转行 AI，如何给自己定位？

1. 为什么人工智能如此之火

众所周知，互联网行业的变化快到让所有从业者知悉并惶恐。就在笔者还在读研的时候，大家都在议论：Java还是C++更好就业的问题、移动互联网时代正在到来、安卓还是iOS开发，当然也有更前沿的大数据Hadoop与Spark等技术的横行。然而就在2016年的一场人机大战举行，随着 Google 的 AlphaGo 打败韩国围棋棋手李世乭之后，机器学习尤其是深度学习的热潮席卷了整个IT界。所有的互联网公司，尤其是 Google、 微软、百度、阿里、腾讯等巨头，无不在布局人工智能技术和市场。前有百度深度学习研究院，腾讯的AILab，最近几天阿里的“达摩院”又占据了各大新闻头条。

确实，人工智能时代正在到来。人工智能已经被各个国家提升到国家战略的高度！中国AI目前人才短缺，国内的供求比例仅为1：10，供需严重失衡。工信部教育考试中心副主任周明也曾在2016年向媒体透露，中国人工智能人才缺口超过500万人。

的必备基础，几乎如出一辙地都在强调数学的重要性。于是乎许多想转头AI学习机器学习的小伙伴们就被高深莫测的数学难倒了，吓得退出了，事实真的是这样吗？不学高深的数学就无法在工业界应用机器学习算法去解决实际问题了吗？答案显然不是这样的。

毫无疑问，数学是机器学习的基础。以至于传统机器学习最重要的流派叫做“统计学习理论”。但是这和转行AI学机器学习必须要具备深厚的数学“看起来是合情合理”，但事实却并非如此！想要解释清楚这个问题，首先我们需要解释掌握机器学习的三个层次到底是什么？每个层次是怎么定义的？掌握每个层次需要的必备知识是什么？

2.1 机器学习第一层：懂原理的调库君

虽然这个级别被我定义为最低的等级。但是大家可不要忘记，在中国目前的互联网公司中处于我定义的这一级别的从事AI工作的算法工程至少占到 **75%** 以上。所以，这也是我建议的绝大多数数学不是特别好的，没读过博士或者硕士的同学转行AI给自己定位的建议。这样，你就不需要去学习那些你很难搞懂的高深莫测的数学知识，也不需要掌握各种数学原理的推导。你只需要知道理工科本科毕业的那三门数学基础知识就完全可以了。这时候，知道常见算法的基本原理，以及各个参数的含义。OK，能用机器学习算法解决实际项目中的任务才是王道嘛！

举个栗子吧！燕哥本人曾经在公司做过的实战性数据挖掘任务，数据在文件“lppz.csv”中，第8列（为了与代码一致，从0计数）为预测数据。第10列开始的为特征向量列。于是如下的程序就能够完成预测任务，其实就这么简单。


```
from xgboost import XGBRegressor
from sklearn.model_selection import train_test_split
from xgboost import plot_importance
from matplotlib import pyplot as plt

# 读取文件原始数据
data = []
labels = []
labels2 = []
with open("lppz.csv", encoding='UTF-8') as fileObject:
```

```
model = XGBRegressor(max_depth=5, learning_rate=0.1,
n_estimators=145, silent=True, objective='reg:gamma')
model.fit(X_train, y_train)

ans = model.predict(X_test)

plot_importance(model)
plt.show()
```

2.2 机器学习第二层：会推公式的学术君

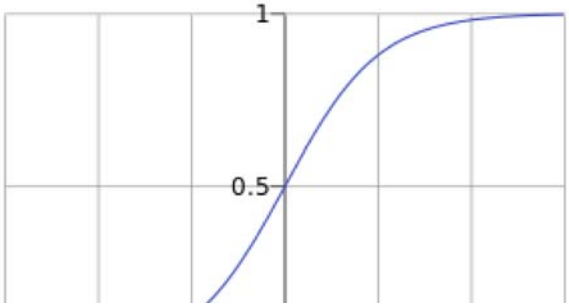
如果你想去BAT核心部门去做算法工程师，也许调库君是不够资格的。那么你需要满足一定的学术水平。也就是说，你不仅要懂得算法的大致原理，比如说决策树的分支是基于信息熵和信息增益的策略的。你还得知道常见算法的优化过程中使用的损失函数以及他的推导过程，这时候对大家的要求就相对来说比较高了。接下来我就以BAT机器学习算法面试必考题（这是福利，燕哥亲自经历的经验，大家要认真点哦！）。

