* 信用评分卡

分类：申请A（Application Score Card）、行为B（Behavior Score Card）、催收C（Collection Score Card）

* 1. 问题的定义

客户群体分类的定义：风控策略（bad：逾期90天、good：12个月未出现逾期90天）、催收策略

* 1. 选取样本，好客户、坏客户、被拒绝的客户（good、bad、rejected）
  2. 数据抽取、清洗、整理
  3. 对好、坏客户样本进行数据分析（不包括被拒绝客户，因为他们没有score），通过分组，选择强预测变量
  4. 得到初始回归模型，对被拒绝的客户进行评分，得到其成为好、坏客户的概率，作为权重，将被拒绝样本放入DataSet中。重新对变量分组，回归
  5. 在第二次模型的基础上，将模型的结果（概率）转换成score，即最终评分卡
  6. 检验模型预测能力。交换曲线、K-S值（30%以上）、Gini指数、AR值
  7. 监测模型稳定性。比较新申请客户群体与以往的分值分布
  8. 监控特征分布
  9. 监控不良贷款。评估不同分数段的不良贷款
* 分类
  1. 二元分类：sigmoid。

真实=1，预测成功，接近1，L=-1

预测失败，接近0，L=0

预测成功的L<预测失败的L，合理。但不是凸函数，无法gd优化

* 1. 多元分类：softmax
* 回归
  1. OLS：
  2. Ridge：
  3. Lasso：
* 距离度量公式
  1. 闵可夫斯基：
  2. 曼哈顿：令p=1
  3. 欧几里得：令p=2
* 贝叶斯公式

先验概率：已经知道的，P(类别C1)

条件概率（似然度）：实验获得的，P(特征1|类别C1，特征2,3…)，需要学习的对象

要解决的实际问题（后验概率）：P(类别C1|特征1,2,3…)

朴素：特征与特征之间相互独立。

总结：朴素贝叶斯：通过似然度和先验概率预测后验概率。后验概率正比于似然度，提高似然度，等价于提高后验概率。通过比较不同特征与类之间的似然关系，把似然度最大的类作为预测结果。

可构成似然函数，通过调节参数来表示不同的似然度，用于训练模型

* 1. 10个女同学，身高、衣服相似，看背影是A的概率：先验概率P(A)=10%
  2. 班上女生有3中发型，统计得到扎马尾的概率：P(马尾)=30%
  3. A喜欢扎马尾，统计得到她扎马尾的概率：条件概率（似然度）P(马尾|A)=70%
  4. 后验概率P(A|扎马尾)

统计先验概率、似然度

对待测样本对不同类分别计算后验概率，最大的类为预测输出

* 决策树

ID3、C4.5、CART

都是为了度量纯度

* ID3：信息增益。分布越乱，熵越大。，当p=0/1时，信息熵最小=0、当P=0.5时，信息熵最大=1。

目的：选择某个特征进行划分，是为了把包含多种类别的集合尽可能划分成只包含一种类别的多个子集。

划分的好坏可以用划分前后的纯度来衡量。

但是，ID3会优先选择特征值比较多的feature，比如身份证号

* C4.5：使用信息增益比

从特征维度自身属性出发，确保特征值越多，其固有值（Intrinsic Value）越大，相应的增益比的效果越小

* CART：不使用信息熵，使用基尼指数

基尼指数和信息熵类似，越小越好

按a对D分类后的基尼指数：

* SVM

最大间隔、高维映射、核方法

目标：间隔最大

损失函数：间隔

点到超平面的距离：

假定支持向量到超平面的距离为d，其他点到超平面的距离>d

，

约束条件：

（支持向量满足）

间隔：支持向量到超平面距离

目标：最大化间隔

求解：用拉格朗日乘子法转化成拉格朗日函数，再分别对参数w和b求导，并令导数=0.再使用SMO算法求解。

Tips：

* 最大间隔：使用拉格朗日乘子法+SMO算法求解
* 转化式的末尾是两个向量的内积，正因如此才能使用核技巧进行高维计算

核方法（kernel method）：将非线性可分数据映射成高维线性可分

核技巧（kernel trick）：提高核方法的计算效率。

如果先进行高维映射，再进行高维向量点积操作，运算量很大

向量经核函数映射后转变成，求解最大间隔时向量内积变成

通过核技巧，原始向量可直接得到正确的点积结果

存在函数，向量内积可直接通过函数K求出