2017年3月29日阿里巴巴数据分析师(北京)

一、10道填空，每题3分

1、小松鼠采到了100颗坚果要运回家。家离放坚果的地方有100米远。小松鼠每次最多运50颗。BUT!小松鼠很馋。。。每走2米就要吃一颗坚果。。。问小松鼠最多能运回家多少颗坚果?

A 0 B 10 C 25 D50

答案：应该是25颗吧!(先运50颗50米，吃了25颗，返回去，回去的途中没吃的了，再运50颗到50米的地方，又吃了25颗，再把剩下的运回家，又吃25颗，还剩25颗。)吐槽一下，题目应该说明：小松鼠足够聪明，至少比参加考试的人聪明。。。

2、标号12345的5个球，一次取两个，和为3或者6的概率是多少? 答案：0.3。不解释。

3、考了logistic回归

# [逻辑回归模型(Logistic Regression, LR)基础](http://www.cnblogs.com/sparkwen/p/3441197.html)

逻辑回归(Logistic Regression, LR)模型其实仅在线性回归的基础上，套用了一个逻辑函数，但也就由于这个逻辑函数，使得逻辑回归模型成为了机器学习领域一颗耀眼的明星，更是计算广告学的核心。本文主要详述逻辑回归模型的基础，至于逻辑回归模型的优化、逻辑回归与计算广告学等，请关注后续文章。

#### 1 逻辑回归模型

    回归是一种极易理解的模型，就相当于y=f(x)，表明自变量x与因变量y的关系。最常见问题有如医生治病时的望、闻、问、切，之后判定病人是否生病或生了什么病，其中的望闻问切就是获取自变量x，即特征数据，判断是否生病就相当于获取因变量y，即预测分类。

    最简单的回归是线性回归，在此借用Andrew NG的讲义，有如图1.a所示，X为数据点——肿瘤的大小，Y为观测值——是否是恶性肿瘤。通过构建线性回归模型，如hθ(x)所示，构建线性回归模型后，即可以根据肿瘤大小，预测是否为恶性肿瘤hθ(x)≥.05为恶性，hθ(x)<0.5为良性。

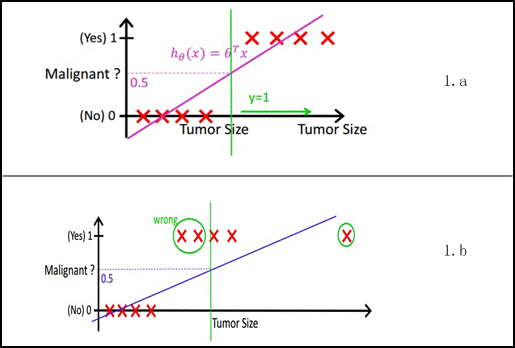
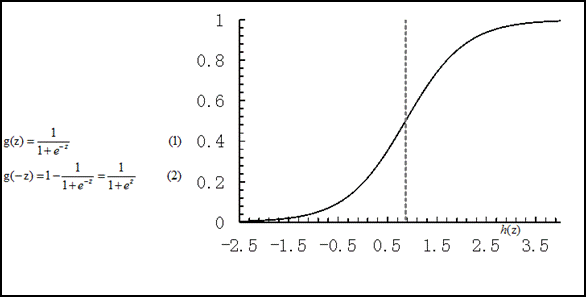
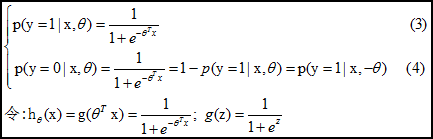
[](file:///C:\Users\wensaiping\AppData\Local\Temp\WindowsLiveWriter1286139640\supfilesCE8402E\image8.png)

图1 线性回归示例

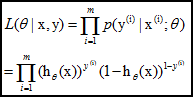
    然而线性回归的鲁棒性很差，例如在图1.b的数据集上建立回归，因最右边噪点的存在，使回归模型在训练集上表现都很差。这主要是由于线性回归在整个实数域内敏感度一致，而分类范围，需要在[0,1]。逻辑回归就是一种减小预测范围，将预测值限定为[0,1]间的一种回归模型，其回归方程与回归曲线如图2所示。逻辑曲线在z=0时，十分敏感，在z>>0或z<<0处，都不敏感，将预测值限定为(0,1)。

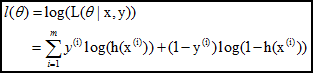
[](file:///C:\Users\wensaiping\AppData\Local\Temp\WindowsLiveWriter1286139640\supfilesCE8402E\image28.png)图2 逻辑方程与逻辑曲线

