浅析机器学习原理

郑梓豪

“如果说统计学习、数据挖掘、机器学习有什么区别，然后加上个模式识别(模式分类)的话，就是基本上这四者的关系实在太亲近了。统计学习应该学院味道比较重，背后有强大的概率论、随机过程、统计学做理论支撑，模型结构比较漂亮，体现数学之美，数据挖掘商业色彩最浓，因为大企业都需要数据挖掘，后跟数据库系统紧密相关，且为商业决策服务，从各个学科拿来各种方法改着用，算是四者最强调实用目的。机器学习和模式识别在工程应用广泛，有神经网络、图形处理等等其它。另外这几个都是多学科交叉的，基本上四个名字有时候仅仅充当不同领域的人对相同事物的不同称谓。[1]”

笔者试图以自身四周的学习体验，介绍机器学习的同时，向读者展示机器学习的核心思想。最后以“数字图像的神经网络识别”为例，试图回答“为什么机器会学习”这个问题。

1 什么是机器学习?

> *“Field of study that gives computers the ability to learn without being explicitly programme”*
>
> *Arthur Samuel*

这是让笔者记忆最深的定义，这也是最能体现“机器学习”的特点的：不必“清楚地”告诉机器如何去做。这也是最吸引人的地方，所谓人工智能本该如此。事实上，冷冰冰的机器真的有自主学习的能力吗？本文将探讨这个问题。

> *“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”*
>
> *Tom M. Mitchell*

第二段关于机器学习的定义便有较强的学术味道，这里有几个关键词：experience E, tasks T, performance P. 在此先不必急着解释这三个方面，待文章进行下去便会有答案。

2 机器为什么会学习?

可以想象，机器（计算机）在面对数据时的窘态，因为它只不过将这堆数据视作内存中的01序列而已。事实上，脱离了编程语言的机器一无是处，软体是硬体能成功运行的保证。那么，机器的“学习”也一定是在机器学习算法的指导下进行的。

[1] 上述文字引自一篇博文(见[2])

或许有人会说，这算什么学习？不过是人们在教机器“如何学习”罢了。要回答这个疑惑，只管想想我们自己就可以了：想象一下，若有一个生活在一个完全封闭[2]的环境里的婴儿会怎样？答案很明显，婴儿什么也不懂。其实，人类的学习过程一直是个谜，关于大脑学习的机理还没有太多认识，所以关于“学习”的定义也是不清晰的。所以不妨把它视作一个广义的概念。

有了对“学习”概念的把握，便可以开始介绍算法了。

2.1 简单的机器学习算法

通过对Linear Regression, Logistic Regression的介绍，揭示其内在通性。对于线性回归的机器学习解释，让读者更易切入。

2.1.1 线性回归(Linear Regression)

事实上，这个简单的算法就是“一元线性回归(Linear Regression)”。回归分析早已在高中便接触过，大学时的数理统计课上给出了进一步的介绍和应用。

先回顾“线性回归”的主要建模思路[3]：

1. 对于一组具有线性特征的数据集[4](x_i, y_i)，用一元线性函数$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$对其进行拟合。
2. 其中$\hat{\beta}_0, \hat{\beta}_1$的计算采用最小二乘法：

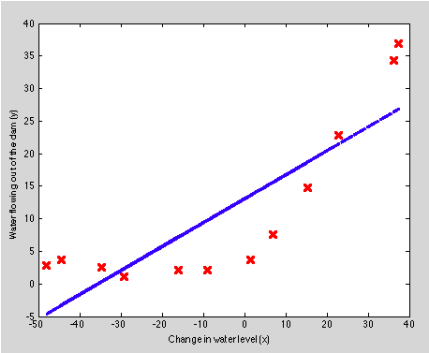
$$J(\hat{\beta}_0, \hat{\beta}_1) = \min \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

   要求选取适当的$\hat{\beta}_0, \hat{\beta}_1$使$J(\hat{\beta}_0, \hat{\beta}_1)$最小。不妨称$J(\hat{\beta}_0, \hat{\beta}_1)$为Cost Function

3. Cost Function为二元函数，其极值问题可通过其偏导数得到：

$$\begin{cases} \frac{\partial J}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial J}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{cases}$$

