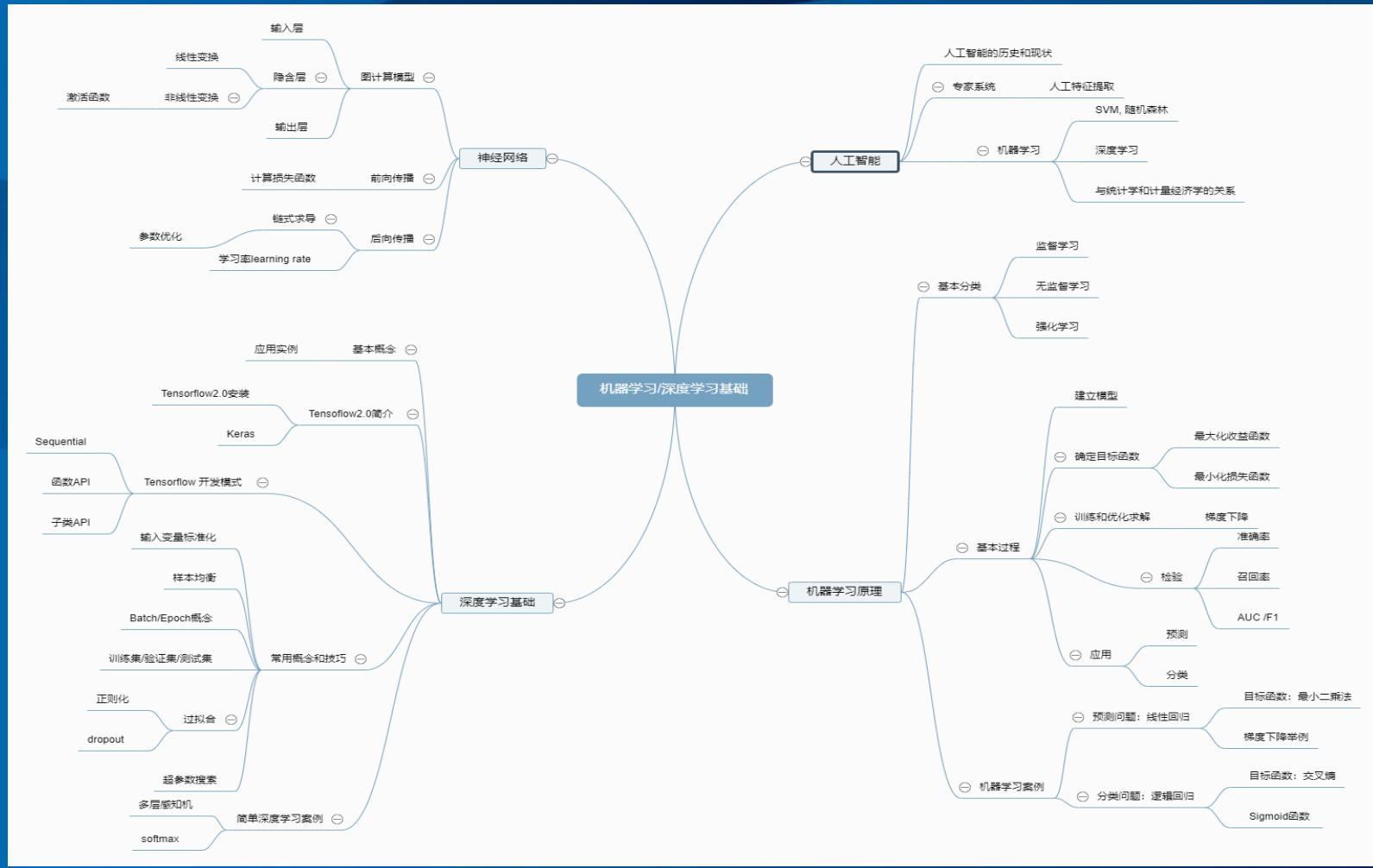
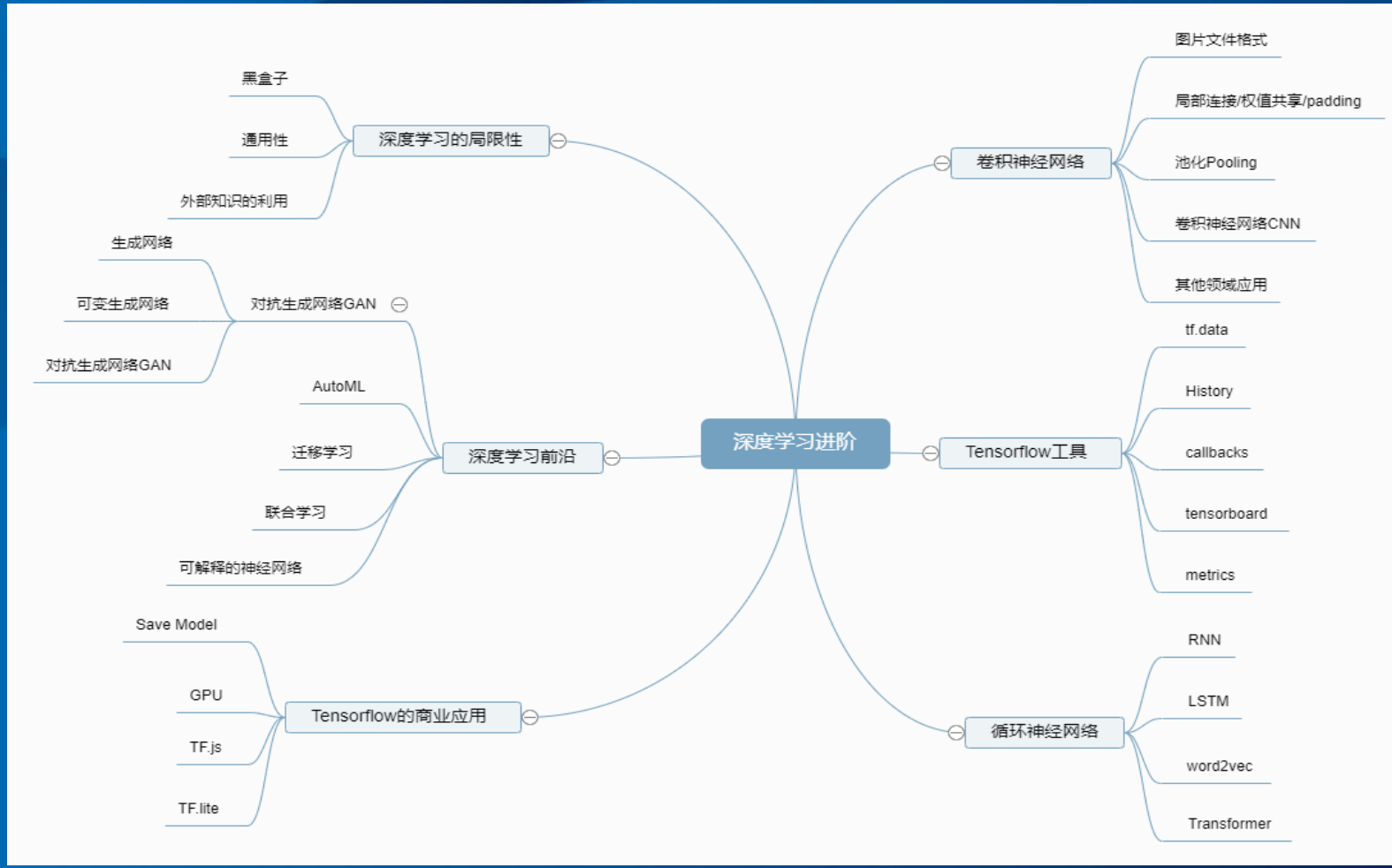
机器学习和深度学习原理和课程概述