    逻辑回归其实仅为在线性回归的基础上，套用了一个逻辑函数，但也就由于这个逻辑函数，逻辑回归成为了机器学习领域一颗耀眼的明星，更是计算广告学的核心。对于多元逻辑回归，可用如下公式似合分类，其中公式(4)的变换，将在逻辑回归模型参数估计时，化简公式带来很多益处，y={0,1}为分类结果。

    对于训练数据集，特征数据x={x1, x2, … , xm}和对应的分类数据y={y1, y2, … , ym}。构建逻辑回归模型f(θ)，最典型的构建方法便是应用[极大似然估计](http://www.cnblogs.com/sparkwen/p/3199728.html)。首先，对于单个样本，其后验概率为：

[clip_image008](file:///C:\Users\wensaiping\AppData\Local\Temp\WindowsLiveWriter1286139640\supfilesCE8402E\image141.png)    那么，极大似然函数为：

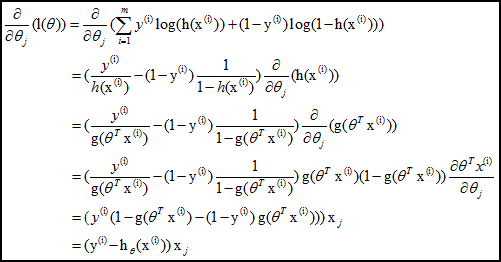
[](file:///C:\Users\wensaiping\AppData\Local\Temp\WindowsLiveWriter1286139640\supfilesCE8402E\image20.png)    log似然是：

[](file:///C:\Users\wensaiping\AppData\Local\Temp\WindowsLiveWriter1286139640\supfilesCE8402E\image23.png)

#### 2 梯度下降

    由第1节可知，求逻辑回归模型f(θ)，等价于：

[clip_image014](file:///C:\Users\wensaiping\AppData\Local\Temp\WindowsLiveWriter1286139640\supfilesCE8402E\image31.png)    采用梯度下降法：

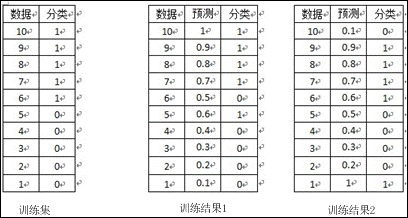
[](file:///C:\Users\wensaiping\AppData\Local\Temp\WindowsLiveWriter1286139640\supfilesCE8402E\image37.png)     从而迭代θ至收敛即可：

[clip_image018](file:///C:\Users\wensaiping\AppData\Local\Temp\WindowsLiveWriter1286139640\supfilesCE8402E\image40.png)

#### 3 模型评估

    对于LR分类模型的评估，常用AUC来评估，关于AUC的更多定义与介绍，可见参考文献2，在此只介绍一种极简单的计算与理解方法。

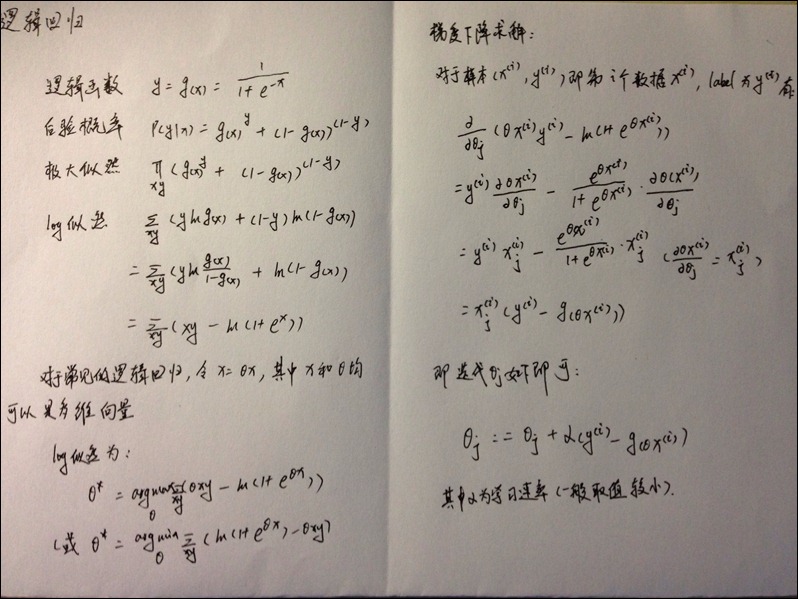
    对于下图的分类：

[](file:///C:\Users\wensaiping\AppData\Local\Temp\WindowsLiveWriter1286139640\supfilesCE8402E\image14.png)     对于训练集的分类，训练方法1和训练方法2分类正确率都为80%，但明显可以感觉到训练方法1要比训练方法2好。因为训练方法1中，5和6两数据分类错误，但这两个数据位于分类面附近，而训练方法2中，将10和1两个数据分类错误，但这两个数据均离分类面较远。