   求解回归系数，便可得回归方程。



[2]比如说一个只有墙的房子里
[3]这里的思路可能会与教材有所不同，但保持了其核心思想
[4]事实上对于数据集呈现出来的特性并不一定要求为线性，非线性亦可

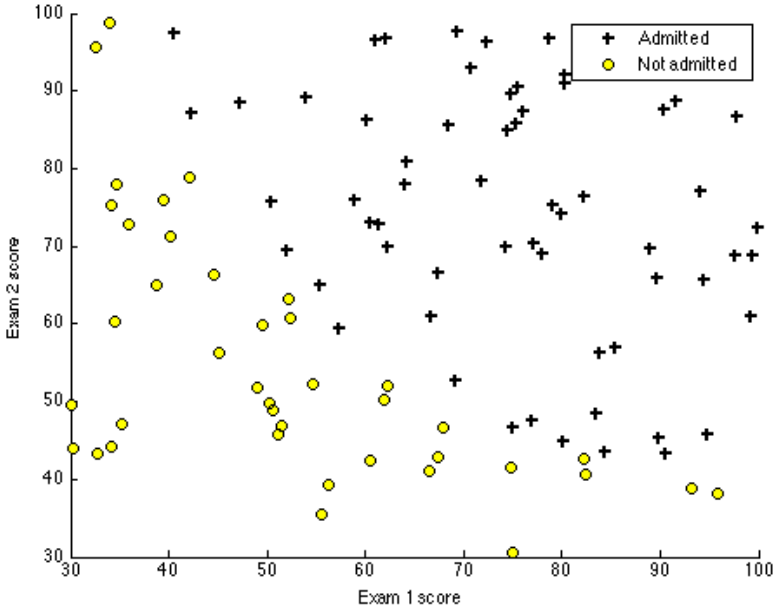
事实上，回归方程提供了一种拟合原数据的方案。留意到其回归系数求法的$(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$，这个简单的式子强调了对近似值$(\hat{\beta}_0 + \hat{\beta}_1 x_i)$与真实值$(y_i)$的要求，最接近！所以求得的回归方程对原数据的拟合是误差最小的。

再来着重理解Cost Function的命名意图，为什么称其为Cost？其实，拟合效果的好坏，可以视作此模型的“成本”或者“损耗”，成本、损耗越低便拟合效果越好。

这类问题，用机器学习的观点来看，是应用十分广泛的“预测”问题。初看起来，这似乎与机器的“学习”没有太多关系。不过不必担忧，下面将介绍机器学习中另一类重要问题——分类(Classification)问题。

2.1.2 Logistic Regression

给定数据集（如图2），其分类问题可用另一种回归分析法进行。



假设你是一位懂机器学习的老师，你希望通过学生两次平时测试的成绩来推断其期末的表现。现在你手头上有历年来的数据资料，上面登记了学生的平时成绩：横纵分别为第一、二次测验的结果，点的颜色代表了期末是否通过(Admitted, Not admitted)。

那么你便可以通过这个历史资料，运用以模式分类为目标的机器学习算法，建立分类模型。模型建立后，你便可以通过对目前学生的平时表现，去推测其期末表现。

下面介绍Logistic Regression算法:

1. 对已知数据集$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$，其中 $y_i = \{0, 1\}$ 这里0,1的表示是否为某一特定类型（如Admitted为1）
2. 假设有一个可以将两种类型(Admitted,Not admitted)区分开的函数，不妨称之为hypothesis:

