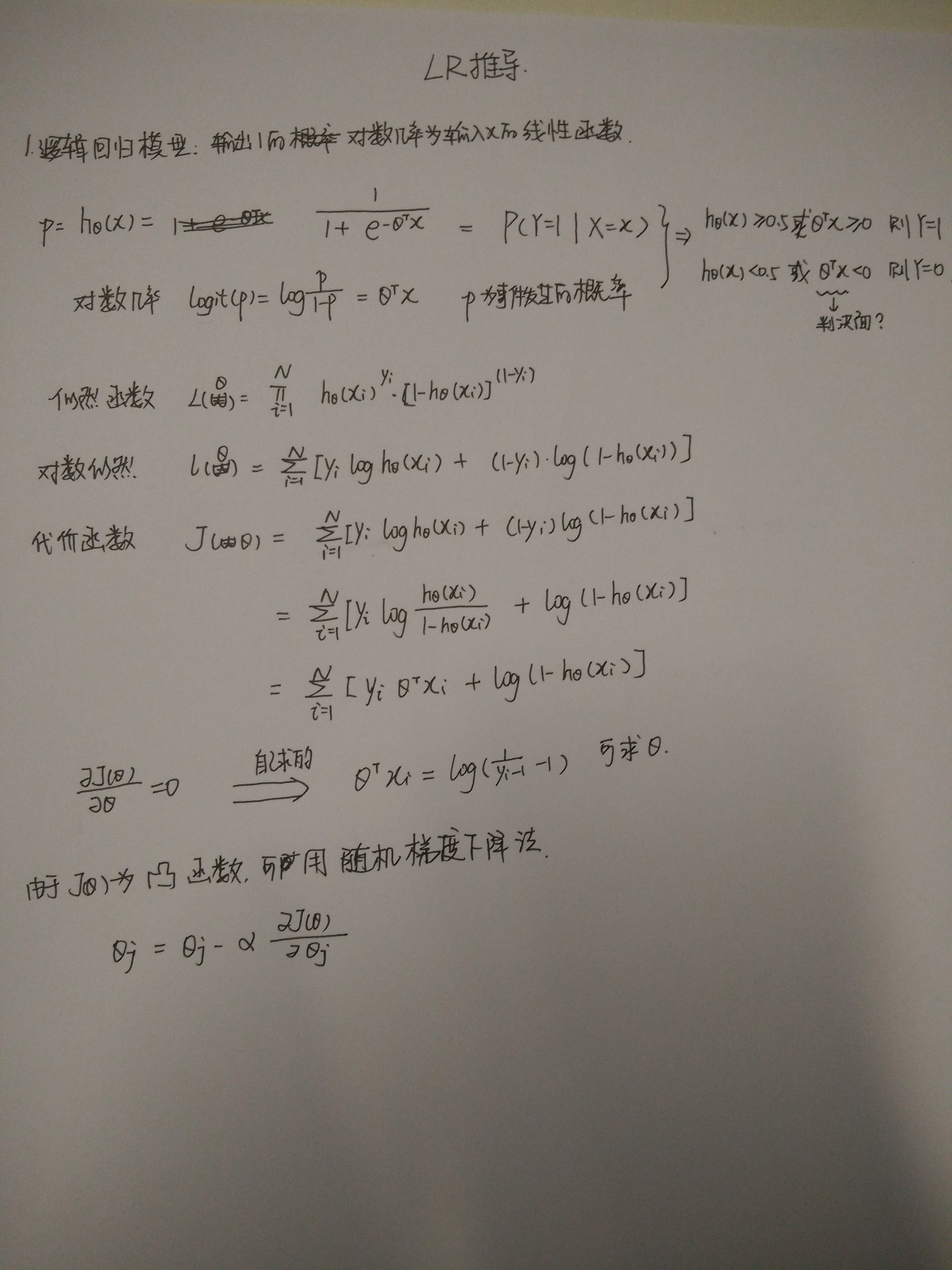
**LR**

****

**问题与答案**

1. **我知道校正R2或者F值是用来评估线性回归模型的。那用什么来评估逻辑回归模型？**

由于逻辑回归是用来预测概率的，我们可以用AUC-ROC曲线以及混淆矩阵来确定其性能。

此外，在逻辑回归中类似于校正R2的指标是AIC。AIC是对模型系数数量惩罚模型的拟合度量。因此，我们更偏爱有最小AIC的模型。

空偏差指的是只有截距项的模型预测的响应。数值越低，模型越好。残余偏差表示由添加自变量的模型预测的响应。数值越低，模型越好。

1. **广义线性模型是怎被应用在深度学习中?**

 深度学习从统计学角度，可以看做**递归的广义线性模型**。

 广义线性模型相对于经典的线性模型(y=wx+b)，核心在于引入了连接函数g(.)，形式变为：y=g−1(wx+b)。

 深度学习时递归的广义线性模型，神经元的激活函数，即为广义线性模型的链接函数。逻辑回归（广义线性模型的一种）的Logistic函数即为神经元激活函数中的Sigmoid函数。

