机器学习 Notes 1

本文给出一些纲领性的内容。

人工智能简介

三大方法：搜索、推理、学习。前两者：基于规则的方法。

- 基于规则的方法：直接编程实现，借鉴人类启发式学习的思想。
- 基于数据的方法：两个阶段。
    - 专家系统：专家或者数据科学家基于数据创造用于预测或者决策的准则。需要人的参与。
    - 机器学习：直接基于数据进行预测或者决策。和数据科学相辅相成。自动化。

人工智能探讨对机器进行设计的方法论，使其可以去完成基于智能的任务。

数据科学

比一般的自然科学更加广泛。

目标：发现数据的基本原理，利用原理进行服务。（对比物理学：发现世界的基本原理）

解决方法：从观测结果中构建数据模型（对比物理学：从观测结果构建世界模型）

从数学上来讲，数据科学想要解决的问题大致就是**从现有的数据中找出一个联合概率分布，从而得到条件概率分布。**

数据本身没有价值，有价值的是数据服务！

机器学习简介

定义

> 学习是系统通过经验提升性能的过程。

什么是机器学习？有一个非常押韵的英文定义：

> A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if **its performance at tasks in T, as measured by P, improves with experience E**.
>
> By Tom Mitchell

还有一个更加古典的定义：

> The field of study that gives computers the ability to learn **without being explicitly programmed**.
>
> By Arthur Samuel

要点有 4 个：在某些任务 T 上，通过经验 E，提升性能 P，以**非显式编程**的形式。

编程 vs 机器学习

编程：人直接写程序，然后程序接受输入并产生输出。

机器学习：人只写学习算法，然后利用数据进行训练得到模型，模型接受输入并产生输出。

机器学习的优势

应用情形：

- 模型基于大量数据。如网络搜索。
- 输出具有个性化。如推荐系统。
- 人类也解释不清楚的。如语音、人脸识别。
- 人类根本不知道的。如火星上导航（Try-and-error）。

机器学习的任务类型

分为两大类型：**预测**和**决策**。

- 预测（prediction）。比较简单，起辅助作用。
  - 给定特征数据，预测目标概率分布（监督学习）
  - 生成数据实例（无监督学习）
- 决策（decision-making）。比较重要，起主导作用。
  - 在动态环境中采取行动（强化学习）。一般包含下面几个方面：
    - 转变到新的状态。
    - 获得即时的奖励。
    - 随时间推移最大化积累奖励。

机器学习的应用

预测类：网页搜索、人脸识别、推荐系统、在线广告、信息提取、医疗图像分析、金融预测、社交分析等。

决策类：交互式内容推荐、机器人控制、自动驾驶、游戏 AI、多智能体协作等。

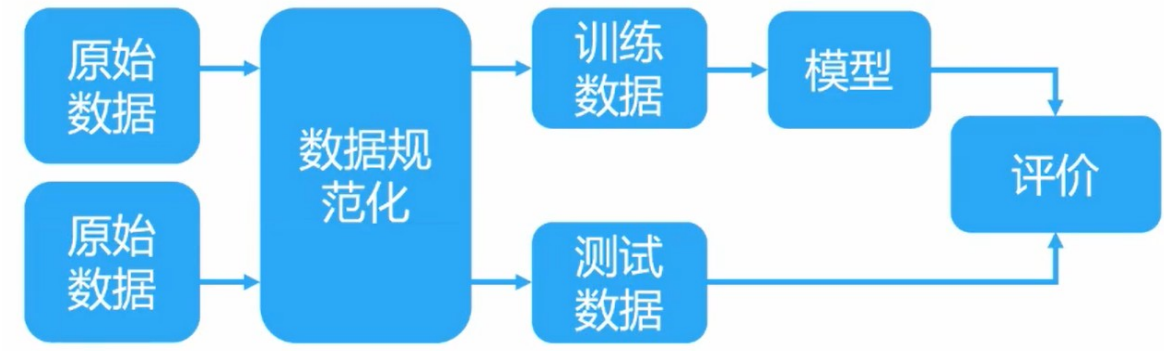
机器学习的学习类型

分为**监督学习**、**无监督学习**和**强化学习**。

- 监督学习：给定数据和标签，对输入预测所需的输出。
- 无监督学习：分析和利用隐式数据模式/结构，建模数据的联合分布。
- 强化学习：学习在动态环境中动作执行的决策，并获得尽可能多的奖励值。

