



# 机器学习简介与学习路线

---

机器学习入门到BAT系列群

推荐QQ群：342942219

群主跟据群友们的高频需求整理



# 前言

---

- 关于本人：6年机器学习相关经验，工作履历有腾讯、百度等，目前从事机器学习在计算广告方面的应用。
- 一门为入门级学员和初学者准备的机器学习基础课程，这门课程主要是讲解机器学习的基本概念、经典算法和模型、顺带讲解一下神经网络的基础和应用
- 为大家后续学习机器学习的高级高级课程、学习深度学习打下扎实的基础。

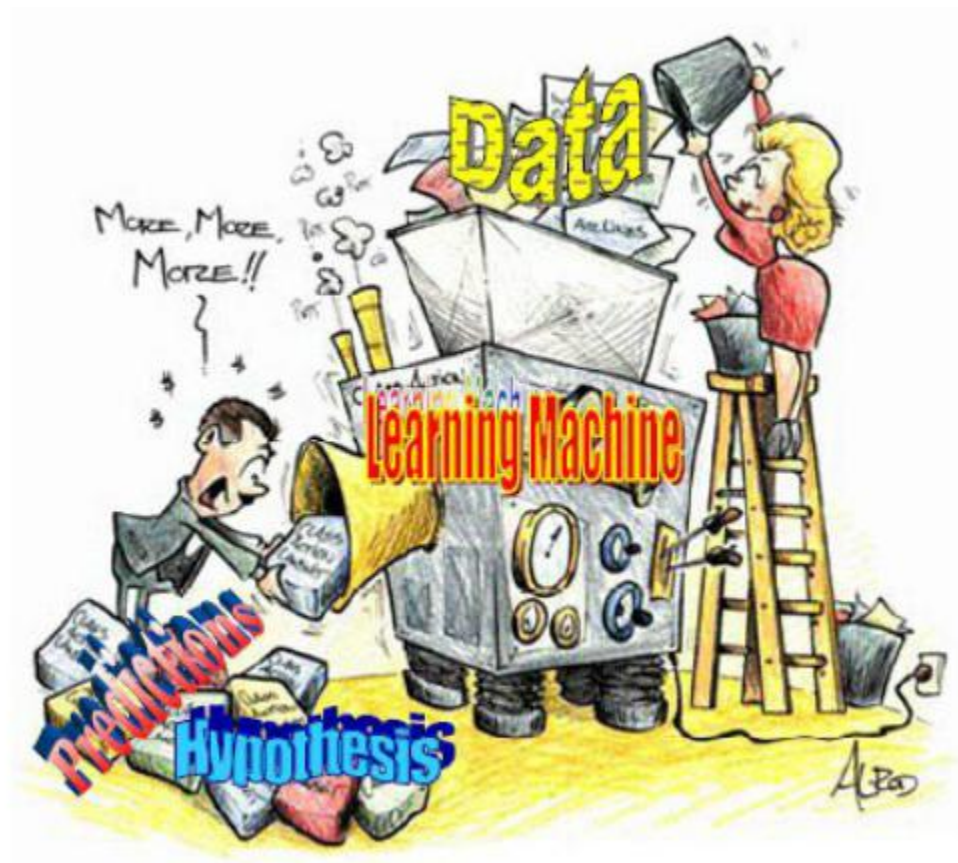


# 本次目标

---

- (一) 机器学习是什么
- (二) 机器学习能干什么
- (三) 机器学习基本概念
- (四) 机器学习理论简介
- (五) 机器学习的局限性
- (六) 思考题
- (七) 面试求职

# 机器学习是什么？





# 机器学习是什么？

- 探究和开发一系列算法来如何使计算机不需要通过外部明显的指示，而可以自己通过数据来学习，建模，并且利用建好的模型和新的输入来进行预测的学科。

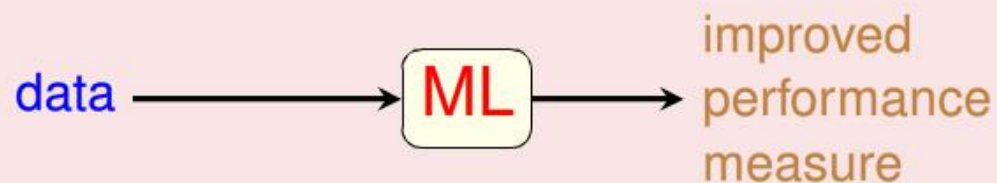
- Which of the following is best suited for machine learning?