1. **logistic的核心公式以及理解；**
2. **写出LR的伪代码；**
3. **比较一下SVM和LR各有哪些优缺点？**

见第10题。

1. **对数几率回归（logistics regression）和一般回归分析有什么区别？**

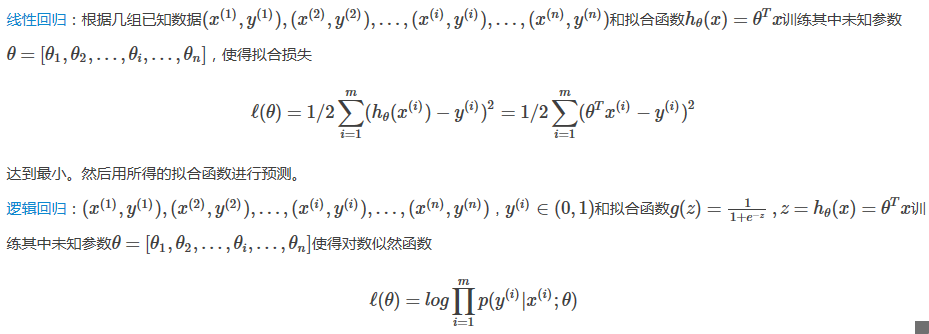
LR：对数几率回归/逻辑回归。

**Logistic**回归目的是从特征学习出一个0/1分类模型，而这个模型是将特性的线性组合作为自变量，由于自变量的取值范围是负无穷到正无穷。因此，使用logistic函数（或称作sigmoid函数）将自变量映射到(0,1)上，映射后的值被认为是属于y=1的概率

http://img.my.csdn.net/uploads/201304/05/1365174192_8325.png

当我们要判别一个新来的特征属于哪个类时，只需求http://img.my.csdn.net/uploads/201304/05/1365175136_8232.png即可，若http://img.my.csdn.net/uploads/201304/05/1365175136_8232.png大于0.5就是y=1的类，反之属于y=0类。

**概念**: 面对一个回归或者分类问题，建立代价函数，然后通过优化方法迭代求解出最优的模型参数，然后测试验证我们这个求解的模型的好坏.





<http://blog.csdn.net/viewcode/article/details/8794401>

1. **写出LR的目标函数和损失函数**

损失函数/目标函数区别与联系：

有人理解是：损失函数/代价函数是对模型或者算法的准确性的最直接的描述（对于二分类问题来说就是能不能正确分类的能力），而目标函数则是损失函数/代价函数的递进，目标函数考虑的可能不仅仅有模型算法的准确性，还可能要考虑模型的复杂度，可解释度等等的问题（比如目标函数在损失函数的基础之上加上各种正则项）。当然最简单的情况就是目标函数和损失函数等价。（例如：目标函数=损失项+正则项）

也有人理解是：没什么不同

<https://www.zhihu.com/question/52398145>

1. **如何判断自己实现的 LR、Kmeans 算法是否正确？**

使用平均误差准则函数E作为聚类结果好坏的衡量标准之一。（没找到）

1. **线性回归对于数据的假设是怎样的**（LR一般指逻辑回归，区别线性回归）

在谈线性回归模型的时候被问到，在线性回归中，有三个假设，是哪三个？  
当时回答出自变量x和因变量y间是线性关系。

也答道了样本点之间要求是独立同分布的（根据MLE准则，假定对样本加上高斯白噪声e的情况下）。  
但是第三个最终还是没有答上来。  
LR的wiki页面（http://en.wikipedia.org/wiki/Linear\_regression）中，有提到了LR的假设，分别是：  
Weak exogeneity：弱外生性。看意思是说假设用来预测的自变量x是没有测量误差的。这一假设当然是不现实的，不过如果没有这个假设的话，模型的复杂度会大大增加。  
Linearity：线性。因变量y是多个自变量x之间的线性组合。  
Constant variance (aka homoscedasticity)：同方差性。意思是说不同的因变量x的方差都是相同的。  
Independence of errors：误差独立性。即是变量之间是独立的

Lack of multicollinearity in the predictors：预测变量之中没有多重共线性。多重共线性我的理解是说，没有特征之间的交叉或者多次重复某些互相相关的特征。

感觉这几个假设中，我提到了线性假设（线性假设应该也算是包含了没有多重共线性吧，个人猜测），独立同分布假设中应该算是包含了同方差性、误差独立性。所以，可能没提到的那个指的是弱外生性，也就是假设样本点的取得是没有测量误差的，这一个假设一般现实中都不满足，但是都会假设满足来降低模型复杂度。

