机器学习基础

--原理、方法与实践

王志鹏

- 课程目的
  - 基本概念
  - 新问题：判断，选择，解决
  - 技巧：数学、编程

- 鸣谢
  - eCafe
  - 成臣

- 四次课
  - → 1.定义、前沿成果、基础方法
  - 2. 基础方法讲解与实践
    - 逻辑回归（logistic regression）
    - 凸优化的一阶方法，二阶方法
    - 正则化
    - 随机梯度下降
    - k-means、 梯度提升树(gradient boosting decision trees)
  - 3. 神经网络原理讲解与实践
    - Multi-layer perceptron: tensorflow
      - 反向传播、激活函数、dropout及其他相关知识点
    - Convolutional neural network: keras
    - Recurrent neural network: keras
  - 4. 强化学习方法介绍与实践
    - Alpha go论文介绍
    - Policy gradient

- 自我介绍
  - 十年AI路，漫漫无坦途，
    梦想做生物，非常爱吃鱼

- 手←眼，笔←嘴

- 资源推荐
  - Andrew Ng
    - Machine learning https://www.coursera.org/learn/machine-learning

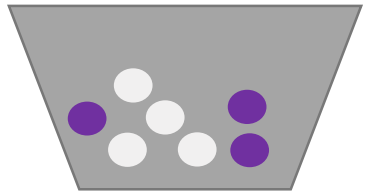
  - 林轩田
    - 机器学习基石
      https://www.youtube.com/playlist?list=PLXVfgk9fNX2I7tB6oIINGBmW50rrmFTqf

    - 机器学习技法
      https://www.youtube.com/playlist?list=PLXVfgk9fNX2IQOYPmqjqWsNUFl2kpk1U2

- 资源推荐
  - 数学基础
    - 《深度学习》 第二章到第五章
  - 参考书
    - 周志华
      - 《机器学习》
    - Chapman & Hall
      - *Machine Learning: An Algorithmic Perspective, Second Edition*
  - 读了三遍的书
    - Richard S. Sutton
      - Reinforcement Learning: An Introduction
      - http://incompleteideas.net/sutton/book/the-book-2nd.html

- 机器学习定义
  - 对于某类任务T和（对任务的）性能指标P，一个计算机程序能够从经验E里学习，也就是说，基于经验E，（计算机程序）在任务T上的性能指标P有所提升。 -- Tom Mitchell
    - T P E
      - 学习：从经历（历史数据）里面找到道理，来做的更好
  - 机器学习就是不直接编程而让计算机有学习（解决问题）的能力 -- Arthur Samuel
    - 自动从**数据**中发现**规律**，并使用规律**解决问题**
      - 使用**优化方法**找到**模型**基于**数据**的**最适合的参数**，使用得到的参数通过模型**完成任务**

- 机器学习与统计
  - 实用中心主义盛行
  - 与计算机硬件的进步结合紧密



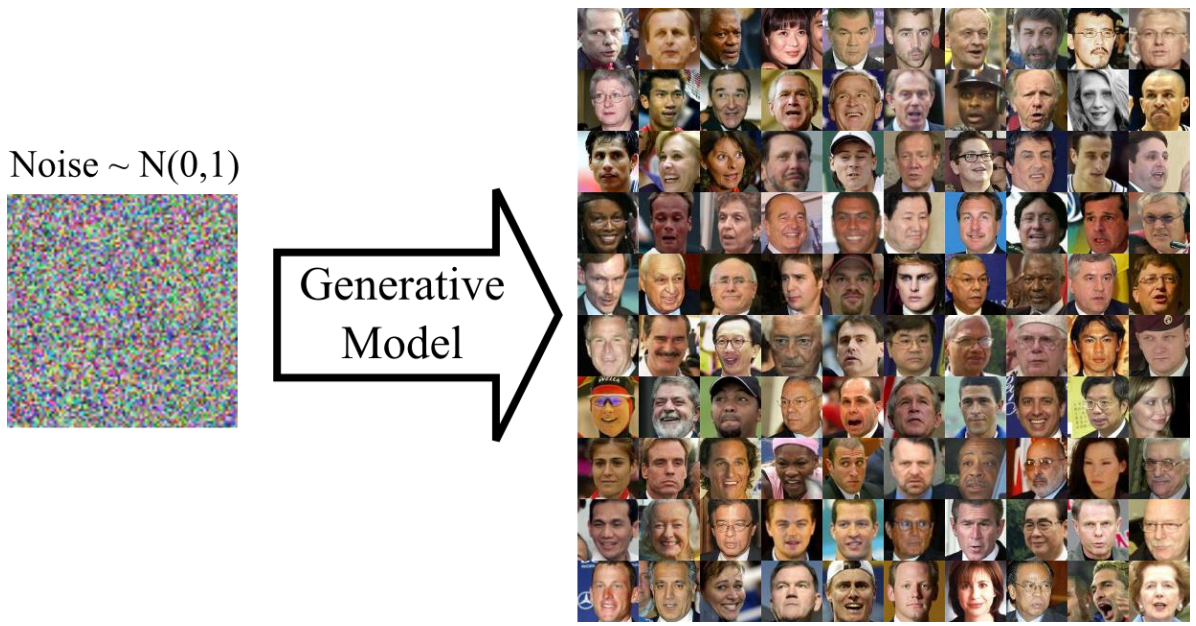
- 机器学习与人工智能（使机器有类人的智能）
  - 子集
  - 近年来进展非常多