- ① predicting whether the next cry of the baby girl happens at an even-numbered minute or not
- ② determining whether a given graph contains a cycle
- ③ deciding whether to approve credit card to some customer
- ④ guessing whether the earth will be destroyed by the misuse of nuclear power in the next ten years



# 机器学习是什么？

**machine learning**: improving some performance measure  
with experience **computed** from data



- 1 exists some 'underlying pattern' to be learned  
—so 'performance measure' can be improved
- 2 but **no** programmable (easy) **definition**  
—so 'ML' is needed
- 3 somehow there is **data** about the pattern



# 它能干什么



# 它能干什么







# 演示

---

- caffe demo
- <http://demo.caffe.berkeleyvision.org/>



# 我能干什么

- 互联网公司都需要大量的机器学习工程师，很多的创业公司都已经开始搞机器学习和大数据了 这是一个非常有想象空间的领域。当然 大疆创新 face++ 第四范式 地平线 这些非互联网公司也做的很不错
- 根据处理的数据类型不同 有文本处理（NLP，这个需求最大）、语音识别（如百度语音搜索、讯飞语音）、视频识别（如无人车）以及其他的数据挖掘，如金融征信、量化交易、智能硬件中的数据挖掘等。以熟知的互联网公司举例，今日头条做个性化推荐、滴滴打车做智能调度算法

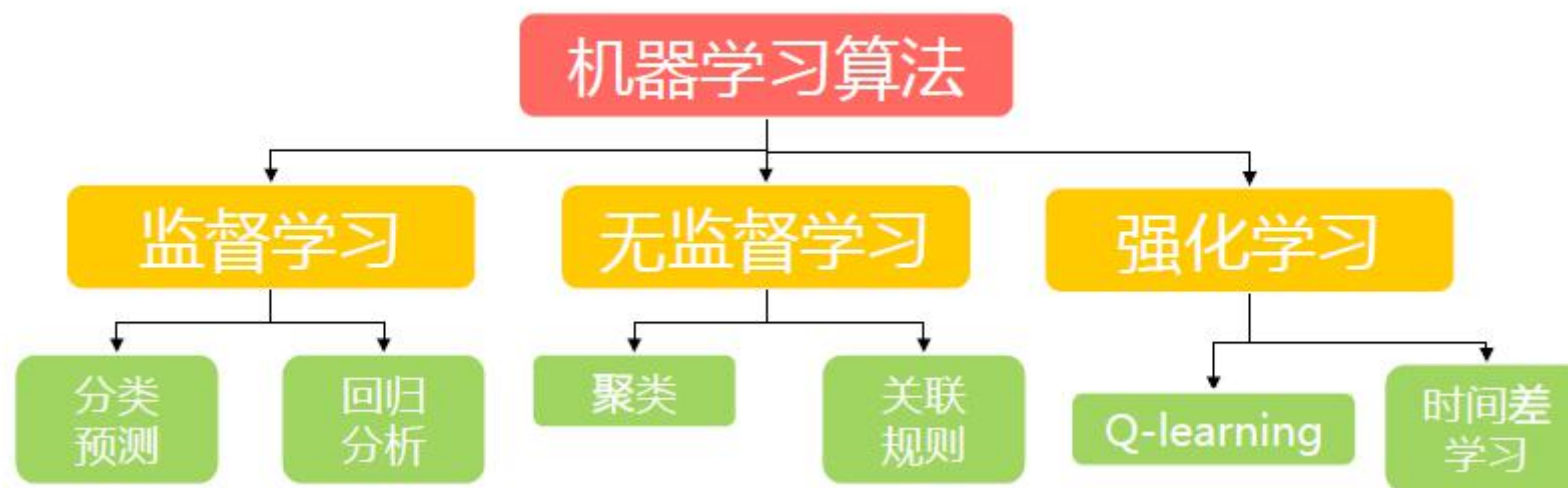
# 这堆概念都是什么鬼？

- 机器学习
- 数据挖掘
- 模式识别
- 人工智能



- PR(模式识别)、DM(数据挖掘)属于 AI 的具体应用
- 人工智能是一种应用领域，机器学习是实现人工智能的一种手段，但是不限于此。

# 理论框架





# 监督学习 (Supervised Learning)

- 从标记的训练数据来推断一个功能的机器学习任务
- • 根据输出变量的类型，监督学习分为以下两类学习问题：
  - Ø回归：定量输出称为回归，或者说是连续变量预测
  - Ø分类：定性输出称为分类，或者说是离散变量预测
- 回归和分类是什么区别？面试官喜欢问