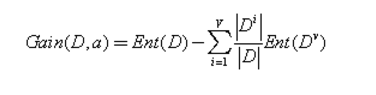
<http://blog.csdn.net/arthur503/article/details/24865405>

1. LR 的推导**，特性？决策树的特性？SVM、LR、决策树的对比？**

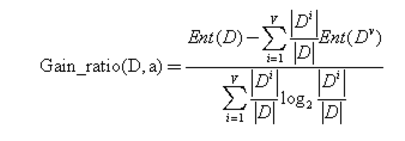
**LR优缺点**：实现简单，计算量非常小，速度很快，存储资源低，缺点就是因为模型简单，对于复杂的情况下会出现欠拟合。

**决策树特性：**

决策树基于树结构进行决策。特点在于需要选择一个属性进行分支，在分支的过程中选择信息增益最大的属性，定义如下



在划分中希望决策树的分支节点所包含的样本属于同一类别，即节点的纯度越来越高。决策树计算量简单，可解释性强，比较适合处理有缺失属性值的样本，能够处理不相关的特征，但容易过拟合，需要使用剪枝或者随机森林。信息增益是熵减去条件熵，代表信息不确定性较少的程度，信息增益越大，说明不确定性降低的越大，因此说明该特征对分类来说很重要。由于信息增益准则会对数目较多的属性有所偏好，因此一般用信息增益率(c4.5)



其中分母可以看作为属性自身的熵。取值可能性越多，属性的熵越大。

Cart决策树使用基尼指数来选择划分属性，直观的来说，Gini(D)反映了从数据集D中随机抽取两个样本，其类别标记不一致的概率，因此基尼指数越小数据集D的纯度越高，一般为了防止过拟合要进行剪枝，有预剪枝和后剪枝，一般用cross validation集进行剪枝。

连续值和缺失值的处理，对于连续属性a,将a在D上出现的不同的取值进行排序，基于划分点t将D分为两个子集。一般对每一个连续的两个取值的中点作为划分点，然后根据信息增益选择最大的。与离散属性不同，若当前节点划分属性为连续属性，该属性还可以作为其后代的划分属性。

**SVM、LR、决策树的对比**

**SVM：**可分类问题（线性、非线性），也可回归问题，可通过核函数快速的计算。软间隔损失函数为hinge loss，对outlier不敏感。小数据

LR：实现简单，训练速度快，但模型较简单。损失函数为带L2正则化项的cross entropy loss（adaboost对应的是exponential loss），对远点敏感。大数据

**决策树**：易过拟合，需要进行剪枝等。

因为只关心support vector，SVM可以将特征映射到无穷维空间，但是LR不可以。一般小数据中SVM比LR更优一点，但是LR可以预测概率，而SVM不可以。SVM依赖于数据测度，需要先做归一化，LR一般不需要，对于大量的数据LR使用更加广泛，LR向多分类的扩展更加直接，对于类别不平衡SVM一般用权重解决，即目标函数中对正负样本代价函数不同，LR可以用一般的方法，也可以直接对最后结果调整(通过阈值)，一般小数据下样本维度比较高的时候SVM效果要更优一些。

<http://www.cnblogs.com/zuochongyan/p/5407053.html>

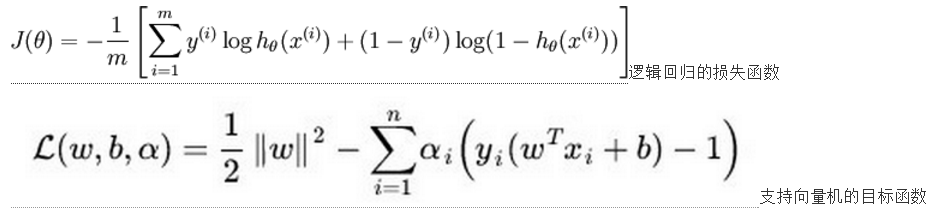
**LR与SVM比较**：

**相同点：**

* 都是分类算法（svm也可以回归）；
* 如果不考虑核函数，都是线性分类，分类决策面都是线性的（决策树是非线性的）；
* 都是监督学习方法，都是判别模型。

**不同点**：

* Loss函数不同则原理不同，LR是基于概率理论（假设样本为1的概率用sigmoid函数表示），SVM是几何间隔最大化的原理；



* SVM只考虑局部边界附近的点（支持向量），LR考虑全局（远离边界的点也起作用）；

解决非线性问题时，SVM通常用核函数（只计算几个少量支持向量的样本参与核函数计算），LR通常不采用核函数（如果运用核函数，每个样本都需要核计算，计算复杂）；

* SVM依赖数据表达的距离测度，所以要先对数据进行normalization，LR不受影响；
* SVM损失函数自带正则项（ ），这就是SVM是结构风险最小化算法的原因，LR必须在损失函数上添加正则项。

<http://www.cnblogs.com/zhizhan/p/5038747.html>