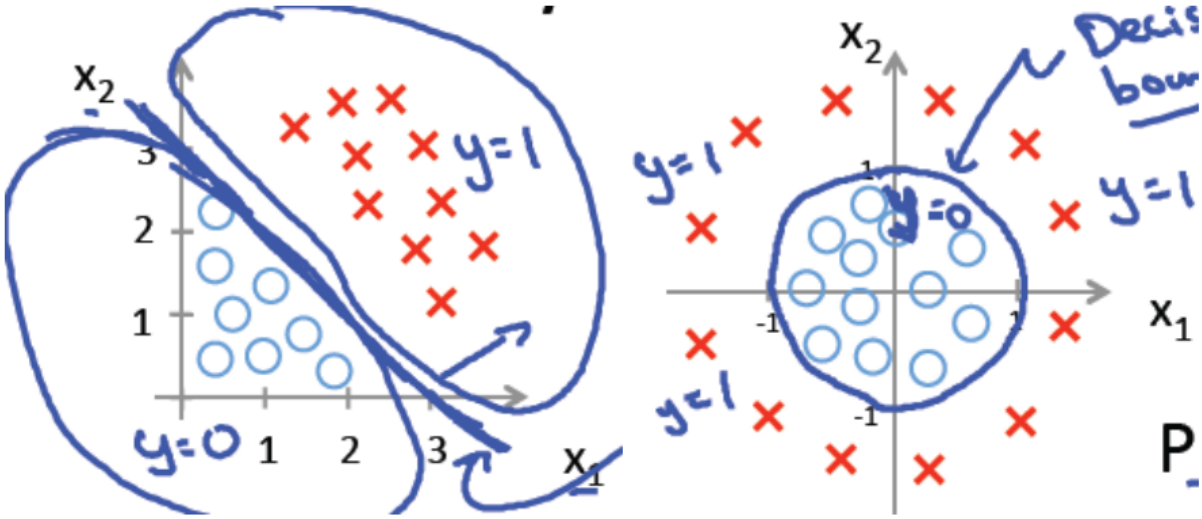
LR逻辑回归算法的损失函数的大致推导过程？

- 逻辑回归函数形式

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

- Sigmoid函数形状





对于线性边界的情况，边界形式如下：

$$\theta_0 + \theta_1 x_1 + ,\ldots, + \theta_n x_n = \sum_{i=1}^{n} \theta_i x_i = \theta^T x$$

构造预测函数为：

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

函数 $h_\theta(x)$ 的值有特殊的含义，它表示结果取1的概率，因此对于输入x分类结果为类别1和类别0的概率分别为：

$$\begin{aligned} P(y=1 \mid x;\theta) &= h_\theta(x) \\ P(y=0 \mid x;\theta) &= 1 - h_\theta(x) \end{aligned} \quad (1)$$

构造如下的损失函数：

$$L(\theta)=\prod_{i=1}^{m}P(y_i\mid x_i;\theta)=\prod_{i=1}^{m}(h_\theta(x_i))^{y_i}(1-h_\theta(x_i))^{1-y_i}$$

对数似然函数为：

$$l(\theta)=\log L(\theta)=\sum_{i=1}^{m}\left(y_i\log h_\theta(x_i)+(1-y_i)\log(1-h_\theta(x_i))\right)$$

最大似然估计就是求使 $l(\theta)$ 取最大值时的θ，其实这里可以使用梯度上升法求解，求得的θ就是要求的最佳参数。但是，在Andrew Ng的课程中将 $J(\theta)$ 取为下式，即：

$$J(\theta)=-\frac{1}{m}l(\theta)$$

因为乘了一个负的系数-1/m，所以取 $J(\theta)$ 最小值时的θ为要求的最佳参数。

梯度下降法求的最小值θ更新过程：

$$\frac{\delta}{\delta\theta_j}J(\theta)=-\frac{1}{m}\sum_{i=1}^{m}\left(y_i\frac{1}{h_\theta(x_i)}\frac{\delta}{\delta\theta_j}h_\theta(x_i)-(1-y_i)\frac{1}{1-h_\theta(x_i)}\frac{\delta}{\delta\theta_j}h_\theta(x_i)\right)$$

$$=-\frac{1}{m}\sum_{i=1}^{m}\left(y_i\frac{1}{g(\theta^{T}x_i)}-(1-y_i)\frac{1}{1-g(\theta^{T}x_i)}\right)\frac{\delta}{\delta\theta_j}g(\theta^{T}x_i)$$

$$=-\frac{1}{m}\sum_{i=1}^{m}\left(y_i\frac{1}{g(\theta^{T}x_i)}-(1-y_i)\frac{1}{1-g(\theta^{T}x_i)}\right)g(\theta^{T}x_i)(1-g(\theta^{T}x_i))\frac{\delta}{\delta\theta_j}\theta^{T}x_i$$

$$=-\frac{1}{m}\sum_{i=1}^{m}\left(y_i(1-g(\theta^{T}x_i))-(1-y_i)g(\theta^{T}x_i)\right)x_i^{j}$$