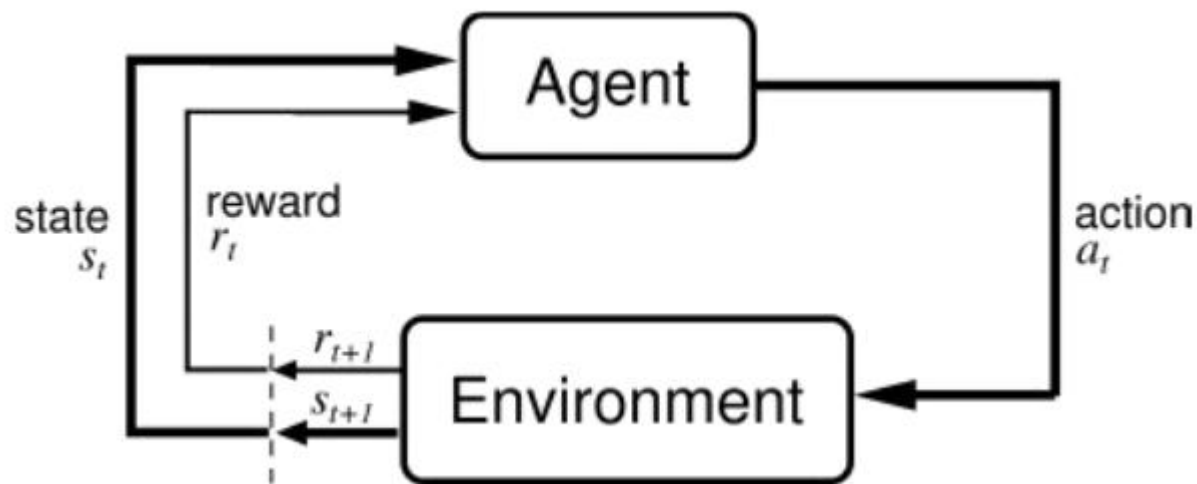


# 无监督学习（Unsupervised Learning）

- 用于处理未被标记的样本集，模型能够自主学习到知识。
- • 常用：GMM、聚类、降维、深度学习的逐层训练等
- K-means是无监督的聚类方法，KNN是有监督的分类方法，不要弄混

# 强化学习（Reinforcement Learning）

- 强化学习就是智能系统从环境到行为映射的学习，以使奖励信号(强化信号)函数值最大
- • 基本组件
- ∅环境
- ∅agent（交互对象）
- ∅动作
- ∅反馈（回报，奖赏
- • 应用：机器人等





# 经典算法

## Machine Learning Algorithms *(sample)*

	<u>Unsupervised</u>	<u>Supervised</u>
<u>Continuous</u>	<ul style="list-style-type: none"><li>• Clustering &amp; Dimensionality Reduction<ul style="list-style-type: none"><li>○ SVD</li><li>○ PCA</li><li>○ K-means</li></ul></li></ul>	<ul style="list-style-type: none"><li>• Regression<ul style="list-style-type: none"><li>○ Linear</li><li>○ Polynomial</li></ul></li><li>• Decision Trees</li><li>• Random Forests</li></ul>
<u>Categorical</u>	<ul style="list-style-type: none"><li>• Association Analysis<ul style="list-style-type: none"><li>○ Apriori</li><li>○ FP-Growth</li></ul></li><li>• Hidden Markov Model</li></ul>	<ul style="list-style-type: none"><li>• Classification<ul style="list-style-type: none"><li>○ KNN</li><li>○ Trees</li><li>○ Logistic Regression</li><li>○ Naive-Bayes</li><li>○ SVM</li></ul></li></ul>



# 机器学习不是万能的

- 在手机上的一些照相app中，有这样一种功能，通过面部拍照可以识别出人的年龄，但是经过亲测发现，在面部光线充沛和光线偏暗两种情况下，程序判断出的人的年龄差别很多，差十年都是很正常的，这是为什么呢？
- 机器学习本质上还是一种统计方法，它只讲求统计意义未必考虑的是事情的本质。



# 机器学习不是万能的

- 对于机器学习模型来说，准确率和召回率都不可能是100%，极端case难以避免
- 还记得“大明湖畔”的GAN么？通过GAN合成一些噪声一样毫无意义的图片，就能轻易骗过你高大上的机器学习模型
- 对于金融交易、自动驾驶等事关大笔资金安全、人身安全的场景中，不要盲目迷信AI。不要把你的安全全部交给模型。正确的做法是？规则（经验）+模型 融合