$$h_\theta(x) = g(\theta^T x)$$

其中函数$g(z)$称为sigmoid function

$$g(z) = \frac{1}{1+e^{-z}}$$

对于这个函数的功能，我们可以让它输出0,1作为标识是否属于某一特定类别的特征函数。sigmoid function的形式也是由这种想法得到的：

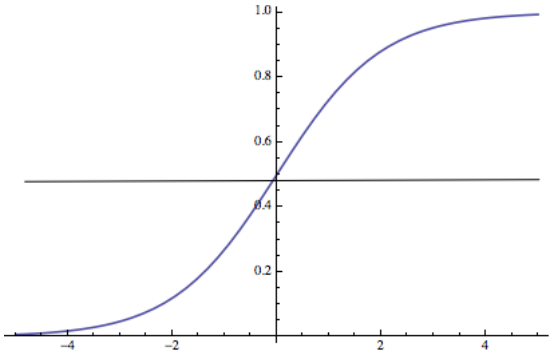


Figure 1: $h_\theta(x) = \frac{1}{1+e^{-(\theta^T x)}}$

通过图像恰好发现，这是一个函数值在[0,1]上的连续函数。进一步地，这个函数的优美之处在于，它被 $h = 0.5$ 平分！这允许我们把 $h > 0.5$ 标识为1，即属于某种特定类型(如Admitted)，这里面有概率的味道——比如有0.7的概率说此数据为Admitted类型。

3. 可以猜想到，黄色点与黑色点之间存在一条分界线——这条分界线正是我们想要的。不妨设$h_\theta(x) = \theta_0 + \theta_1 x$。那么如何从历史数据中学习到这种信息呢？同样地有Cost Function：

$$\begin{cases} J(\theta) = -log(h_\theta(x_i)) & y_i = 1 \\ J(\theta) = 1 - log(1 - h_\theta(x_i)) & y_i = 0 \end{cases}$$

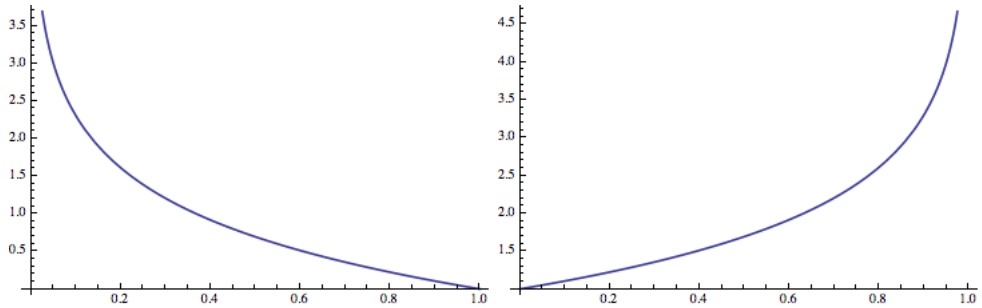


Figure 2: $J(\theta)$

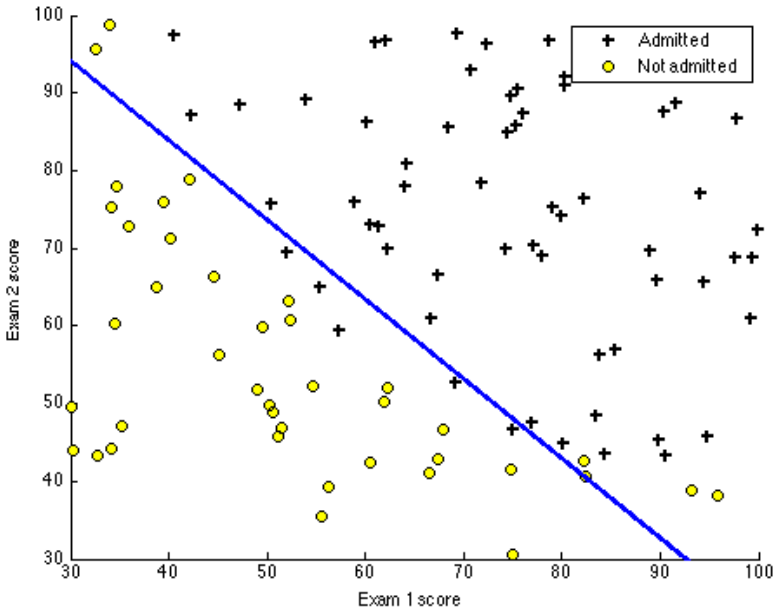
可以看到当$y_i = 1$时，若$h_\theta(x_i) = 0$，则$J(\theta) = +\infty$也就意味着发生“归类错误”的“损耗”十分巨大。当我们进一步对参数θ进行选择时，机器自然就会“主动”避免此情况发生。也就是说，我们得到的θ仍然是最准确的！