机器学习过程

可以用下面的这张图来表示。（图片来源：CS420 课程讲义）



为了保证学习的有效性，一个基本的假设是训练数据和测试数据中存在相同的模式。

监督学习

定义：给定带标签的训练数据集

$$D = \{(x_i, y_i)\}_{i=1,2,\cdots,N}$$

其中 x_i 为**特征数据（features）**，y_i 为对应的**标签（labels）**，需要让机器学习一个映射 f_θ 使得 $y_i \approx f_\theta(x_i)$。

这样的函数一般是一个参数化的函数空间中的某个元素。即：我们有一个参数 θ，不同的 θ 得到不同的 f_θ。这些 $\{f_\theta(\cdot)\}$ 构成**假设空间（Hypothesis space，一般记作 $\mathcal{H}$）**。学习的过程就是不断更新 θ，找到比较好的一个 θ 的过程。

这个时候就出现了两个问题。

问题：学习目标?

确实要让学到的假设接近数据，但学习的具体目标是什么?

我们一般采用量化的方法：使用一个**损失函数（Loss function）**衡量标签和预测结果之间的误差。即求解优化问题

$$\min_\theta \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(y_i, f_\theta(x_i))$$

定性地，应当满足 $y_i, f_\theta(x_i)$ 相差越大，函数值越大。一般采用均方误差函数：

$$\mathcal{L}(y_i, f_\theta(x_i)) = \frac{1}{2}(y_i - f_\theta(x_i))^2$$

这个函数有一些好处，例如：可以容忍小的误差，这容许了噪声的存在，也提升了模型的泛化性能。

问题：参数更新?

采用了量化的方法之后就可以使用一些数学上的手段求解优化问题。常见的方法有梯度下降等。

模型选择

欠拟合、过拟合

当统计模型或者机器学习算法无法捕捉数据的基础变化趋势的时候，就会出现**欠拟合（underfitting）**。

当统计模型把随机误差与噪声也考虑进去，而不仅仅是考虑数据的基础关联时，就会出现**过拟合（overfitting）**。

正则化

添加参数的惩罚项，防止模型对数据过拟合。一般的形式为

$$\min_\theta \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(y_i, f_\theta(x_i)) + \lambda\Omega(\theta)$$

后面那个新的项就是**正则项（regularization term）**。

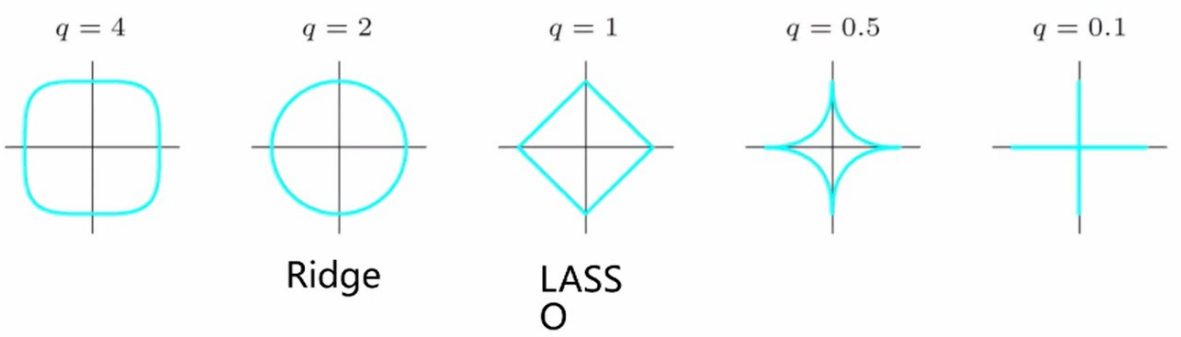
有两个经典的正则化方法：

1. **L2 正则化（岭回归，Ridge）**，即取 $\Omega(\theta) = \|\theta\|_2^2$。

2. **L1 正则化（套索，LASSO）**，即取 $\Omega(\theta) = \|\theta\|_1$（即一维范数）。

（这里 LASSO 的全称是 Least Absolute Shrinkage and Selectionator Operator，直译为最小绝对收敛和选择算子）