# 学会批判热点

- 为了否定和质疑别人的机器学习模型，有哪些思考的角度？
- Facebook聊天机器人开始自创语言了？程序的bug！
- 在通过照片识别同性恋这样的任务中，斯坦福的人通过平均人脸 的模型发现 “同性恋男性更少留胡子” 那么会不会出现这样的情况 一个人模型判定为 “非同性恋” 的家伙剃掉胡子以后再用模型判断，就变成了同性恋，从机器学习的角度，完全会有这样的情况发生



# 思考题

---

- 判断对错
- 1、回归和分类都是有监督学习问题 ()
- 2、对回归问题和分类问题的评价 最常用的指标都是 准确率和召回率 ()
- 3、输出变量为有限个离散变量的预测问题是回归问题；
- 输出变量为连续变量的预测问题是分类问题； ()
- 4、[百度校招试题]给定  $n$  个数据点，如果其中一半用于训练，另一半用于测试，则训练误差和测试误差之间的差别会随着  $n$  的增加而减小 ()



# 机器学习工程师面试题

---

- 大数据基础： 大小表的map-reduce
- NLP基本知识： edit distance
- 数据结构与coding：  
手写快排、二分查找(C++/python写)、leetcode
- 深度学习理论：  
lstm原理 CNN做文本分类的网络结构
- 传统机器学习理论：  
boosting的原理， 手推SVM， 手写造轮子： kmeans的  
hadoop实现



# 怎样成为机器学习工程师

- 辅助技能（linux python 数学）
- 机器学习概念和实战
- 深度学习理论和实战 tensorflow 等工具
- 具体业余领域的训练（NLP、视觉、SLAM等）
- 最好还懂基础编程语言(C++/java)
- 最好有大数据的基础(hadoop spark)

将字符串A变为字符串B所需要的最小操作次数

操作定义：插入、删除、替换

A: 1234sdf  
B: 234345SDFG

<http://collabedit.com/w2td2>

求一个数组的 最大不存在相邻元素的子数组和，时空复杂度尽量小

[1, -3, 4, -2, 2, 9, 4, 5]



# 我该学什么

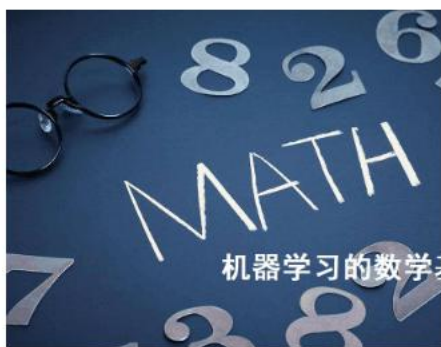
---

- 基本的语言能力： linux Python C++或者JAVA
- 算法和数据结构功底
- 机器学习理论和一定的实战经验
- hadoop 或 spark（加分）
- 实习或者竞赛经验（加分）



# 怎么学

- step1: 拜师 剃发明志
- step2: 扎马步



机器学习的数学基础[AI打怪系列]



机器学习的数据库基础[AI打怪系..]



机器学习的编程基础之python A..

- step3: 学武功



深度学习入门级课程[AI打基础系..]



机器学习入门级学习材料（适合...