将分段函数合在一起：

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[-y_i log(h_\theta(x_i)) - (1 - y_i log(1 - h_\theta(x_i)))]$$

4. 最优化参数选择。这里的目标是通过$\theta = \theta_0, \theta_1$的选择使$J(\theta)$最小。方法类似，不再过多解释。读者可参见文献[1]得到关于参数计算的信息。

分类函数[5]便可以得到



现在，读者应当已经发现了机器学习中最重要的两个方面。

2.2 机器学习的原理

不论是Linear Regression, 还是Logistic Regression，都在试图拉近“预测”与“实际”之间的距离。当然，这个拉近的过程有两个部分：

- Hypothesis 是人们针对数据集的特点给出的一个猜想，人们用这个猜想去实现分类和预测功能。有了这个猜想，就需要验证它的有效性。所以有Cost Function的概念。
- Cost Function 用来衡量猜想与真实之间的距离，人们形象地成其为“成本”或“损耗”。任务就是将成本降到最低，距离拉到最近。

现在，或许可以先回答开头的问题了。机器学习的原理[6]，是希望通过简单的要求，让机器告诉人们数据的内在联系和规律，并将结果运用到实际中去。而这个希望，通过Cost Function大道至简的美，体现得淋漓尽致！

[5]分类函数也可以是非线性的，只需改变hypothesis的形式即可
[6]指有导师的(supervised learning)学习

3 实例：数字图像的神经网络识别

通过一个“对手写数字(0 9)图像的识别”实例，进一步加深对机器学习的理解和认识。

3.1 神经网络的优势

神经网络的优势，在于其对逻辑的表达。

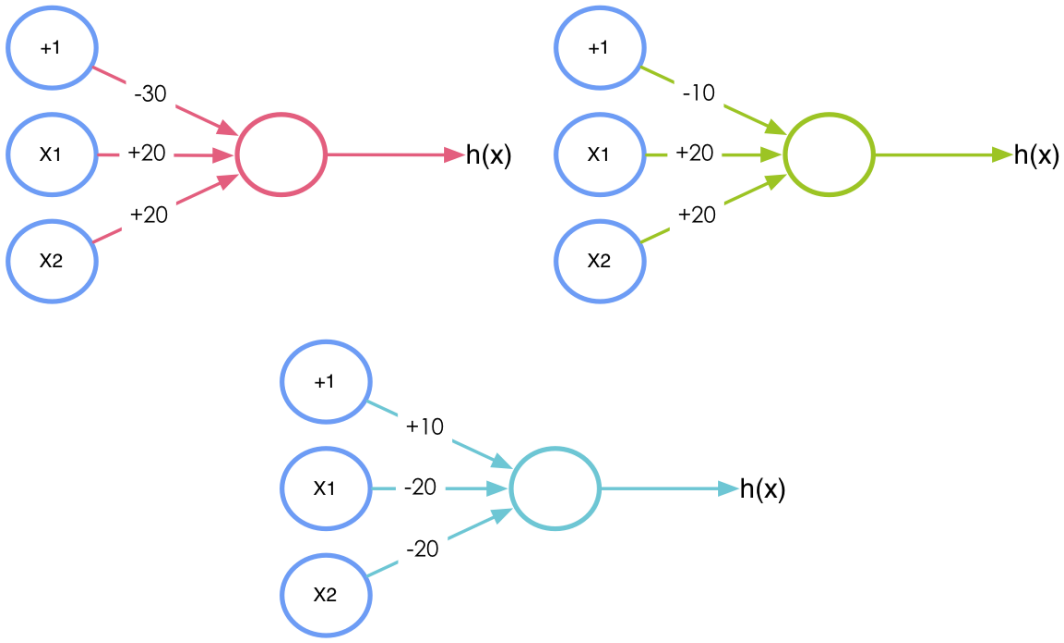


Figure 3: x_1 AND,OR,NOT x_2

对于关系“与、或、非”，可以得到这样的布尔逻辑表

x_1	x_2	$h_\theta(x)$
0	0	0
0	1	0
1	0	0
1	1	1

x_1	x_2	$h_\theta(x)$