可以推广到更加普遍的情况，即使用 $\Omega(\theta) = \|\theta\|_q^q = \sum |\theta_i|^q$（$q$ 维范数的 q 次方）作为正则项。下面展示了 $\sum |\theta_i|^q$ 一定时，θ 的分布。（图片来源：CS420 课程讲义）



当取 $q \leq 1$ 的时候，模型进行稀疏性学习，即利用的特征会比较少。这一点可以和凸优化里面的"最小维度表示"联系起来。

大多数情况下，都是用 L1 或者 L2 正则化。

哲学原则：奥卡姆剃刀

> 如无必要，勿增实体。

有多个**有效**的假设模型时，应当选择假设条件最少的建模方法。

超参数

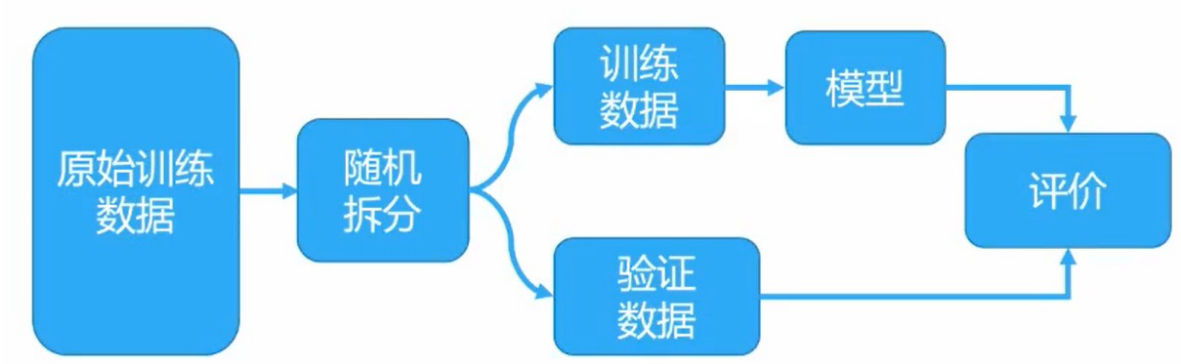
在上述框架下，一个机器学习的解决方案的模型包括参数 θ 和**超参数（hyperparameter）** λ。

超参数定义模型的更高层次的概念，如复杂性或者学习能力（如学习率）。它们在标准的模型训练过程中**无法直接从数据中学习**，需要依靠人为选择。

可以通过不同的参数设置，训练不同的模型以及选择最好的测试结果来进行超参数的选择。

交叉验证

下图展示了交叉验证的过程。（图片来源：CS420 课程讲义）



常见的一种交叉验证方法为 K-折交叉验证（K-fold cross validation）。对于给定的超参数，将数据集随机拆成 K 份。每次用其中 $K-1$ 份训练，一份测试、评估。重复 K 次。最终对得到的分数取平均，作为模型的性能。

注意在验证的时候我们不会用到测试数据。

泛化能力

泛化能力（generalization ability，GA）指的是模型对未观测数据的预测能力。

泛化误差

泛化能力可以通过**泛化误差（generalization error，GE）**来评估，定义为：

$$R(f)=\mathrm{E}(\mathcal{L}(Y,f(X)))=\int_{X\times Y}\mathcal{L}(y,f(x))p(x,y)\mathrm{d}x\mathrm{d}y$$