- 近期进展
  - 图片识别（image net）
    - 百万量级图片
    - 1000类
    - Top 5 guess error rate: < 0.5% （随机猜error rate 99.5%）
  - 语音识别、自动翻译
    - 微信的语音识别
    - Google、facebook等公司的端到端翻译
  - 游戏
    - Alpha Go
    - Atari
      - https://www.youtube.com/watch?v=V1eYniJ0Rnk
    - Dota2
      - https://blog.openai.com/dota-2

- 近期进展
  - 写程序

```
/*
 * Increment the size file of the new incorrect UI_FILTER group information
 * of the size generatively.
 */
static int indicate_policy(void)
{
  int error;
  if (fd == MARN_EPT) {
    /*
     * The kernel blank will coeld it to userspace.
     */
    if (ss->segment < mem_total)
      unblock_graph_and_set_blocked();
    else
      ret = 1;
    goto bail;
  }
  segaddr = in_SB(in.addr);
  selector = seg / 16;
  setup_works = true;
  for (i = 0; i < blocks; i++) {
    seq = buf[i++];
    bpf = bd->bd.next + i * search;
    if (fd) {
      current = blocked;
    }
  }
  rw->name = "Getjbbregs";
  bprm_self_clearl(&iv->version);
  regs->new = blocks[(BPF_STATS << info->historidac)] | PFMR_CLOBATHINC_SECONDS << 12;
  return segtable;
}
```

  - 点石成金、变水为油
    - Generative adversarial networks



- 繁荣
  - 得益于高速计算、廉价存储
  - 工程、经验 与 基础理论 并重
  - 媲美小规模工业革命
    - 以新方法解决旧问题
    - 解决新问题

- 对于某类任务T和性能指标P，一个计算机程序能够从经验E里学习，也就是说，基于经验E，在任务T上的性能指标P有所提升。
  - 怎么保证提升
  - 内在规律存在并被有效发现
  - 一定程度的统计不变性

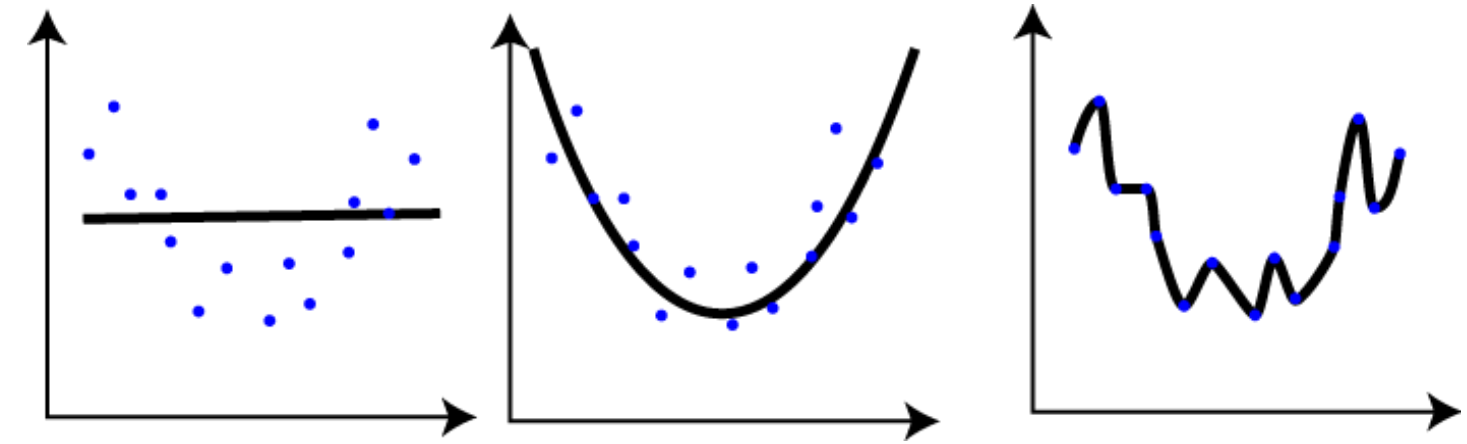
- 何时使用机器学习
  - 火星巡游：人类经验缺乏
  - 语音识别、物体识别：人类经验不易描述
  - 高频决策：超出人类决策速度
  - 大规模推荐系统：问题规模大、没有准确物理定律描述