0	0	0
0	1	1
1	0	1
1	1	1

x_1	x_2	$h_\theta(x)$
0	0	0
0	1	0
1	0	0
1	1	1

对于更复杂的逻辑运算，只需要结合上面的几个运算，给神经网络多加一层即可！

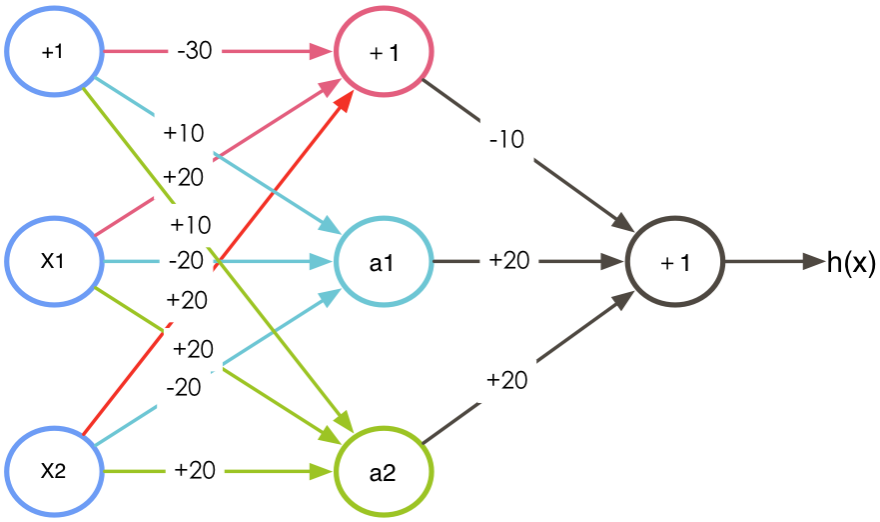


Figure 4: x_1 NOR x_2

x_1	x_2	a_1	a_2	$h_\theta(x)$
0	0	0	1	1
0	1	0	0	0
1	0	0	0	0
1	1	1	0	1

3.2 应用举例

现在将神经网络应用到对数字图像的识别上。这里我们有5000组训练样本，每个样本都是20 × 20的灰度图像。此处的模型有一个中间层，所以

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\left[-y_k^{(i)}log\left((h_\theta(x)_k^{(i)}\right) - \left(1-y_k^{(i)}\right)log\left(1-(h_\theta(x_k^{(i)}))\right)\right]$$

其中的$x_k^{(i)}, y_k^{(i)}$都是向量化的数据。

$$y = \begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\\vdots\\0\end{bmatrix}, \ldots, \begin{bmatrix}0\\0\\0\\\vdots\\1\end{bmatrix}.$$

这里用三层神经网络(图5)对原数据进行学习，是考虑到了多层网络结构对复杂逻辑的表达能力。这里的矩阵$\Theta^{(1)}, \Theta^{(2)}$都是我们需要确定的参数集合。

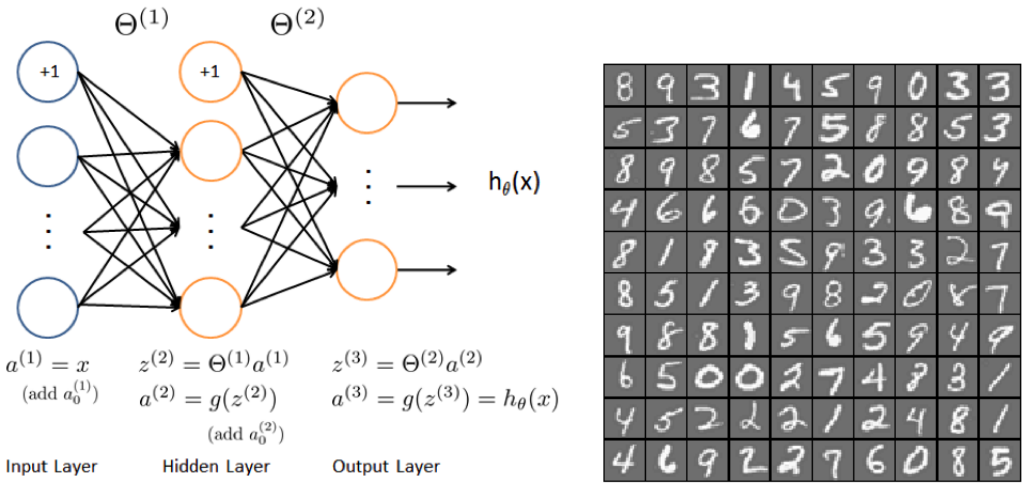


Figure 5: 神经网络模型与训练集

这里对参数求解的思想仍然是让$J(\theta)$在此参数下的损耗最小。然而在具体操作上，考虑到计算复杂度，便不能再使用偏导数的方法；而是以“回溯法(Backpropagation)”进行快速求解(详见[1])。

