吴恩达人工智能课程学习体会

林晓

2017-08-11

世界上有各种事件，记为$x = (x_1, x_2, \cdots, x_n)^T$。这些事件可能会以某种概率导致事件$y$的出现。人工智能研究如何把$x$与$y$的因果关系联系起来的一种量化模型(即一个很长的数学公式)。

$$y = h_\theta(x)$$

其中θ是参数。

人工智能主要研究四个方面的问题：(1)根据具体的行业(如无人驾驶，人脸识别等等)对$x_1, x_2, \cdots, x_n$进行量化定义。(2)提出函数的$h_\theta(x)$的具体形式，专业上叫该函数为学习算法。(3)通过大数据$(x^{(i)}, y^{(i)}), (i = 1, 2, \cdots, m)$ 的回归计算参数θ，专业上叫这个过程为机器学习或者训练。(4)将量化模型的输出y作为整个系统其他构件的输入。由于(1)和(4)与具体的专业密切相关，吴教授的教程主要讲授(2)和(3)方面的知识。一些学者把现阶段称为人工智能的中级阶段，认为(1)和(2)的大部分工作是由人工来完成的。在未来的高级阶段，(1)和(2)主要也会由计算机来完成。

概率统计学是建立一个学习算法$h_\theta(x)$的基础。吴教授在课程中证明了一般的学习算法$y = h_\theta(x)$可以表示为给定x和参数θ的条件下y的数学期望

$$h_\theta(x) = E[y|x;\theta]$$

因此，整个建模过程以及参数θ的求解都与条件概率密度函数$p(y|x;\theta)$密切相关。下面是一些经典例子。

回归模型: 学习算法中目标y取连续函数值

$$y = h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x$$

概率密度函数可取为

$$p(y|x;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(y - \theta^T x)^2}{2\sigma^2} \right)$$

参数θ通过似然函数(Likelihood)的最大化来确定

$$L(\theta) = \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(y^{(j)} - \theta^T x^{(j)})^2}{2\sigma^2} \right)$$

取对数并对θ求导求极值点，得到梯度下降法的迭代求解公式

$$\theta_j := \theta_j + \alpha \sum_{i=1}^{m} \left(y^{(i)} - \theta^T x^{(i)} \right) x_j^{(i)} \qquad (j = 1, 2, \cdots, m)$$

α为步长因子。这个方法可以用最小二乘方法得到，所以也叫最小二乘学习方法。还有一种算法，目标y也是连续的，但取值限于区间$[0, 1]$之内,

$$y = h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

分类模型: 学习算法中目标y取离散函数值。最简单的情况y只取0和1两个值，这是著名的感知模型

$$y = h_\theta(x) = g(\theta^T x) = \begin{cases} 1, & \theta^T x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

概率密度函数可取Bernoulli概率分布

$$p(y|x;\theta) = ((h_\theta(x))^y \ (1 - (h_\theta(x))^{1-y}$$

同样可求得似然函数和参数θ的迭代求解公式

$$\ell(\theta) = \log L(\theta) = \sum_{j=1}^{m} y^{(j)} \log h(x^{(j)}) + (1 - y^{(j)}) \log(1 - h(x^{(j)}))$$

$$\theta_j := \theta_j + \alpha (y^{(i)} - h_\theta(x^{(i)})) x_j^{(i)} \qquad (j = 1, 2, \cdots, m)$$

比较复杂一点的分类模型可以允许y取k个值：$y \in \{1, 2, \cdots, k\}$, 对应的多项分布概率为$\phi_1, \phi_2, \cdots, \phi_k$。通过演算可以得到相应的学习函数$h_\theta(x) = E[y|x;\theta]$。分类模型有着广泛的应用，例如判别一个动物是一头大象还是一条狗，判别一种肿瘤是良性的还是恶性的，等等。

以上建立模型使用的方法是一种直接的方法。还有一种间接的建立模型方法，其思路是应用贝叶斯概率公式。如果用y表示一个动物是一条狗(0)或者一头大象(1)，则$p(x|y = 0)$将模拟狗的特征分布，$p(x|y = 1)$将模拟大象的特征分布。在模拟

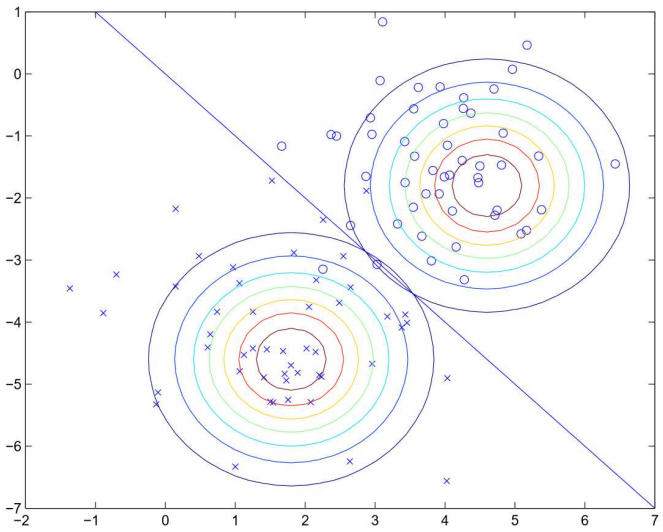

Figure 1: 引自吴恩达的讲义。

了$p(y)$（称为先验概率）和$p(x|y)$之后，我们的学习算法将使用贝叶斯公式得到，其结果是在给定x条件下关于y的后验概率，

$$p(y|x;\theta) = \frac{p(x|y;\theta)\ p(y;\theta)}{p(x;\theta)}$$

上式的分母可以从概率教科书中的全概率公式得到(省去θ)