    AUC正是衡量分类正确度的方法，将训练集中的label看两类{0，1}的分类问题，分类目标是将预测结果尽量将两者分开。将每个0和1看成一个pair关系，团中的训练集共有5\*5=25个pair关系，只有将所有pair关系一至时，分类结果才是最好的，而auc为1。在训练方法1中，与10相关的pair关系完全正确，同样9、8、7的pair关系也完全正确，但对于6，其pair关系(6，5)关系错误，而与4、3、2、1的关系正确，故其auc为(25-1)/25=0.96；对于分类方法2，其6、7、8、9的pair关系，均有一个错误，即(6,1)、(7,1)、(8,1)、(9,1)，对于数据点10，其正任何数据点的pair关系，都错误，即(10,1)、(10,2)、(10,3)、(10,4)、(10,5)，故方法2的auc为(25-4-5)/25=0.64，因而正如直观所见，分类方法1要优于分类方法2。

#### 4 演算手稿

    附演算手稿如下：

[](https://images0.cnblogs.com/blog/540980/201312/04223813-79436d93574e452a9107aca513eba010.jpg)

4、聚类分析法，k\_means。

5、其他条件相同，置信水平越低，则置信区间上下限差值越() A.越大 B.越小 C.为0 D不确定 应该是B吧。

6、precision、ecall、ROC。 剩下的不记得了。

二、三道题，每题10分。

1、已知每10万人中有1人得艾滋病。现在有一种检查，如果被测者患病则一定能查出来。如果被测者没病，有1%的测试出错也显示阳性。现在一个人检查结果是阳性。问真正得病的概率?

答案：貌似所有的讲全概率公式的书上都有这道题。

2、SQL

两张表合并，主键是USER\_ID，然后把深圳市、广州市，大于16岁的，发生在2013年12月的一项挑出来加起来。不会SQL的话就写思路。我就不会。。。。。。

3、层次分析法AHP的含义，具体步骤。并举一个适合用层次分析法的案例。

三、两道题，每题20分。

1、淘宝上有一些小众但是品位高的店铺，怎么把他们筛选出来捏?

2、双十一的时候，商家会发优惠券，从商家和平台两个角度设计评价指标，并分析指标不同结果应该对应怎样的措施。

2011年阿里巴巴数据分析实习生笔试题

[**从阿里巴巴笔试试题看数据分析师的职业要求**](http://dataunion.org/13795.html)

以下试题是来自阿里巴巴招募实习生的2011年笔试题，从笔试题的几个要求我们一起来看看数据分析的职业要求。

一、异常值是指什么？请列举1种识别连续型变量异常值的方法？

异常值（Outlier） 是指样本中的个别值，其数值明显偏离所属样本的其余观测值。在数理统计里一般是指一组观测值中与平均值的偏差超过两倍标准差的测定值。

Grubbs’ test（是以Frank E. Grubbs命名的），又叫maximum normed residual test，是一种用于单变量数据集异常值识别的统计检测，它假定数据集来自正态分布的总体。

未知总体标准差σ，在五种检验法中，优劣次序为：t检验法、格拉布斯检验法、峰度检验法、狄克逊检验法、偏度检验法。

点评：考察的内容是统计学基础功底。

二、什么是聚类分析？聚类算法有哪几种？请选择一种详细描述其计算原理和步骤。

聚类分析(cluster analysis)是一组将研究对象分为相对同质的群组(clusters)的统计分析技术。 聚类分析也叫分类分析(classification analysis)或数值分类(numerical taxonomy)。聚类与分类的不同在于，聚类所要求划分的类是未知的。

聚类分析计算方法主要有： 层次的方法（hierarchical method）、划分方法（partitioning method）、基于密度的方法（density-based method）、基于网格的方法（grid-based method）、基于模型的方法（model-based method）等。其中，前两种算法是利用统计学定义的距离进行度量。

k-means 算法的工作过程说明如下：首先从n个数据对象任意选择 k 个对象作为初始聚类中心；而对于所剩下其它对象，则根据它们与这些聚类中心的相似度（距离），分别将它们分配给与其最相似的（聚类中心所代表的）聚类；然 后再计算每个所获新聚类的聚类中心（该聚类中所有对象的均值）；不断重复这一过程直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数. k个聚类具有以下特点：各聚类本身尽可能的紧凑，而各聚类之间尽可能的分开。