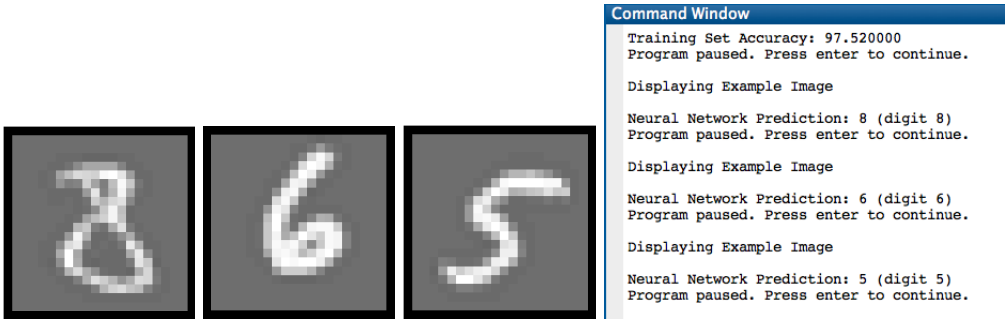


Figure 6: 检验结果

最后，将中间层可视化：

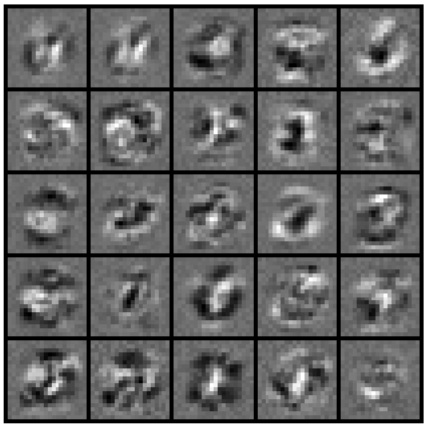


Figure 7: 中间层可视化

这时读者应该感受到了其中的逻辑变换。这里也恰好是体现神经网络算法优越性的地方，而且这其中的过程是人所不能预知的。具体地说，是人们只给机器提出了一个要求：

“让$J(\theta)$最小！”

后面发生的，只有机器知道。到了这里，机器的“学习”从神经网络算法中体现出来。这背后，一定是数据中隐藏的逻辑关系，是机器发现了它。

4 你需要了解的未来

最后一部分，感受机器学习的威力，以及能遇见的人工智能的未来。下面两端节选自文章[3]：

> “有种理论说，人类的智慧来源于一个单一的算法。
>
> 这个理论的实验依据是，人类大脑发育初期，每一部分的职责分工是不确定的，也就是说，人脑中负责处理声音的部分其实也可以处理视觉影像。人脑究其本质来说，是一台可以被调试以执行特定任务的通用型机器。”

这种猜想让人醍醐灌顶，但并非空想——只需再次思考文章开头关于新生婴儿的问题。

> “斯坦福大学人工智能实验室主任Andrew Ng(华裔，中文名吴恩达)领导的Google Brain项目，在人工智能方面走得更加前沿。去年6月，谷歌Google Brain运用深度学习的研究成果，使用 1000 台电脑创造出多达 10 亿个连接的“神经网络”，让机器系统学会自动识别猫，成为国际深度学习领域广为人知的案例之一。”

这意味着，机器的智能越来越高，也许有一天，人工智能将改变人类的生活。

参考文献

[1] Andrew Ng, Machine Learning, Coursera

[2] StatLearning/DataMining/MachineLearning方法系列简介, Stochastic Quant

[3] 你需要了解的未来：Andrew Ng与他的Google Brain项目及人工智能实践