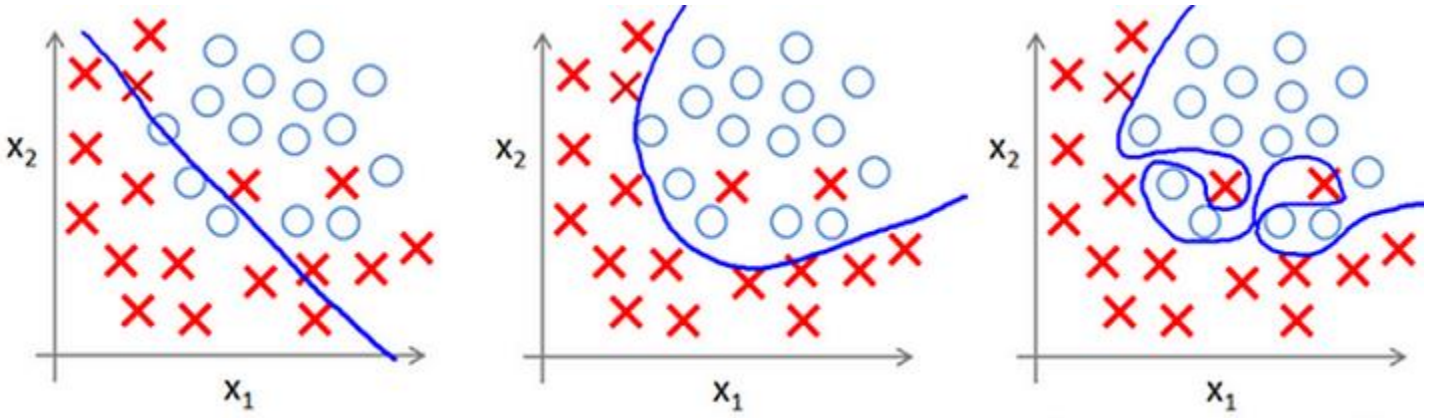
- 机器学习分类
  - 监督学习（supervised learning）：分类与回归
    - 经验E有明确的标签
    - T：数据 → 标签
    - P：从数据得到标签与E标签进行对比
  - 无监督学习（unsupervised learning）：
    - 经验E没有明确的标签
    - T：数据 → 有用的结构来表示数据的内在（聚类，降维，数据生成）
    - P：经常依赖于外部主观判断
  - 强化学习（reinforcement learning）：感知环境、进行决策、获得反馈、达成目的
    - 经验E的标签是环境延迟获得
    - T：状态（数据），反馈 → 决策
    - P：环境给出评判

- 监督学习（使用优化方法找到模型基于数据的最适合的参数，使用得到的参数通过模型完成分类和回归）
  - 通常做法：交叉验证与滚动前测
  - 模型选择：拟合能力（欠拟合、过拟合）与模型泛化
  - 回归模型的评估、分类模型的评估
    - 方法
    - 指标

- 通常做法
  - 汽车销量预测
    - 月份 + 地区 → 汽车销量
    - 已有五年数据 → 对未来预测
      - 选择模型 → 使用数据估计模型参数 → 使用模型和参数做出预测
      - 关键：预见模型的预测能力
    - 五年数据分为训练数据与测试数据（完全不重叠，比例8:2或者7:3）
      - 小目标：在测试数据上的表现好
      - 交叉验证
      - 严格前测