$$p(x) = p(x|y=0)\ p(y=0) + p(x|y=1)\ p(y=1)$$

因此也可以用我们已经获得的$p(x|y)$和$p(y)$表示出来。但是，如果仅仅限于模拟$p(y|x)$以便得到一个学习算法，我们不需要计算分母，因为

$$\begin{aligned} \arg\max_y p(y|x) &= \arg\max_y \frac{p(x|y)\ p(y)}{p(x)} \\ &= \arg\max_y p(x|y)\ p(y) \end{aligned}$$

以上使用贝叶斯方法分类，就是按物体的属性求出一个分界面，将两个物体分开，如图1。从这个思路出发，还可以引申出其他的方法。例如从几何上来看，求一个分界面，使得不同性质的点到分界面的距离最短，等等。

还有一种分类方法，与无监督学习模型有关。一个人工智能模型，如果是给出了一个$y = h_\theta(x)$的学习算法，输入数据$(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$进行训练，这叫做监督

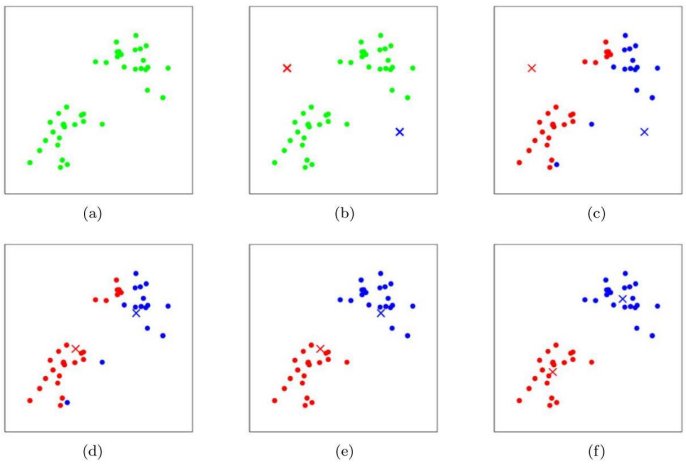

Figure 2: 引自吴恩达的讲义。

学习模型，就是对给定的x,引导计算机输出一个什么样的y。如果我们研究的对象是一堆只有x的数据，$x_1, x_2, \cdots, x_m$, 要对他们进行分类。这种模型叫无监督学习模型。一个应用叫K-期望值学习方法。例如，如果认为这堆数据可分为K类，每类都服从一个正态分布。这样先假定K个中心。按照数据点到中心的距离对数据分为K类，分类后又求各类的中心点（期望值）。然后重复以上过程，按照数据点到中心的距离再次对数据分为K类，这样一直迭代到收敛为止，如图2。

试想一下如何应用这种无监督学习的训练方法来发行余额宝理财产品。余额宝发行理财产品，期限是不规定的，客户可能一天就把钱取走，也可能很多年都不动，并且不断加入更多的资金。于是余额宝可以对大量的客户进行分类，计算出各类客户存钱的久期来，然后按分析结果把资金投到收益高期限长的资产产品上去，得到比较高的收益来报答客户。有一段时间余额宝的理财产品收益很高。当记者问他们怎样投资时，回答说是投了银行的协议存款，估计就是这么操作的。

开发人工智能的思路，就是建立一个模型，让机器通过大量的数据进行学习，确定模型中的参数，最后就能用模型来预测未来。显然模型选得好不好，对产品的质量影响很大。于是，需要一种检验模型的方法。吴教授介绍了一种方法，先选定几个模型，例如一个多项式回归，按多项式的幂次数就可以得到不同的模型。然后将训练数据集三七开：70%用来学习建模。建好之后，用剩下的30%数据来检验计算，看看输出

结果与期望值的偏差大不大，方差大不大，从而判别那个模型比较好。叙述这个方法的定理，据说是人工智能最重要的定理，为以后使用计算机来建立模型和选择模型打下了基础。

最后交流一下个人的学习体会。人工智能也和当年华尔街投资银行开发衍生产品的量化模型一样，数学好会编程的人是比较好切入的。但是，能否真正做出一番事业来，还必须非常熟悉某个具体行业。例如，如果不懂汽车行业，想开发无人驾驶方面的人工智能软件，将是比较困难的。