资本市场研究所
曾刚


机器学习/深度学习基础
人工智能
人工智能的历史和现状
专家系统
人工特征提取
机器学习
SVM, 随机森林
深度学习
与统计学和计量经济学的关系
机器学习原理
基本分类
监督学习
无监督学习
强化学习
基本过程
建立模型
确定目标函数
最大化收益函数
最小化损失函数
训练和优化求解
梯度下降
检验
准确率
召回率
AUC /F1
应用
预测
分类
机器学习案例
预测问题：线性回归
目标函数：最小二乘法
梯度下降举例
分类问题：逻辑回归
目标函数：交叉熵
Sigmoid函数
深度学习基础
神经网络
图计算模型
输入层
隐含层
线性变换
非线性变换
激活函数
输出层
前向传播
计算损失函数
后向传播
链式求导
参数优化
学习率learning rate
Tensorflow 开发模式
Sequential
函数API
子类API
Tensoflow2.0简介
基本概念
应用实例
Tensorflow2.0安装
Keras
常用概念和技巧
输入变量标准化
样本均衡
Batch/Epoch概念
训练集/验证集/测试集
过拟合
正则化
dropout
超参数搜索
简单深度学习案例
多层感知机
softmax


深度学习进阶
深度学习的局限性
黑盒子
通用性
外部知识的利用
深度学习前沿
对抗生成网络GAN
生成网络
可变生成网络
对抗生成网络GAN
AutoML
迁移学习
联合学习
可解释的神经网络
Tensorflow的商业应用
Save Model
GPU
TF.js
TF.lite
卷积神经网络
图片文件格式
局部连接/权值共享/padding
池化Pooling
卷积神经网络CNN
其他领域应用
Tensorflow工具
tf.data
History
callbacks
tensorboard
metrics
循环神经网络
RNN
LSTM
word2vec
Transformer