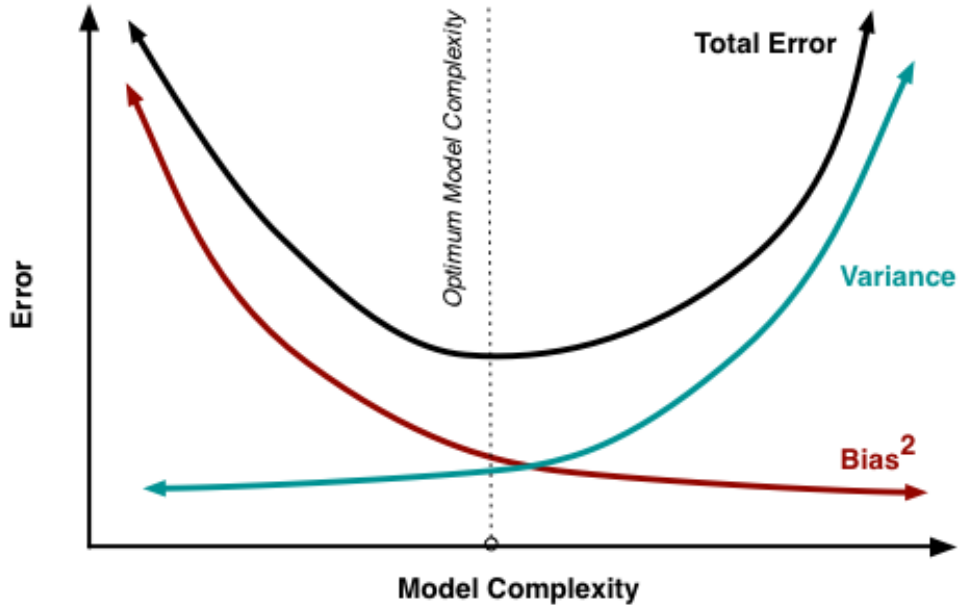
- 模型拟合能力：欠拟合与过拟合
  - 模型复杂度上升 → 模型拟合训练数据的能力越强 & 模型在训练数据上的表现与在测试数据上的表现越难一致

