模式识别与机器学习

Pattern Recognition and Machine Learning

Ding Zhang

March 2019

1 入门、概率论、决策论及信息论

1.1 机器学习的知识框架

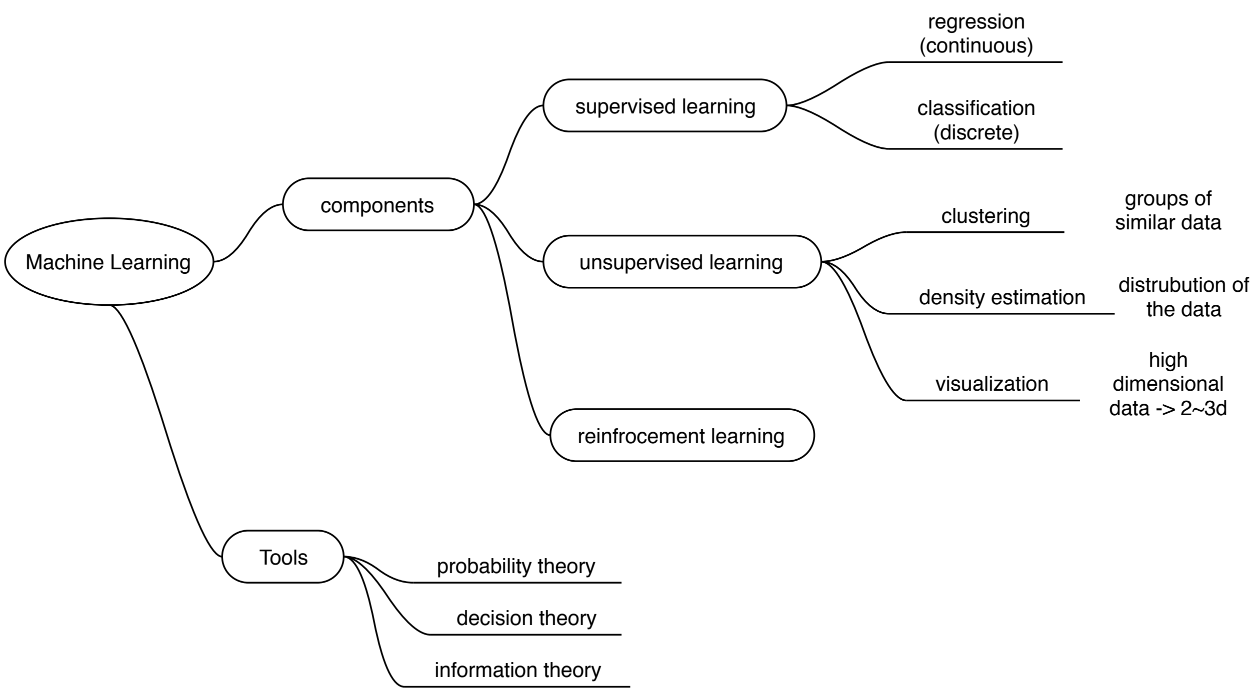


图 1: Outline of Machine Learning

1.2 A regression example: Curve-Fitting

Problem: 给定规模为 N 的数据集 dataset $= \{(x_i, t_i)|i = 1, 2, \cdots, N\}$，希望通过数据集找到 $f : x \mapsto t$，对未知的 x 对应的 t 值进行预测

最朴素的想法就是进行**多项式拟合 (Polynomial Fitting)**, 定义多项式1：

$$f(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M \tag{1}$$

多项式是关于 $\mathbf{w}$ 的线性函数

? 如何去寻找系数 $\mathbf{w}$

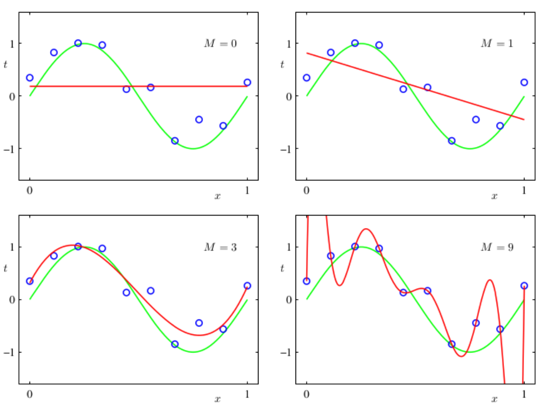
我们定义一个误差函数如2所示：

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

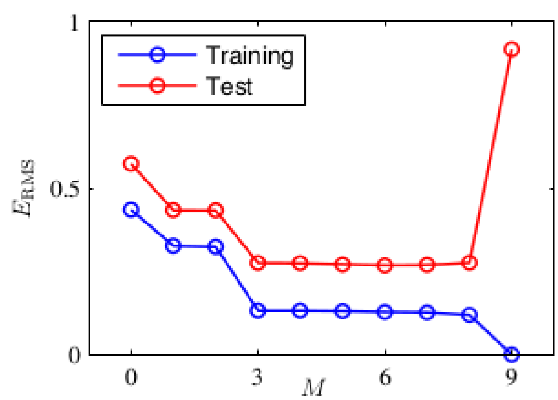
这是一个最优化问题，因为误差函数是关于系数的二次型，导数必然是线性的，对于一个确定的阶数 M，一定存在一个最优解 $\mathbf{w}_*$。