1. 人工智能概述

- 弱人工智能：完成某个特定任务，例如Siri, 人脸识别，机器翻译，自动驾驶
- 通用人工智能：自我学习，通用性
- 强人工智能：超越人类智能

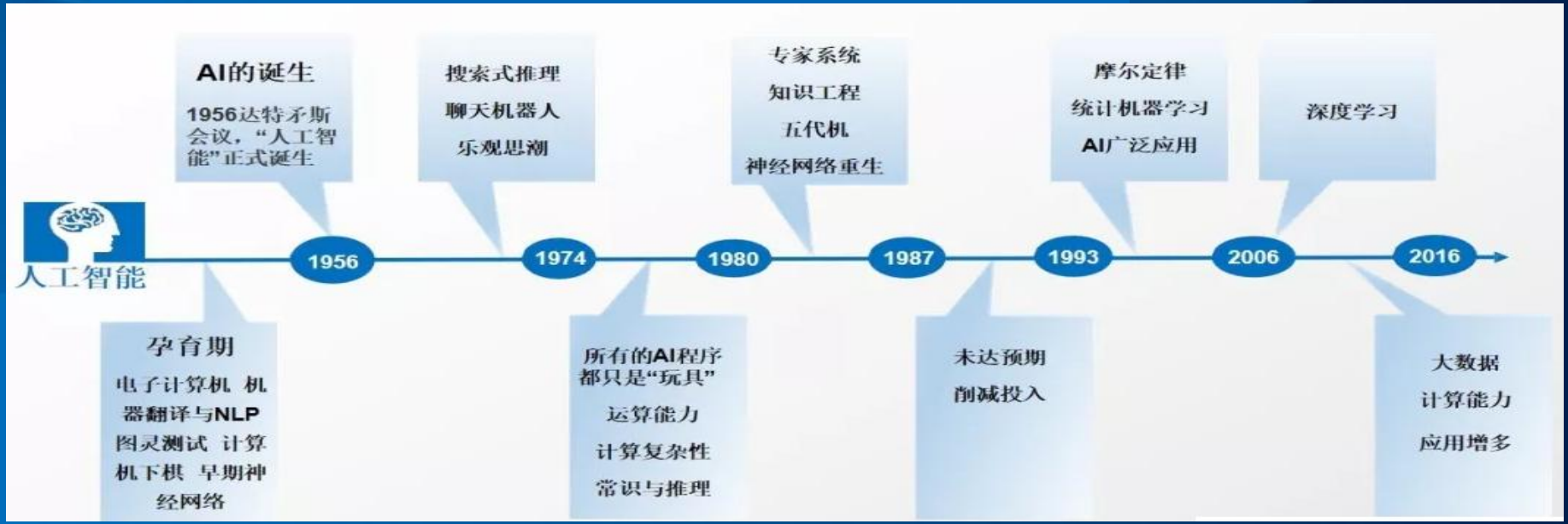
1. 正确理解当前人工智能技术的现状和前景

机器快速，准确但有点“笨”，而人缓慢，不精准却充满创造力！

----李飞飞

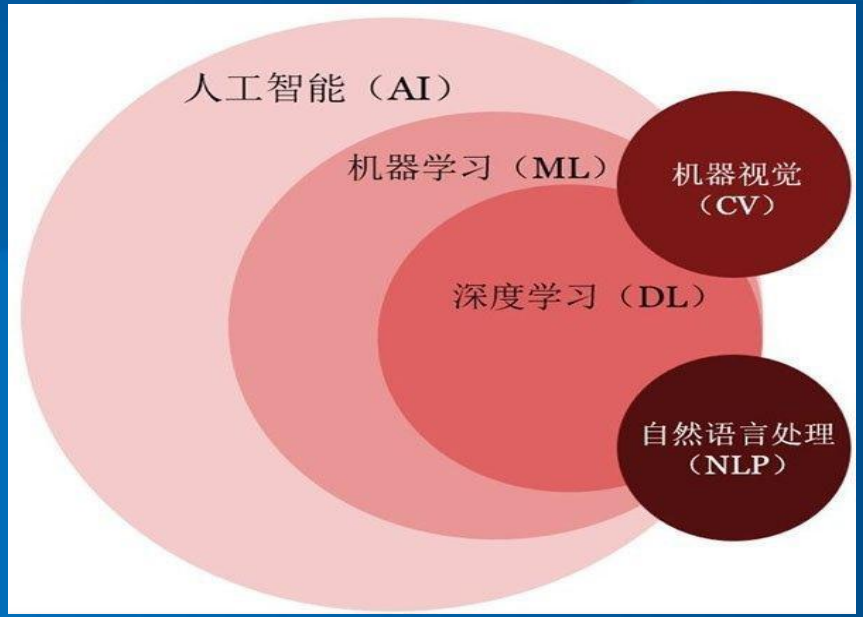
- 机器学习其实只是在总结数据
- 机器学习只能发现相关性，而无法发现因果关系
- 机器学习很难利用现存的知识、
- 计算机训练得到的模型人类很难理解
- 机器学习同人类智慧内在机制差异就像鸟类飞行与人类飞机的差异