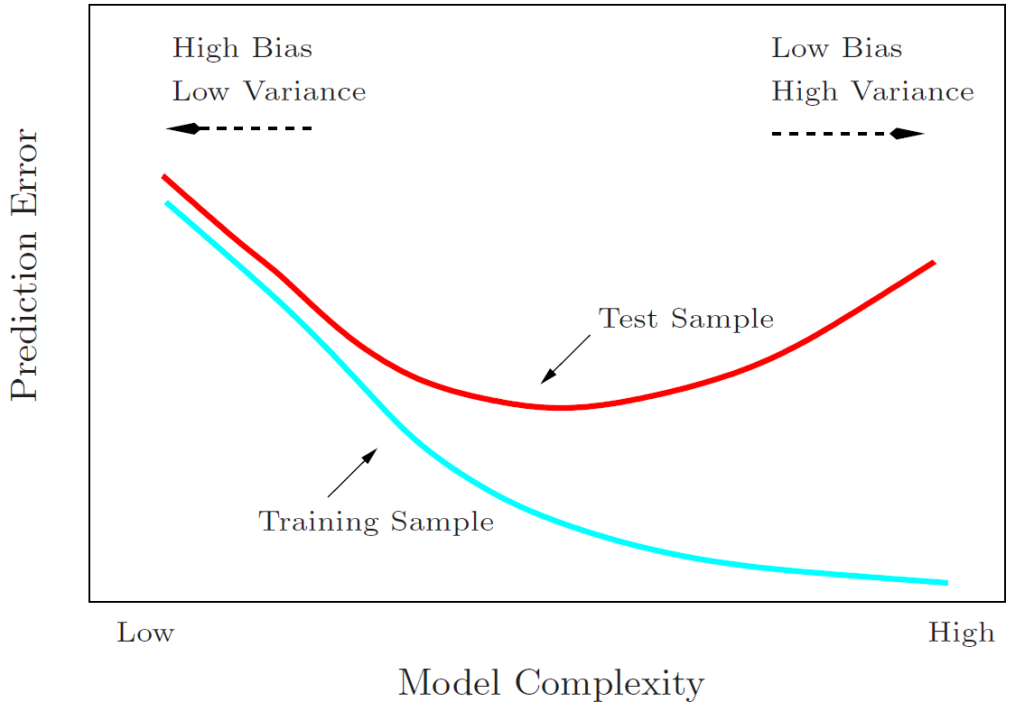




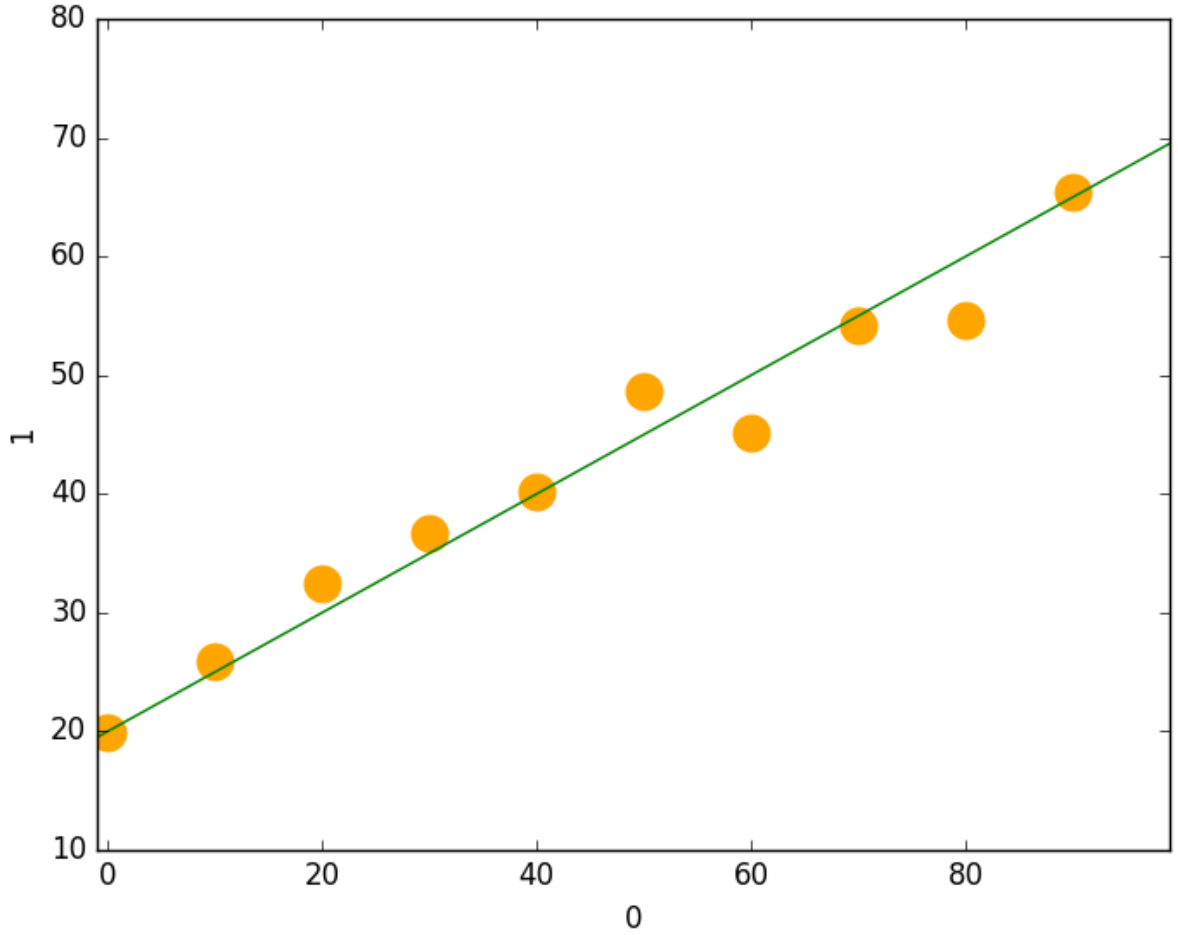
- 模型泛化：偏差（bias）与方差（variance）
  - 偏差：在训练数据上预测与实际标签的偏差
    - 预测的准确性
  - **方差**：不同训练数据集训练出来的模型，预测间的差异
    - 预测的稳定性