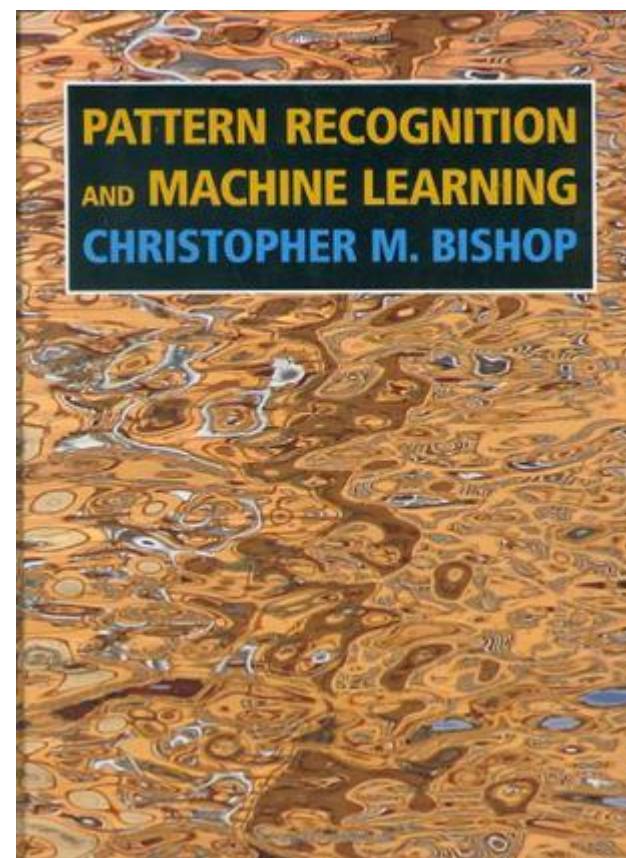
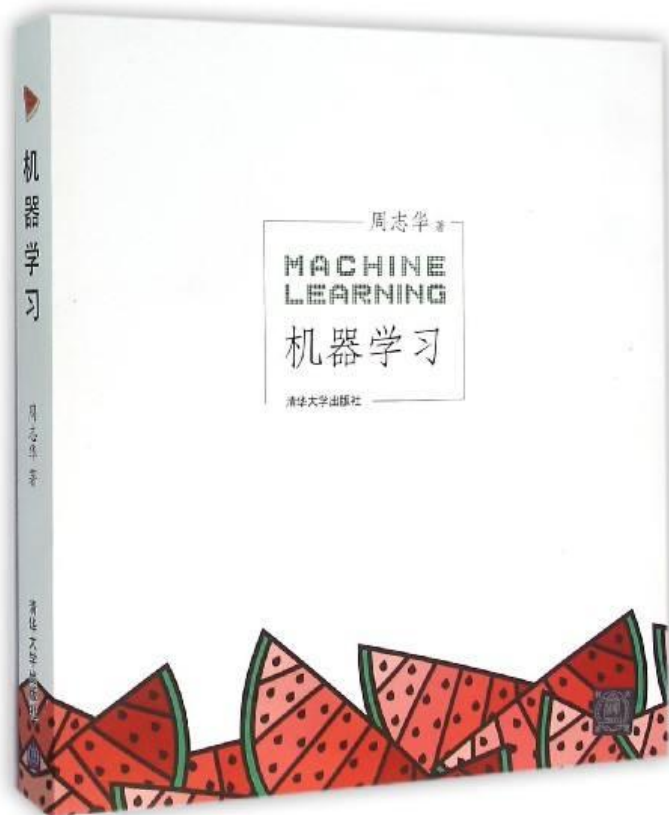