人工智能发展历程



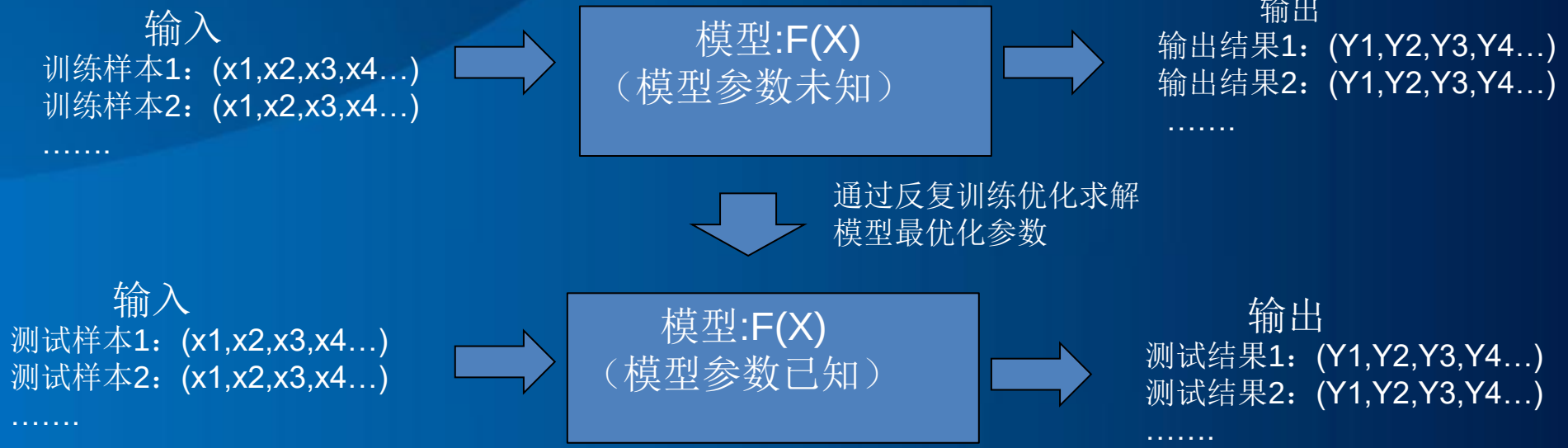
- 80年代神经网络出现，90年代末SVM等机器学习算法
- 2006年Geoffery Hinton、Yann LeCun和Yoshua Bengio教授论文使深度学习技术成为趋势
- 2016年AlphaGo带动深度学习热潮

人工智能，机器学习和深度学习的关系



- 机器学习是人工智能的一种途径或子集，它强调学习而不是计算机程序。
- 机器学习之父Tom Mitchel：每个机器学习都可以被精准地定义为：1.任务；2.训练过程；3.模型表现。
- 机器学习强调通过Optimization实现Prediction, 统计学强调Statistically modeling and inference, 计量经济学则重在causal inference。

2. 机器学习的概念流程图

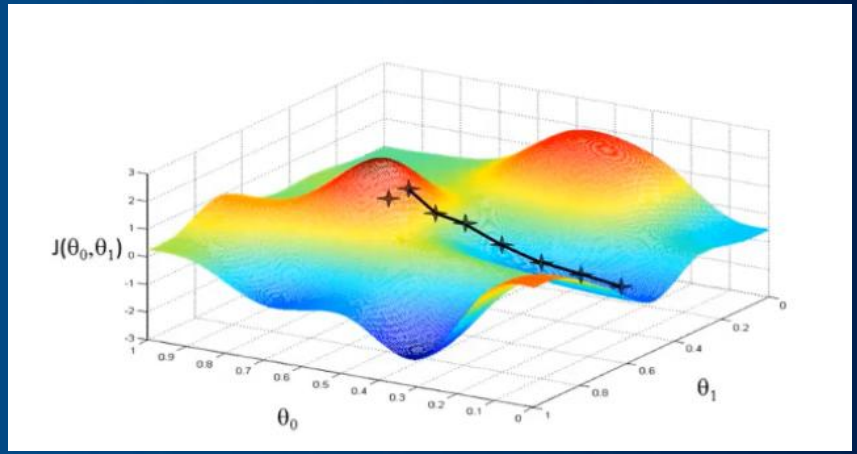


2. 机器学习的种类

- 监督学习（Supervised learning）：监督学习是使用已知正确答案的示例来训练网络。已知数据和其一一对应的标签，训练一个预测模型，将输入数据映射到标签的过程。监督式学习的常见应用场景如分类问题和回归问题。
- 无监督学习（Unsupervised learning）：训练数据并不被事先标识，适用于具有数据集但无标签的情况。学习模型是为了推断出数据的一些内在结构特点。常见的应用场景包括聚类、主成分分析等。
- 强化学习（Reinforcement Learning, RL）：属于无监督学习的一种，用于描述和解决在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题，例如棋类比赛、自动驾驶。