图片来自网络

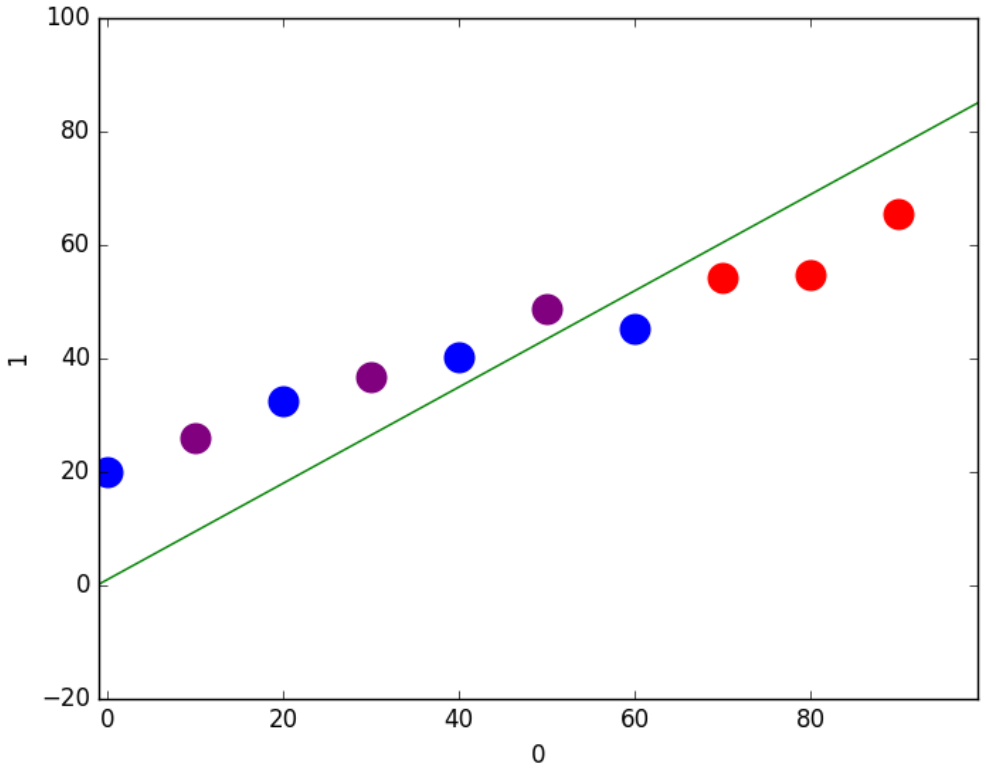


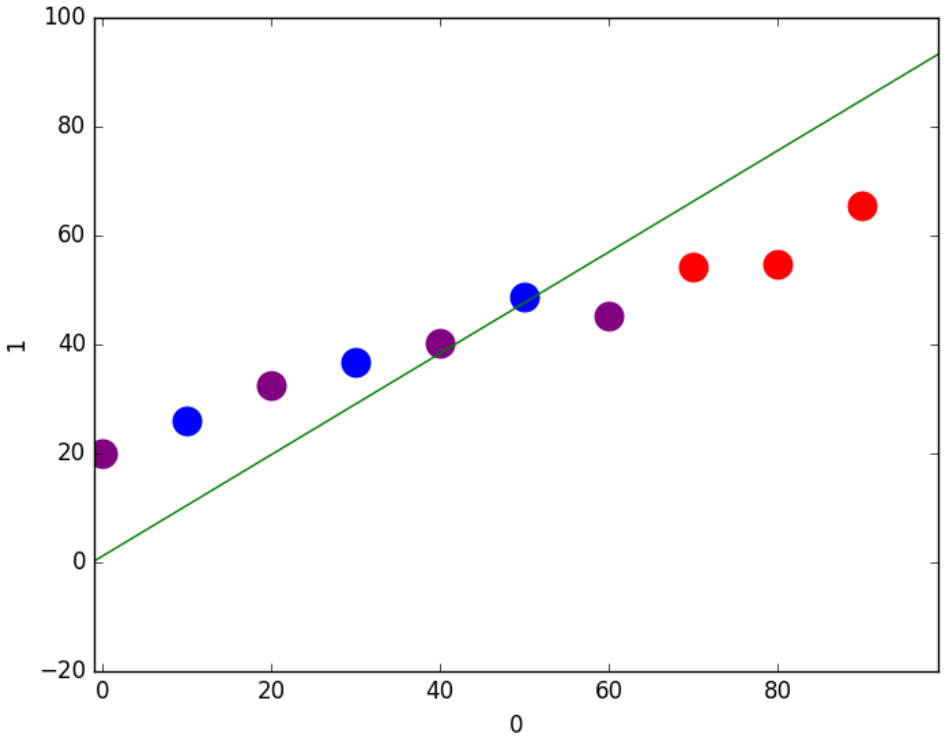


- $y = 0.5x + 20 + noise$

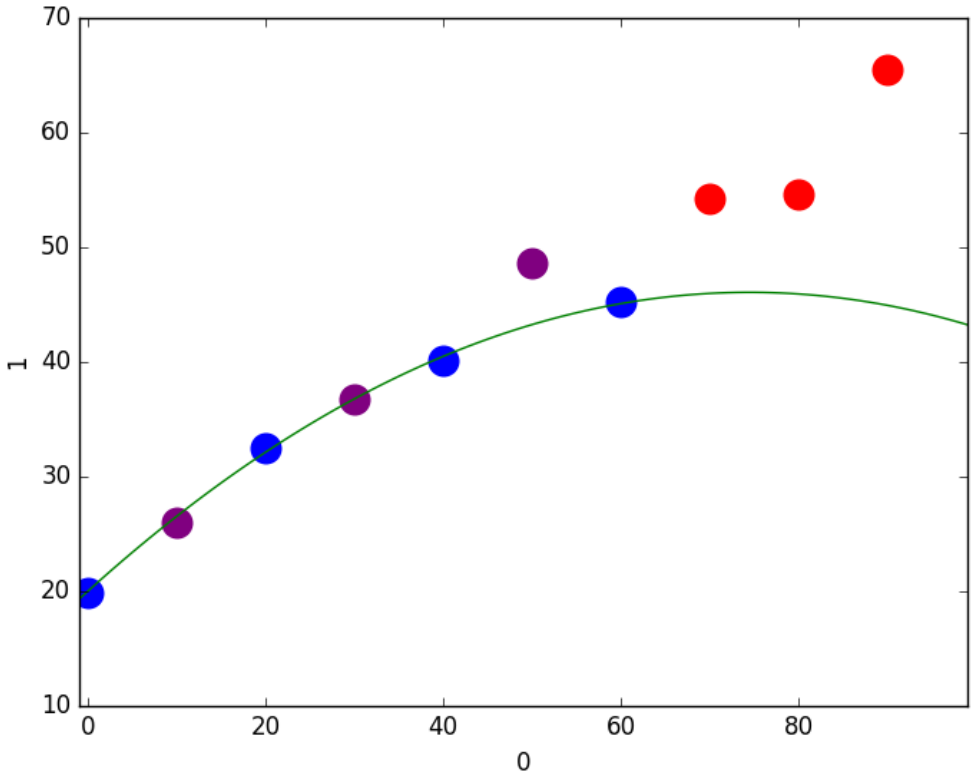


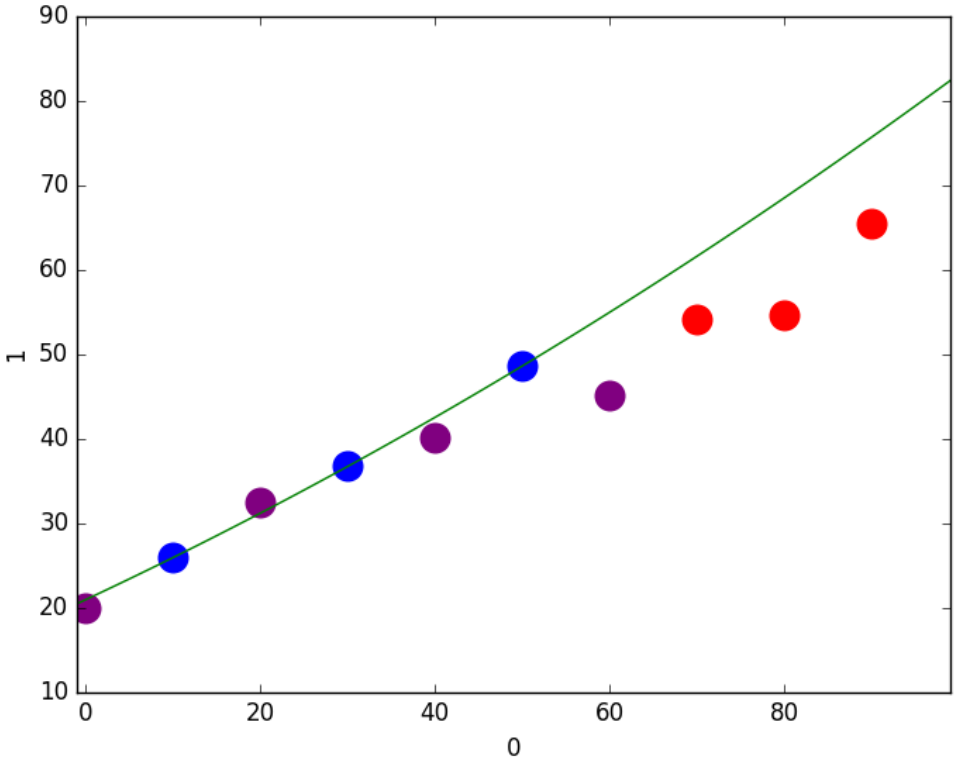
- 模型：$y = ax + b$
- 蓝点：训练数据 红点：测试数据





- 模型：$y = ax^2 + bx + c$
- 蓝点：训练数据 红点：测试数据





- 回归方法的评估（距离）
  - 预测第二天的股价
  - 进行100天预测
    - $\frac{1}{100}\sum|y-f(x)|$
    - $\frac{1}{100}\sum(\mathrm{y}-f(x))^2$
    - $\frac{1}{100}\sqrt{\sum(\mathrm{y}-f(x))^2}$

- 分类方法的评估
  - 预测第二天股价涨的概率

  - 预测100天
    - 方法一：(涨, 0.7), (没涨, 0.6), (涨, 0.2) …

    - 方法二：(涨, 0.9), (没涨, 0.3), (涨, 0.4) …

- 对方法一，设阈值0.65
- 对方法一，设阈值0.5
- 对方法一，设阈值0.1

真实情况	**预测结果**	
	正例（P）	反例（N)
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