需要注意 $\mathbf{w}_*$ 和 M 是对应的，不同的阶数 M 对应着不同的模型复杂度，也对应着不同的 $\mathbf{w}_*$。在不同的模型（curve-fitting 中的 M 的选择）之中作出抉择即**模型比较或模型选择 (Model Comparison/Model Selection)** 问题是机器学习中的一类重要问题

欠拟合 (Under-fitting) 与过拟合 (Over-fitting) 问题：在 curve-fitting 问题中，当选择的模型阶数过低，模型无法反映出数据真值的变化; 当选择的模型的阶数过高时，尽管模型对数据集的误差非常小，但是多项式的系数会非常之大，造成严重的震荡。



(a) 不同复杂度（不同阶数）的模型的拟合效果



(b) $E_{\text{RMS}}(\mathbf{w}_*)$ M

图 2: 不同的模型之间的比较

这里需要一个衡量拟合效果的标准，我们采用 N 个数据得到了模型 $f(x, \mathbf{w}_*)$，当需要检测模型的效果时，可以测量得到更大的数据集，计规模为 N_{test}，则测试集计算得到的误差函数为:$E(\mathbf{w}_*)$，为了比较不同阶数多项式的拟合效果，可以用**均方根值 (Root Mean Square, RMS)** 来进行比较：

$$E_{\text{RMS}}(\mathbf{w}_*) = \sqrt{2E(\mathbf{w}_*)/N_{test}} \tag{3}$$

如图2所示，当 $N = 10$ 时，M 在超过 9 后（$\mathbf{w}_* \in \mathbb{R}^{10}$ ），拟合效果变差。当我们增加数据集的规模 N，$M = 9$ 的拟合效果会慢慢变好。

对 $E(\mathbf{w})$ 的最优化问题本质即是**最小二乘法（Least Square Root）**，而最小二乘法是**极大似然 (Maximum Likelyhood)** 的一个特例。可以看到最小二乘法模型复杂度受到数据集大小的限制，而之后会介绍的**贝叶斯（Bayesian）**则可以避免过拟合的发生，它对数据集的规模是自适应的。

这里先暂时不介绍贝叶斯的内容，而尝试将最小二乘法方法进行优化，通过观察发现 $M > N$ 造成的过拟合震荡主要源于异常大的系数，这里考虑在 $E(\mathbf{w})$ 中加入惩罚项 (Penalty Term)，再对新的目标函数进行最优化，从而来避开 $\|\mathbf{w}_*\|$ 过大的求解域, 这一过程称之为**正规化（Regularization）**

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}[f(x_i, \mathbf{w}) - t_i]^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}^1, \lambda \geq 0 \tag{4}$$

这种正规划方法在不同的领域有不同的名称，在统计学领域，因为它能将系数缩小，称之为 **Shrinkage**；这个特殊的正规项也称 **ridge regression**；在神经网络中，这种方法叫做**权重衰减 (Weight Decay)**

1.3 概率论

概率论是我们用来度量不确定度 (Uncertainty) 的理论工具，其中包括不确定度的量化 (quantification) 以及操作（Manipulation）。概率论和之后介绍的决策论将共同形成一套最优预测的方案（Optimal Prediction）。

1.3.1 基础概念及重要性质

<u>**Def**</u> **Probability** of an event is the fraction of times that event occurs out of total number of trials, in the limit the total number of trials goes to infinity.

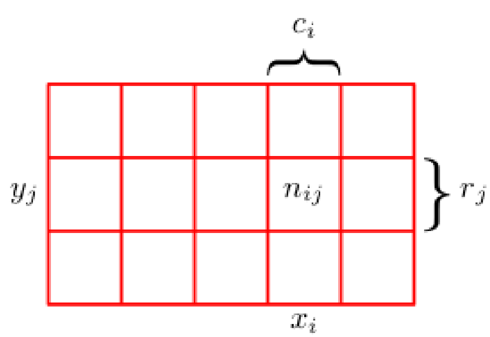


图 3: 两个随机变量 X 与 Y

<u>常见的几个概率</u>

概率 $p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$

边缘概率 (Marginal Probability) $p(X = x_i) = p(X = x_i, Y) = \frac{c_i}{N}$

$p(x_i) = \int_{-\infty}^{+\infty} p(x_i, y)\,\mathrm{d}y$

[1] ω_0^2 常常省略，具体 5.5.1 会讲到

条件概率 (Conditional Probability) $p(X=x_i|Y=y_j)=\frac{n_{ij}}{r_j}=\frac{n_{ij}/N}{r_j/N}=\frac{p(X=x_i,Y=y_j)}{p(Y=y_j)}$

probability of A given B:

$$p(A|B)=\frac{p(AB)}{p(B)}$$

<u>两条重要规则</u>

加法规则 (sum rule): $p(X)=\sum_Y p(X,Y)$

乘法规则 (product rule): $p(X)=p(XY)/p(Y|X)$

1.3.2 贝叶斯（Bayesian）

基于先验信息的概率计算问题

$$\begin{aligned} p(X|Y) &= \frac{p(XY)}{p(Y)} \\ &= \frac{p(Y|X)p(X)}{p(Y)} \end{aligned}$$

在 Y 发生的条件下，X 的概率。这里 $p(X)$ 称之为先验概率（Prior Probability），它是基于历史观测和经验得到的；$p(X|Y)$ 称之为后验概率 (Posterior Probability)，该概率是对 Y 的观测完成之后才得到的，可以通过贝叶斯定理得到。

1.3.3 概率密度