3.机器学习的具体过程

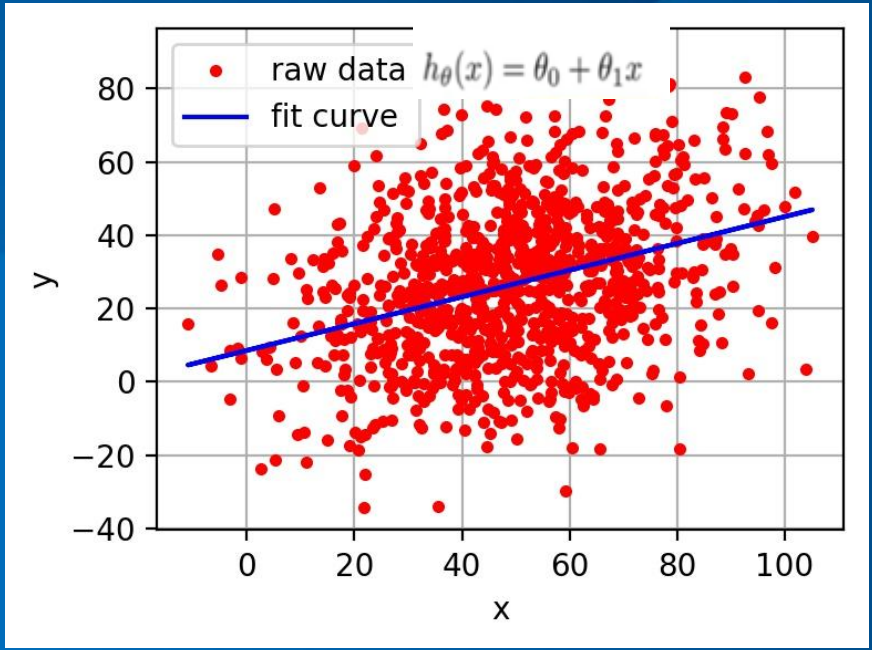
- 建立模型
- 确定最优化目标函数
  (损失函数loss Function)
- 使用训练集样本数据进行训练和优化求解
  (梯度下降Gradient Decent)
- 使用测试集样本对模型效果检验
- 模型应用



梯度下降过程

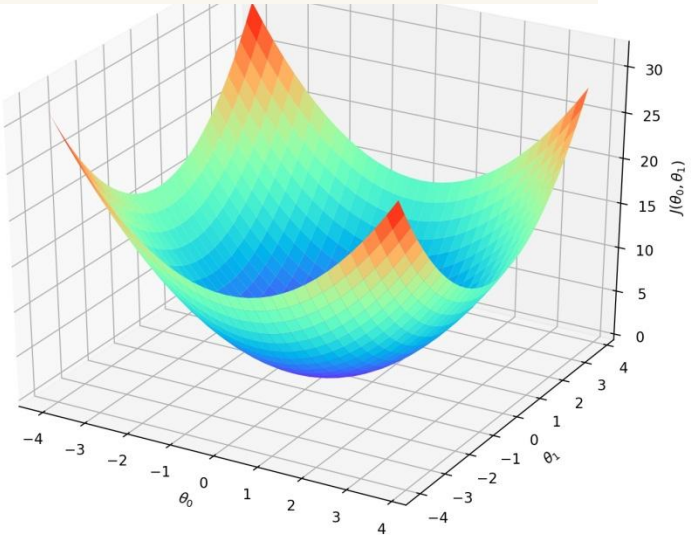
3.机器学习过程举例：线性回归（预测问题）

线性回归模型表示



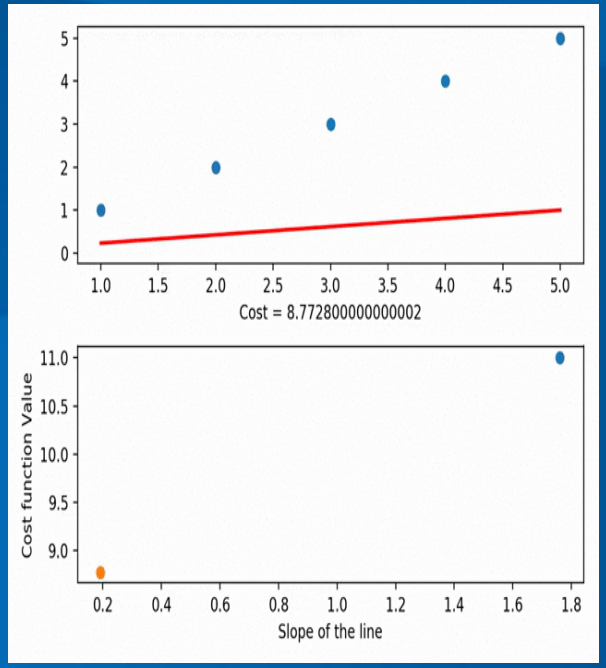
损失函数（最小二乘法）

$$J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})^2$$



资料来源https://www.cnblogs.com/zhoubindut/p/12160116.html

3.机器学习过程举例：线性回归—模型梯度下降优化求解过程



Gradient descent algorithm

repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

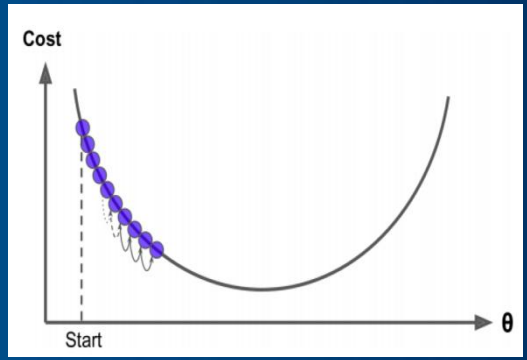
(for $j = 1$ and $j = 0$)