- 查准率：$precision = \frac{TP}{TP+FP}$

- 查全率：$recall = \frac{TP}{TP+\mathrm{FN}}$

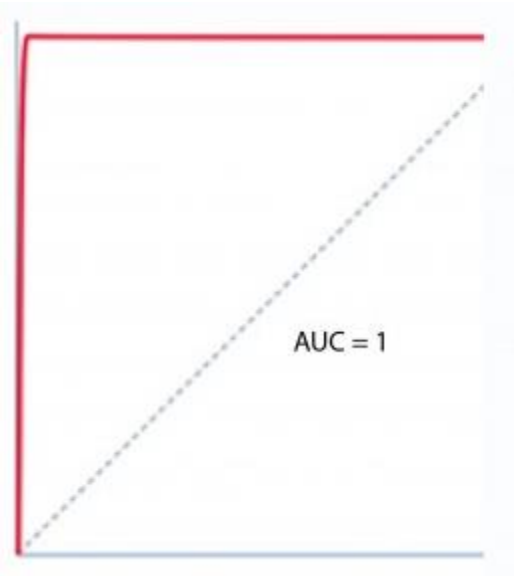
真实情况	预测结果	
	正例（P）	反例（N)
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

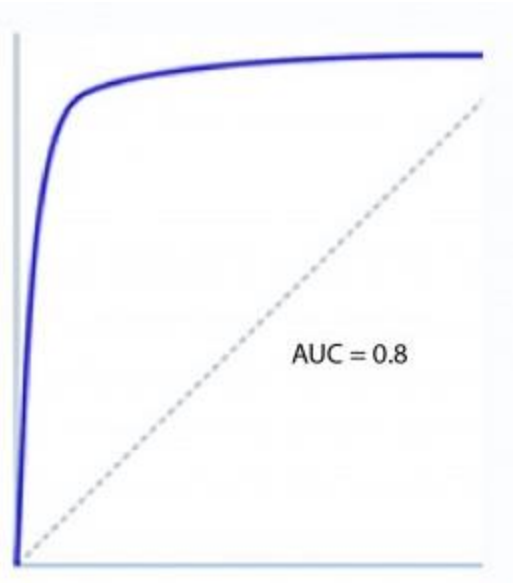
- Precision-recall曲线、auPR
  - 查准率：$precision = \frac{TP}{TP+FP}$
  - 查全率：$recall = \frac{TP}{TP+\mathrm{FN}}$

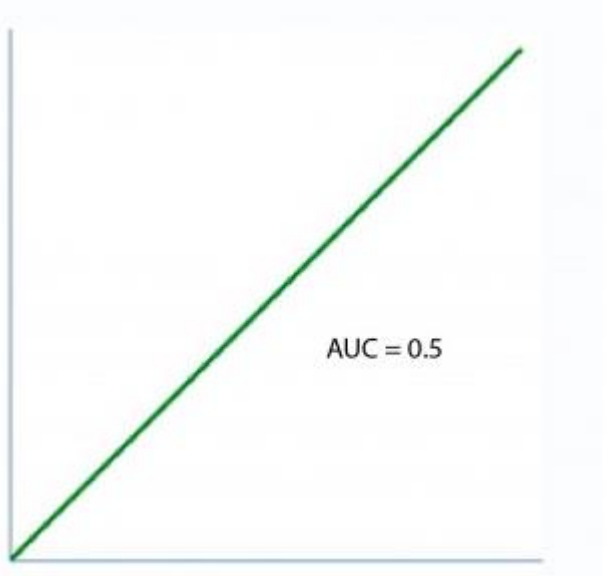
真实情况	预测结果	
	正例（P）	反例（N)
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

- Receiver Operating Characteristic曲线、AUC(area under ROC curve)
  - 真正例率：$\mathrm{T}PR = \frac{TP}{TP+F\mathrm{N}}$
  - 假正例率：$FPR = \frac{\mathrm{F}P}{\mathrm{FP}+TN}$

几种ROC曲线：
完美、有预测能力、无预测能力







- 代码（auPR、AUC）

```
from sklearn.metrics import roc_auc_score, average_precision_score

# labels [1, 0, 1]
#pred [0.4, 0.3, 0.2]

roc_auc_score(labels, pred)

average_precision_score(labels, pred)
```

- 好的模型
  - 学习快速、硬件要求低
  - 符合常识（domain knowledge）

  - 在测试数据上评价高

  - 可以拟合训练数据又可以泛化到测试数据
    - 模型复杂度适中（参数量、正规化）
    - 训练程度适中
      - 一般在训练数据集与测试数据集的评价相仿
      - 在测试数据集的表现不再继续上升

- 以上