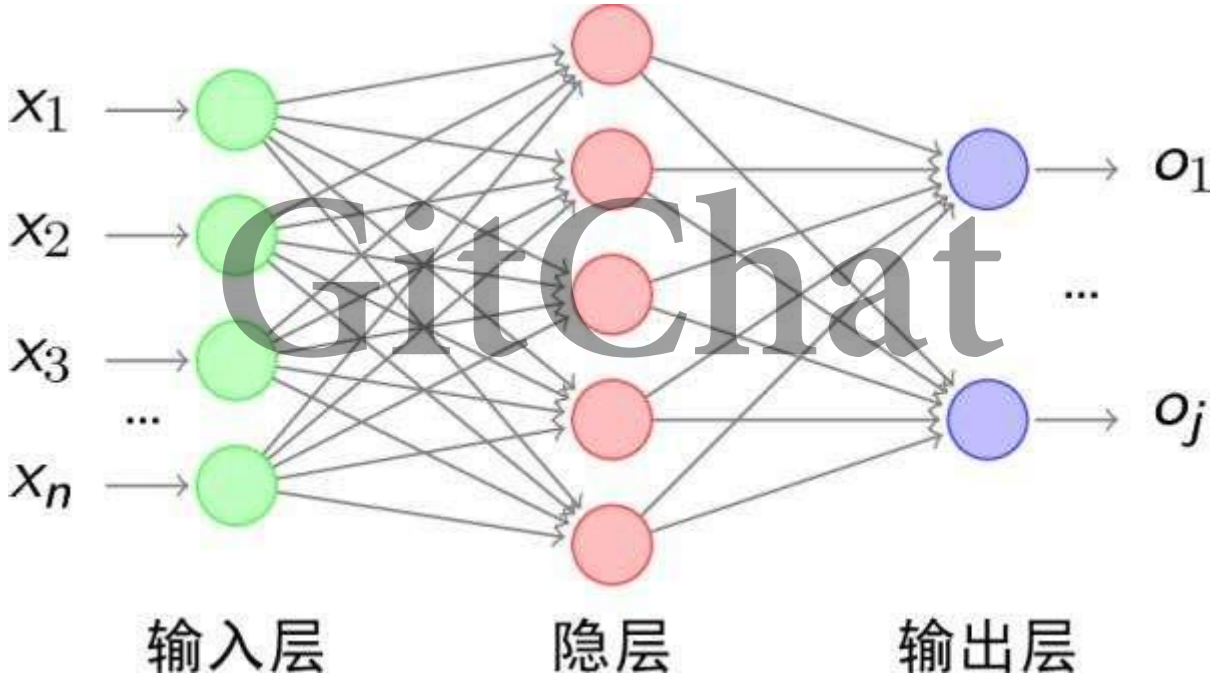
$$=-\frac{1}{m}\sum_{i=1}^{m}\left(y_i-g(\theta^{T}x_i)\right)x_i^{j}$$

2.3 机器学习第三层：创造新算法的大神君

这一层次的典型人物就是 **吴恩达**、**Hinton**、**Lecun**等大神级别的人物，这并非是仅仅是数学好或者是普通硕士生就能达到的水平，需要从博士生的艰苦数学基础的积累和长期不懈的研究探索才能达到的境界，一般如果转行的话，我个人建议别忘这个level上去想，容易伤害你自己！但是，如果你是本科生，而且数学也非常不错的话，向往这方面发展还是很有希望的。

那么，这个层级的大神到底如何牛呢？还是举个栗子吧！

著名的神经网络发明人 Hinton ，在别人发现感知机的时候，他觉得感知机能力有限，于是想出了神经网络来表示更复杂的学习任务。只是提出这么个模型不行，还得解决我该怎么优化参数，想出（与上面的会推是两回事）一个损失函数并且找到很好的BP反向传播算法来优化神经网络参数，这就是他牛的地方。并不是理解别人的思路，会推别人想出来的公式，而是从头到尾自己创造，并且算法实用性非常好，能解决很多现实的问题。