}

Linear Regression Model

$$h_\theta(x) = \theta_0 + \theta_1 x$$

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$



资料来源：Bilibili网站吴恩达机器学习公开课，
https://baijiahao.baidu.com/s?id=1627880370781148770&wfr=spider&for=pc

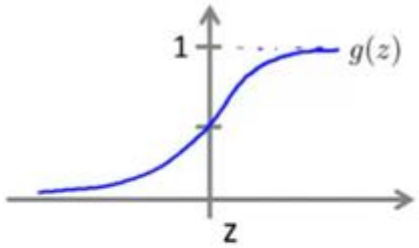
3.机器学习过程举例：逻辑回归（分类问题）

模型表示（函数g称为Sigmoid函数）

Logistic regression

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

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



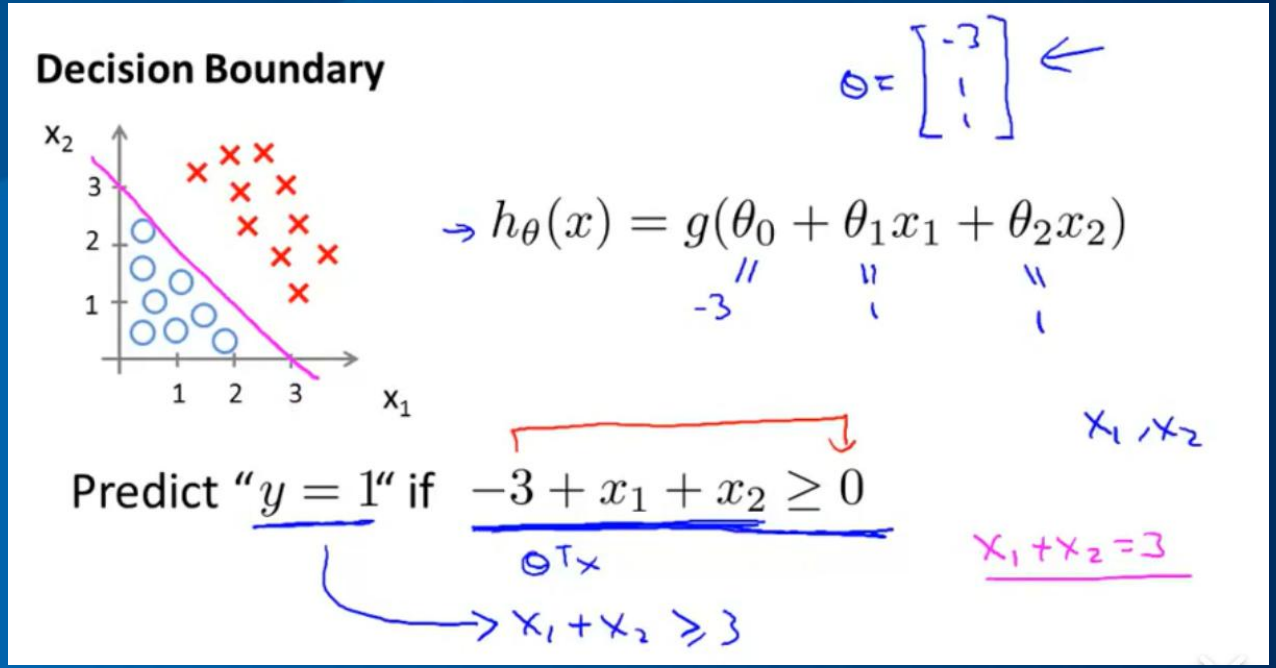
Suppose predict "$y = 1$" if $h_\theta(x) \geq 0.5$

