机器学习里有哪些回归模型

机器学习任务可以大体上分为两大类：一类是预测型的，基于观察数据对类别或其它变量进行预测，主要包括分类任务和回归任务；另一类是描述型的，通过估计数据的概率分布来描述数据的自有特性，主要包括聚类任务和流形学习任务。

所谓回归（Regression），是指给定一个变量集 A，基于 A 中变量的取值来预测另一个变量 b 的取值，其中 A 中变量称为自变量，b 称为因变量。虽然看起来比较复杂，其实就是通过一些观测值来预测未知变量，如通过照片预测年龄，通过地理位置和地形预测降水量，通过昨天的股市情况预测今天的开盘价等。

1.“回归”的由来

“回归”一词来源于 19 纪世英国遗传学家弗朗西斯-高尔顿（Frramcia Galton），他在研究父母和子女身高之间的关系时发现，父母身高越高子女也倾向于更高，这个并不难理解；有趣的是，他还发现身高越高的父母，自己的子女倾向于比自己矮一些，而身高较低的父母，其子女倾向于比自己高一些。也就是说，后代的身高倾向于“回归平均身高”。“回归均值”是生物界的基本规律，例如很多天才人物的孩子只是平庸之辈，很多高智商的父母发现子女的脑子不灵，都是这一规律的体现。

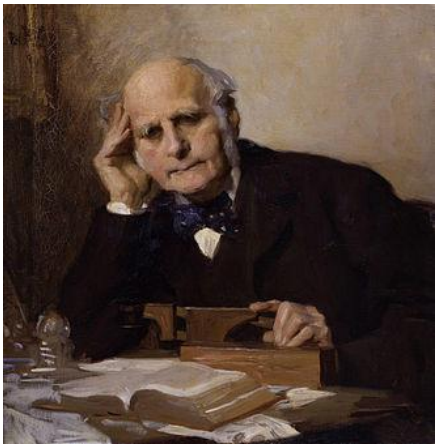

图 1：弗朗西斯-高尔顿（Frramcia Galton），英国博学家、人类学家、统计学家。

回归分析可追溯到 19 世纪初法国数学家勒让德（Legendre）和德国数学家高斯（Gauss）提出的最小二乘法。卡尔-皮尔逊（Karl-Pearson）将回归分析解释成联合分布为高斯的最大似然估计，罗纳德-费舍尔（Ronald Fisher）进一步将回归分析解释成条件概率为高斯的最大似然估计，这也是机器学习所持有的观点。

2. 线性回归

线性回归是最简单的回归模型。以预测子女的身高为例，设父亲向高为 F, 母亲身高为 M，子女的性别为 G，则子女身高的回归模型为：

$$C = \alpha F + \beta M + \gamma G + \epsilon$$

其中ϵ为一个均值为 0 的高斯分布，α、β、γ为模型参数。如果我们可以收集到若干父母-子女的身高值作为训练样本，即可得到一组优化的参数α、β、γ，这些参数将确定一个优化的回归模型。有了这一回归模型，即可对任意一对父母的子女进行身高预测了。

线性模型虽然简单，但因为变量之间的依赖关系清晰明确，可解释性强，在很多场景下有广泛应用。如在金融信号分析中，线性回归模型被广泛用于趋势预测、因子敏感性分析、资产定价等。

3. 非线性回归

将变量间的线性依赖关系推广到非线性形式即得到非线性回归模型，表示如下：

$$C = f(F, M, G, \theta) + \epsilon$$

其中 f 为任意非线性映射函数，θ为该函数的参数，ϵ依然为一个高斯分布。如果采用神经网络来实现 f，则θ对应网络的连接权重。

进一步，也可以允许ϵ不再是一个高斯分布（如高斯混合分布），此时模型的预测有可能不再向均值回归。另外，模型训练时可以引入更复杂的目标函数或更复杂的约束条件，此时模型不再是传统的回归模型，而是一个通用的预测模型。