回归算子**。回归算子有很多种类型，这里称为 L_2 范数回归，也叫做岭回归 *(ridge regression)*。

4. 模型的评估方法

由训练集学习出某个模型后，通常需要使用测试集对学得的模型进行评估。那么对于一个给定的数据集 $\mathbf{X}$，如何得到训练集 $\mathbf{X}_t$ 和测试集 $\mathbf{X}_s$？

4. 模型的评估方法

- **留出法 (hold-out)**：直接将 $\mathbf{X}$ 分解为两个不想交的集合，其中一个作为训练集，另一个作为测试集。常常将 $\frac{2}{3}--\frac{4}{5}$ 的样本用于训练，其余用于测试。
- **交叉验证 (cross validation)**: 将数据集 $\mathbf{X}$ 分解为 k 个互补相交的子集，即 $\mathbf{X}_1 \bigcup \mathbf{X}_2 \bigcup \cdots \bigcup \mathbf{X}_k = \mathbf{X}$。然后每次用 $k-1$ 个子集训练，剩余一个做测试，最终返回 k 个测试结果——k 折交叉验证。
  - 留一法 (Leave-One-Out): 假设 $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$, 满足 $k = n$。

5. 错误率与精度

在分类任务中，精度和错误率是最常用的模型性能度量。现在假设数据集 $\mathbf{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\}$，学习得到的模型为 $f : \mathbf{x} \to y$。

- 错误率：$E(f; D) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(f(\mathbf{x}_i \neq y_i)$
- 精度：$acc(f; D) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(f(\mathbf{x}_i = y_i) = 1 - E(f; D)$

其中，$\mathbb{I}$ 代表一个指示函数。

Thanks !

My Email：weilai@shmtu.edu.cn

My office: Room 212, College of Information Engeering