predict "$y = 0$" if $h_\theta(x) < 0.5$

资料来源：吴恩达机器学习公开课

3.机器学习过程举例：逻辑回归（分类问题）

模型表示

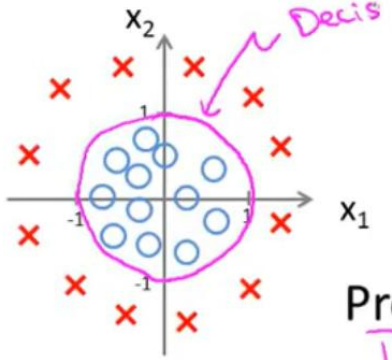


资料来源：吴恩达机器学习公开课

3.机器学习过程举例：逻辑回归（分类问题）

非线性边界情形

Non-linear decision boundaries



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

$$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$

$x_1^2 + x_2^2 = 1$

$x_1^2 + x_2^2 \geq 1$

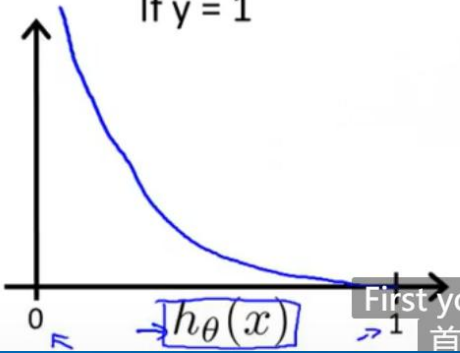
资料来源：吴恩达机器学习公开课

3.机器学习过程举例：逻辑回归（分类问题）

逻辑回归的损失函数：交叉熵Cross Entropy（确保损失函数凸性.，易于最优化求解）

Logistic regression cost function

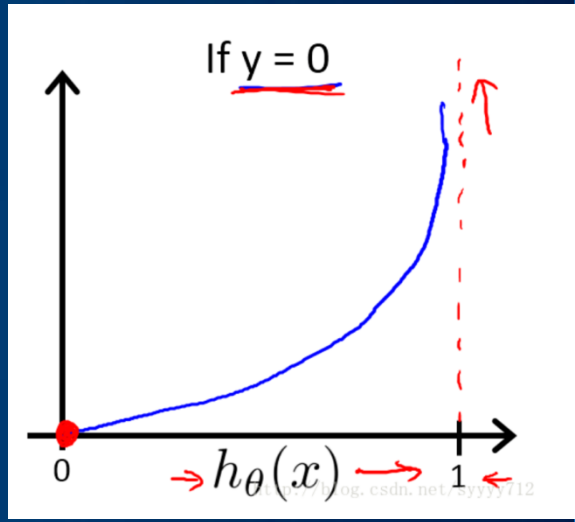
$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



Cost = 0 if $y = 1, h_\theta(x) = 1$
But as $h_\theta(x) \to 0$
$Cost \to \infty$

Captures intuition that if $h_\theta(x) = 0$, (predict $P(y = 1|x; \theta) = 0$), but $y = 1$, we'll penalize learning algorithm by a very large cost.

First you notice that if
首先 你注意到



资料来源：吴恩达机器学习公开课

3.机器学习过程小结

线性回归模型(连续变量预测)

$$h_\theta(x) = \theta_0 + \theta_1 x$$

逻辑回归模型（离散分类判定）

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

线性回归模型损失函数

$$J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})^2$$

逻辑回归模型损失函数

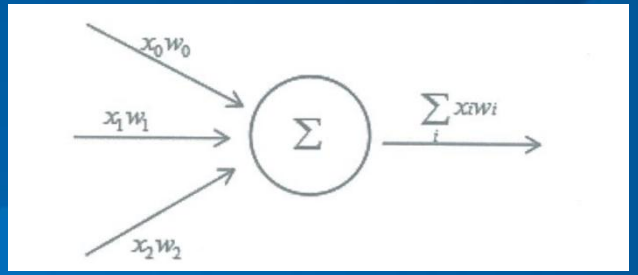
$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

参数优化（梯度下降）

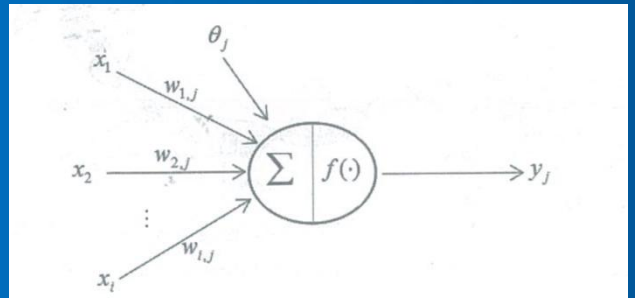
$$\begin{aligned} &\text{repeat until convergence } \{ \\ &\quad \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \\ &\quad \text{(for } j = 1 \text{ and } j = 0) \\ &\} \end{aligned}$$