典型的神经网络结构

前面我也说了，其实现在互联网公司从事AI工作的 75% 以上的都是调库君，所以如果你数学确实不是很好，而且你又不是机器学习科班出身。AI之路并不是不可行，只是说，你不要对自己要求太高。不要相信所谓的数学不行就不能转行AI，也不要觉得转行AI我就一定能成为机器学习专家，这两者都是很极端的例子，正视现实才是唯一的出路。转行AI，准确定位自己很重要。

3.2 定位学术君

如果你是计算机或者数学博士毕业，虽然当初不是科班出身，但是你数学很棒！我觉得这样的小伙伴可以给自己定位为我所说的第二级别学术君。但是，你得专心有耐心的看专业书籍，并在有必要的时候研究原始算法论文搞清楚这个算法的来龙去脉。这样你才能对算法的理解达到一个新的高度。长期积累下去，你肯定能成为机器学习专家级别的人物，但是一定不要着急，慢慢来！

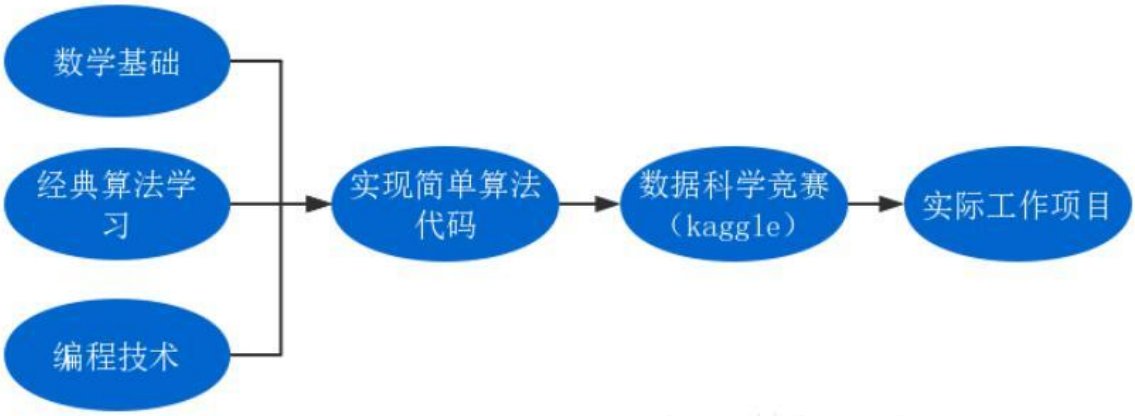
4. AI转行者的入门与提升

- **关于入门**

  对于入门，可能是所有机器学习初学者的痛点，这是一个艰辛的过程，因为这并没有一个统一的答案。对于转行者来说，我个人的建议是选择一个比较正规的视频课程全套教程学习班去学习一下。别着急，我知道很多人会发现学了一遍之后还是蒙圈的，觉得太难崩溃了。

  不怕，燕哥是个过来人，当初我研究生阶段的时候上了我们学校一位很著名的机器学习大牛的课程。其实上完了几乎所有学生都是蒙圈状态，当然我也不例外。那个老师本身就是北大数学系（中国数学最牛的院校）博士毕业的，在机器学习领域已经有20多年的研究。上课的时候，他觉得所有人都跟他一样，对基本的数学理论了如指掌。于是乎，这位大侠上课的时候就沉醉在自己的公式推导之中，自己无比的崇拜和陶醉，然而学生一个个的蒙圈状态。

  但是，后来在我自己买书自学的时候，我发现，其实当初上课还是很有用的，当我



机器学习的学习过程

俗话说，完事开头难。其实在机器学习的道路上提升也是一个不轻松的工作。毅力坚持是一方面，经验也是一方面。毅力是每个人自己的事，看你自己。关于经验，我觉得多写博客，多做实战，多关注比较著名的机器学习公众号。现在很多机器学习公众号的文章质量其实是相当不错的，尤其是近年来AI大火，大家的激情也是蛮高涨的。我推荐我个人的微信公众号《机器学习算法全栈工程师》，里面每一篇技术文章都是我严格把关的，作者也是我海里淘针试的寻找的。

最后，如果大家有任何问题，可以在我的读者圈提问，我会及时的回复的！