# 怎么学

---

- step4: 结合具体应用场景（如NLP 计算机视觉）进行实践 做项目 参加竞赛或者实习
- step5:找工作 面试
- step6:修炼圆满 走进BAT

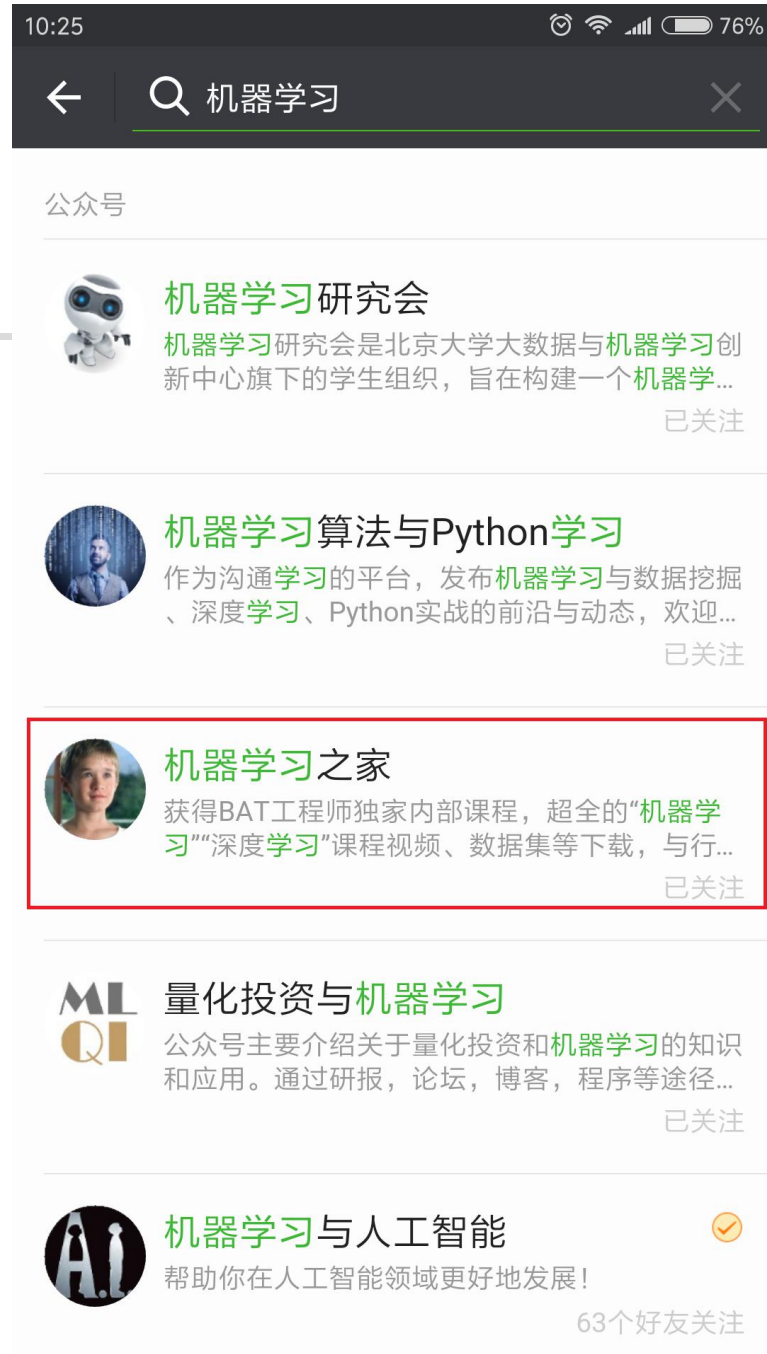
# 推荐资料



平台全覆盖： 公众号

公众号：机器学习之家

以“机器学习”关键字排名第三



# 平台全覆盖：论坛

## ■ 人工智能A7论坛

网址：<http://www.aqinet.cn>



扫码访问：



采集

逻辑回归模型的前世今生

### 精选视频

更多>>



Hinton深度学习视频课程  
下载

[重磅]斯坦福大学自然语言处理(NLP)课程视

智普教育python就业培训视频教程

Coursera林轩田《机器学习基石》课程的视频

### 行业资讯

更多>>

### 精华文章

更多>>



逻辑回归模型的前世今生

神经机器翻译NMT的提升方法和代码实例

CS 294: Deep Reinforcement Learning, Spr  
斯坦福Deep Learning for Natural Language

### 电子书

更多>>



[机器学习入门经典]统计  
学习方法pdf下载

[首发]Windows 10 下Caffe框架的安装、配置

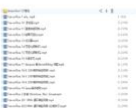
周志华《机器学习》pdf下载

深度学习(最全的中文版)\_2017年新书.pdf百

### 本站公告

更多>>

### 最新文章



[莫烦]Tensorflow模  
频教程(带源码)

机器学习的数学基础

李宏毅-Generative Adversarial Netw  
Alex Li高清Python入门视频教程

更多精彩内容，关注微信公众号



激  
转至

机器学习讨论群(QQ)送100金币



# 平台全覆盖：微信群矩阵

- 加入微信群矩阵
- 与超过一千名微信小伙伴一起成长
- 扫码，备注申请原因



机器学习公开课1群



机器学习公开课2群



机器学习技术交流-2群



机器学习技术交流-总群



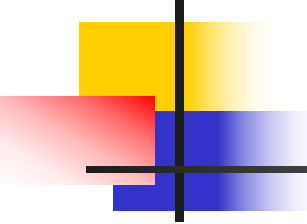
人工智能技术交流-1群



人工智能线下活动（全国）



人工智能线下沙龙（北京）



# 平台全覆盖：QQ社群矩阵

---

## 机器学习综合交流群：

两千人入群 239146371（已满）

推荐群 **342942219**（推荐）

## 分方向交流群：

自然语言处理 524640947

计算机视觉 145700860

资源分享论坛群 239640103

大数据竞赛交流群 195334079

## 分地区交流群（不定期组织线下交流）：

北京地区机器学习 423183544

武汉地区机器学习 281073115

上海地区机器学习 497108144

广东地区机器学习 285331102

西安地区机器学习 163103729

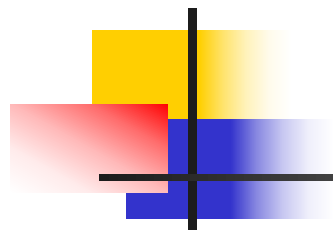
成都地区机器学习 373082415

南京地区机器学习 335064162

## 分高校交流群：

华中科大机器学习 **377867390**

中南大学机器学习 245661601



感谢大家！

恳请大家批评指正！