其流程如下：

（1）从 n个数据对象任意选择 k 个对象作为初始聚类中心；

（2）根据每个聚类对象的均值（中心对象），计算每个对象与这些中心对象的距离；并根据最小距离重新对相应对象进行划分；

（3）重新计算每个（有变化）聚类的均值（中心对象）；

（4）循环（2）、（3）直到每个聚类不再发生变化为止（标准测量函数收敛）。

优点：本算法确定的K 个划分到达平方误差最小。当聚类是密集的，且类与类之间区别明显时，效果较好。对于处理大数据集，这个算法是相对可伸缩和高效的，计算的复杂度为 O(NKt)，其中N是数据对象的数目，t是迭代的次数。一般来说，K<<N，t<<N 。

缺点：1. K 是事先给定的，但非常难以选定；2. 初始聚类中心的选择对聚类结果有较大的影响。

点评：考察的内容是常用数据分析方法，做数据分析一定要理解数据分析算法、应用场景、使用过程、以及优缺点。

三、根据要求写出SQL

表A结构如下：

Member\_ID（用户的ID，字符型）

Log\_time（用户访问页面时间，日期型（只有一天的数据））

URL（访问的页面地址，字符型）

要求：提取出每个用户访问的第一个URL（按时间最早），形成一个新表（新表名为B，表结构和表A一致）

createtable B as select Member\_ID, min(Log\_time), URL from A group by Member\_ID ;

点评：SQL语句，简单的数据获取能力，包括表查询、关联、汇总、函数等。

另外，这个答案其实是不对的，实现有很多方法，任由大家去发挥吧。

四、销售数据分析

以下是一家B2C电子商务网站的一周销售数据，该网站主要用户群是办公室女性，销售额主要集中在5种产品上，如果你是这家公司的分析师，

a) 从数据中，你看到了什么问题？你觉得背后的原因是什么？

b) 如果你的老板要求你提出一个运营改进计划，你会怎么做？

表如下：一组每天某网站的销售数据

[20150402105837159](http://dataunion.org/wp-content/uploads/2015/04/20150402105837159.jpg)

a) 从这一周的数据可以看出，周末的销售额明显偏低。这其中的原因，可以从两个角度来看：站在消费者的角度，周末可能不用上班，因而也没有购买该产品的欲望；站在产品的角度来看，该产品不能在周末的时候引起消费者足够的注意力。

b) 针对该问题背后的两方面原因，我的运营改进计划也分两方面：一是，针对消费者周末没有购买欲望的心理，进行引导提醒消费者周末就应该准备好该产品；二是，通过该产品的一些类似于打折促销等活动来提升该产品在周末的人气和购买力。

点评：数据解读能力，获取数据是基本功，仅仅有数据获取能力是不够的，其次是对数据的解读能力。

五、用户调研

某公司针对A、B、C三类客户，提出了一种统一的改进计划，用于提升客户的周消费次数，需要你来制定一个事前试验方案，来支持决策，请你思考下列问题：

a) 试验需要为决策提供什么样的信息？

c) 按照上述目的，请写出你的数据抽样方法、需要采集的数据指标项，以及你选择的统计方法。

a) 试验要能证明该改进计划能显著提升A、B、C三类客户的周消费次数。

b) 根据三类客户的数量，采用分层比例抽样；

需要采集的数据指标项有：客户类别，改进计划前周消费次数，改进计划后周消费次数；

选用统计方法为：分别针对A、B、C三类客户，进行改进前和后的周消费次数的，两独立样本T-检验（two-sample t-test）。

点评：业务理解能力和数据分析思路，这是数据分析的核心竞争力。

综上所述：一个合格的数据分析应该具备统计学基础知识、数据分析方法、数据获取、数据解读和业务理解、数据分析思想几个方面能力，即将成为数据分析师的亲们，你们准备好了吗？

【作者简介】

郑来轶，中国统计网创始人，江湖人称“数据哥”，微博/微信/今日头条大V帐号@数据分析精选，国内某知名互联网公司无线数据分析专家，前中国移动阅读基地资深数据分析师、每日轩昂BI总监，多年来深耕移动互联网行业的数据体系建设与数据分析应用。

文章出处：http://www.itongji.cn/article/0401435R015.html

2018.5.8 阿里巴巴数据分析模拟笔试

3、下面哪些模型是生成式模型？

A 判别式分析 B 朴素贝叶斯 C 神经网络 D SVM E 以上都是 F 以上都不是

4、如何评估天猫商家成长能力，可以用哪些指标、哪些算法？

5、一个学校有7个年级，每个年级有10个班，每个班有20名学生，如何用SQL随机抽取10%学生参加讲座？