1. 名词解释

LR (Logistic Regression) 逻辑回归

GBDT简介 GBDT(Gradient Boosting Decision Tree) 是一种迭代的决策树算法,

CNN(卷积神经网络)、RNN(循环神经网络)、DNN(深度神经网络) fuzzy neural network 模糊神经网络;

Recommender 推荐系统

ReLu(Rectified Linear Units)激活函数 修正的线性单位

TanH 双曲正切函数

CTR（Click-Through-Rate）即点击通过率，是互联网广告常用的术语，指网络广告（图片广告/文字广告/关键词广告/排名广告/视频广告等）的点击到达率，即该广告的实际点击次数（严格的来说，可以是到达目标页面的数量）除以广告的展现量（Show content）。

one-hot 独热码

FTRL(Follow-the-regularized-Leader)

先验概率（prior probability）是指根据以往经验和分析得到的概率，如全概率公式，它往往作为"由因求果"问题中的"因"出现的概率·

先验概率是通过以往的经验对未知情况的预测。

倒排索引（inverse index） 常规的索引 文档ID –》关键字

倒排索引是 关键字 –》文档ID 可以方面通过关键字知道 在那些文章中出现过。

准确率与召回率（Precision & Recall）

训练数据集Training set is a subset of the dataset used to build predictive models.

验证数据集Validation set is a subset of the dataset used to assess the performance of model built in the training phase. It provides a test platform for fine tuning model's parameters and selecting the best-performing model. Not all modeling algorithms need a validation set.

测试数据集Test set or unseen examples is a subset of the dataset to assess the likely future performance of a model. If a model fit to the training set much better than it fits the test set, overfitting is probably the cause.

偏差：描述的是预测值（估计值）的期望与真实值之间的差距。偏差越大，越偏离真实数据，如下图第二行所示。

期望E：随机变量的平均值为该变量的数学期望

方差D：描述的是预测值的变化范围，离散程度，也就是离其期望值的距离。方差越大，数据的分布越分散。

方差 = 平方的期望 – 期望的平方

D(x) = E(x2) – ( E(x) )2

标准差σ(sigma)： 方差的算数平方根 作用与方差相同 描述 数据的离散程度。用σ表示

协方差Cov(x,y): 协方差表示的是两个变量总体误差的期望

http://c.hiphotos.baidu.com/baike/s%3D340/sign=6005955212dfa9ecf92e501352d1f754/359b033b5bb5c9ea8718f4add639b6003bf3b391.jpg

1. 笔记

2.1 准确率与召回率

准确率和召回率是广泛用于信息检索和统计学分类领域的两个度量值，用来评价结果的质量。其中精度是检索出相关文档数与检索出的文档总数的比率，衡量的是检索系统的查准率；召回率是指检索出的相关文档数和文档库中所有的相关文档数的比率，衡量的是检索系统的查全率。

一般来说，Precision就是检索出来的条目（比如：文档、网页等）有多少是准确的，Recall就是所有准确的条目有多少被检索出来了。

正确率、召回率和 F 值是在鱼龙混杂的环境中，选出目标的重要评价指标。不妨看看这些指标的定义先：

1. 正确率 = 提取出的正确信息条数 / 提取出的信息条数

2. 召回率 = 提取出的正确信息条数 / 样本中的信息条数

两者取值在0和1之间，数值越接近1，查准率或查全率就越高。

先验概率与后验概率的区别

先验概率不是根据有关自然状态的全部资料测定的，而只是利用现有的材料(主要是历史资料)计算的；[后验概率](http://baike.baidu.com/view/336754.htm)使用了有关自然状态更加全面的资料，既有先验概率资料，也有补充资料；

先验概率的计算比较简单，没有使用[贝叶斯公式](http://baike.baidu.com/view/541856.htm)；而后验概率的计算，要使用贝叶斯公式，而且在利用[样本](http://baike.baidu.com/view/315109.htm)资料计算逻辑概率时，还要使用理论[概率分布](http://baike.baidu.com/view/45323.htm)，需要更多的[数理统计](http://baike.baidu.com/view/21137.htm)知识。

2.2 评估预测函数

为了检验算法是否过拟合，我们将数据分成训练集和测试集，通常用 70%的数据作为训练集，用剩下 30%的数据作为测试集。很重要的一点是训练集和测试集均要含有各种类型的数据，通常我们要对数据进行“洗牌”，然后再分成训练集和测试集。

使用 60%的数据作为训练集，使用 20%的数据作为交叉验证集， 使用 20%的数据作为测试集

其中训练集用来估计模型，验证集用来确定网络结构或者控制模型复杂程度的参数，而测试集则检验最终选择最优的模型的性能如何。一个典型的划分是训练集占总样本的50%，而其它各占25%，三部分都是从样本中随机抽取。

高偏差/欠拟合的情况下，增加数据到训练集不一定能有帮助

2.3构建一个学习算法的推荐方法

1. 从一个简单的能快速实现的算法开始，实现该算法并用交叉验证集数据测试这个算法

2. 绘制学习曲线，决定是增加更多数据，或者添加更多特征，还是其他选择

3. 进行误差分析：人工检查交叉验证集中我们算法中产生预测误差的实例，看看这些实例是否有某种系统化的趋势

2.4 Support Vector Machine SVM 支持向量机

“支持向量”：所有在最大边界上行的点都叫支持向量，在后续的分类中这些点要参与预测。

LibSVM 输出信息

optimization finished, #iter = 162 迭代次数

nu = 0.431029 选择的核函数类型的参数

obj = -100.877288, rho = 0.424462 SVM文件转换为的二次规划求解得到的最小值

nSV = 132, nBSV = 107 nSV 为标准支持向量个数(0<a[i]<c)，nBSV为边界上的支持向量个数(a[i]=c)

Total nSV = 132 支持向量总个数

Model参数信息

svm\_type c\_svc //所选择的svm类型，默认为c\_svc

kernel\_type rbf //训练采用的核函数类型，此处为RBF核

gamma 0.0769231 //RBF核的参数γ

nr\_class 2 //类别数，此处为两分类问题

total\_sv 132 //支持向量总个数

rho 0.424462 //判决函数的偏置项b

label 1 -1 //原始文件中的类别标识

nr\_sv 64 68 //每个类的支持向量机的个数

SV //以下为各个类的权系数及相应的支持向量

1 1:0.166667 2:1 3:-0.333333 … 10:-0.903226 11:-1 12:-1 13:1

0.5104832128985164 1:0.125 2:1 3:0.333333 … 10:-0.806452 12:-0.333333 13:0.5

-1 1:-0.375 2:1 3:-0.333333…. 10:-1 11:-1 12:-1 13:1

-1 1:0.166667 2:1 3:1 …. 10:-0.870968 12:-1 13:0.5