3. 线性回归 + 逻辑回归奠定了神经网络的基础

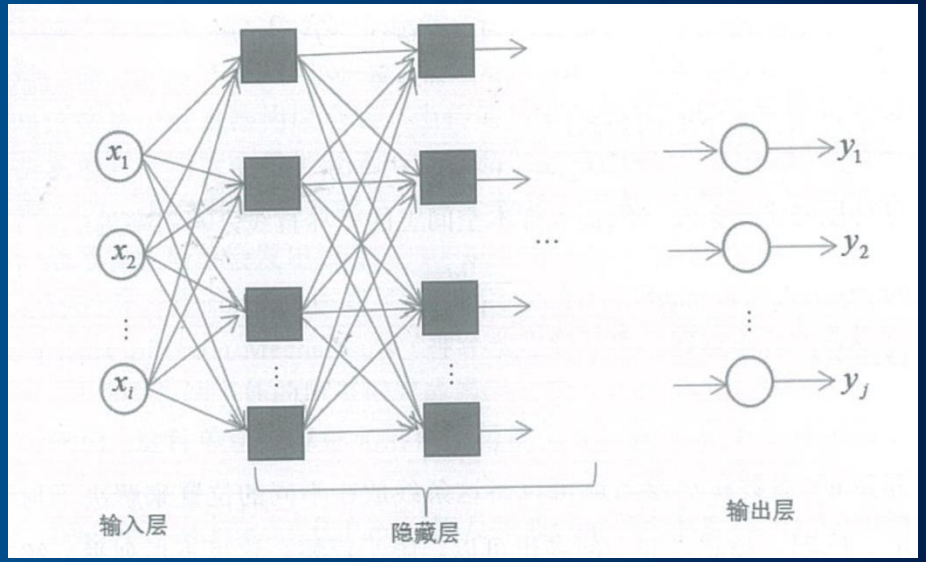
单个神经元
（类似线性回归模型）



单个神经元加激活函数
（类似逻辑回归模型）



多层神经元构成神经网络



4.机器学习模型判定准确度检验

- 将已经打好标签的样本分为训练集和测试集两部分
- 训练集用于训练模型
- 测试集用于检验模型有效性
- 常见的模型评价术语（假设样本分为正例（positive）和负例（negative））：

➢ True positives(TP):

被正确地划分为正例的个数，即实际为正例且被分类器划分为正例的实例数；

➢ False positives(FP):

被错误地划分为正例的个数，即实际为负例但被分类器划分为正例的实例数；

➢ False negatives(FN):

被错误地划分为负例的个数，即实际为正例但被分类器划分为负例的实例数；

➢ True negatives(TN):

被正确地划分为负例的个数，即实际为负例且被分类器划分为负例的实例数。

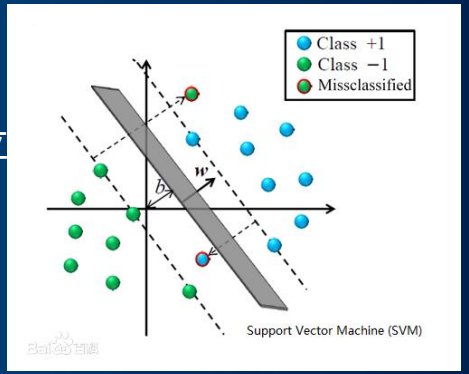
4. 机器学习模型检验指标

- 正确率（accuracy）：accuracy = (TP+TN)/(P+N)，正确率是被分对的样本数在所有样本数中的占比，通常来说，正确率越高，分类器越好。
- 错误率（error rate)：描述被分类器错分的比例，error rate = (FP+FN)/(P+N)，对某一个实例来说，分对与分错是互斥事件，所以accuracy =1 - error rate。
- 精度（precision）：precision=TP/(TP+FP)，也称为查准率，精度是精确性的度量，表示被分为正例的示例中实际为正例的比例。
- 召回率（recall）：recall=TP/(TP+FN)=TP/P，召回率是覆盖面的度量，也称为查全率，度量有多少个正例被分为正例。
- ✓ F1 score：精度和召回率反映了分类器分类性能的两个方面。如果综合考虑查准率与查全率，可以得到新的评价指标F1-score，也称为综合分类率。
- ✓ ROC（Receiver Operating Characteristic）曲线是以FP_rate和TP_rate为轴的曲线，ROC曲线下面的面积我们叫做AUC，面积越大模型越可靠

5. 其他常用机器学习方法

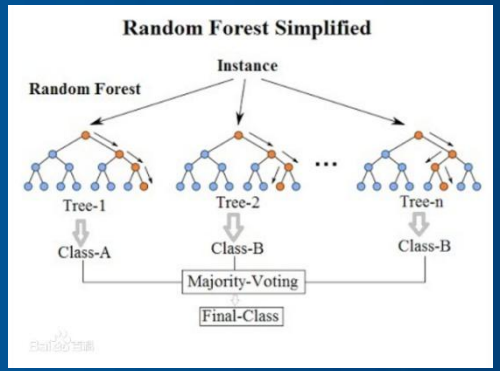
- 支持向量机（SVM)

是一类按监督学习方式对数据进行二元分类的广义线性分类器（generalized linear classifier），其决策边界是对学习样本求解的最大边距超平面（maximum-margin hyperplane）. SVM可以通过核方法（kernel method）进行非线性分类。



- 随机森林(Random Forest)

指的是利用多棵决策树对样本进行训练并预测的一种分类器，是一种通过建立多个分类器模型，各自独立学习和预测并将结果合成单一预测的集成学习方法。



参考文献

- https://www.bilibili.com/video/BV1rJ411S7z4?from=search&seid=9083110598564107417 吴恩达机器学习课程公开课
- https://createmomo.github.io/2018/01/23/Super-Machine-Learning-Revision-Notes/#tableofcontents 英文版机器学习笔记
- https://www.cnblogs.com/zhoubindut/p/12130350.html 机器学习系列博客
- https://tf.wiki/ 简单粗暴 TensorFlow 2： A Concise Handbook of TensorFlow 2

谢谢！