其中 $p(x,y)$ 是潜在的（可能是未知的）联合数据分布。

但显然我们无法穷尽概率空间，于是只能在训练数据上做一个经验性质的估计，即

$$\hat{R}(f)=\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(y_i,f(x_i))$$

（顺便说一下，一般带帽子的都表示基于经验的平均数据。这一点后面可能会用到。）

泛化误差约束定理

我们希望知道 $\hat{R}(f)$ 和 $R(f)$ 之间相差多少。

设有限假设集 $\mathcal{F}=\{f_1,f_2,\cdots,f_d\}$，且训练数据有 N 项。$\forall f\in\mathcal{F}$，有不小于 $1-\delta$ 的概率，下面的式子成立：

$$R(f)\leq\hat{R}(f)+\epsilon(d,N,\delta)$$

其中

$$\epsilon(d,N,\delta)=\sqrt{\frac{1}{2N}\left(\log d+\log\frac{1}{\delta}\right)}$$

我们可以通过 Hoeffding 不等式来证明这个结论。该不等式的内容是：对于独立同分布的一系列随机变量 $X_1,X_2,\cdots,X_n$，且 $X_i\in[a,b]$，设 $Z=\frac{1}{n}\sum_{i=1}^{n}X_i$。那么

$$\forall t\geq 0,\Pr(|Z-\mathrm{E}Z|\geq t)\leq\exp\left(\frac{-2nt^2}{(b-a)^2}\right)$$

方便起见，我们设 $R(f)\in[0,1]$。这可以通过归一化得到。由 Hoeffding 不等式，将 Z 替换为 $\hat{R}(f)$，将 $\mathrm{E}Z$ 替换为 $R(f)$，则有

$$\forall f,\Pr(R(f)-\hat{R}(f)\geq\epsilon)\leq\exp(-2N\epsilon^2)$$

由于数据有限，从而

$$\begin{aligned}\Pr(\exists f\in\mathcal{F},R(f)-\hat{R}(f)\geq\epsilon)&=\Pr(\cup_{f\in\mathcal{F}}\left\{R(f)-\hat{R}(f)\geq\epsilon\right\})\\&\leq\sum_{f\in\mathcal{F}}\Pr(R(f)-\hat{R}(f)\geq\epsilon)\\&\leq d\exp(-2N\epsilon^2)\end{aligned}$$

取上面那个事件的相反事件得到

$$\Pr(\forall f\in\mathcal{F},R(f)-\hat{R}(f)<\epsilon)\geq 1-d\exp(-2N\epsilon^2)$$

再取 $\delta=d\exp(-2N\epsilon^2)$，就得到了上面的定理。

判别模型和生成模型

判别模型

判别模型，对可观测变量和未知变量的关联性进行建模。也被称为条件模型。判别的结果分两种：

- 确定性判别：直接将函数值作为预测。即 $y = f_\theta(x)$。
- 随即判别：计算条件概率，即 $p_\theta(y|x)$。

它是非常直接的。可以用于直接建模预测标签和已知特征的关联。我们可以**只考虑局部的目标**，只使用特定的特征产生模型，而不需要考虑全局，不需要建模出整个联合分布。它实际上可以产生更高的预测性能。

它常用于监督学习中，如线性回归，逻辑回归，k近邻，支持向量机，（多层）感知机，决策树，随机森林等。

生成模型

生成模型：对数据的联合概率分布进行建模。给定的数据是一些隐参数或者隐变量 $p_\theta(x, y)$。可以进行条件推断，即根据联合概率和边缘概率计算条件概率。如：

$$p_\theta(y|x) = \frac{p_\theta(x,y)}{p_\theta(x)} = \frac{p_\theta(x,y)}{\sum_{y'} p_\theta(x,y')}$$

可以认为生成模型对于整个数据集有比较全面的认知。它是在探寻数据的分布，比较接近数据科学的本质。它往往需要领域的专业知识以找到有用的隐变量帮助建模。

常见的生成模型包括朴素贝叶斯，隐马尔可夫模型，混合高斯，马尔可夫随机场，隐狄利克雷分布 (LDA) 等。