## 第三章 线性回归

3.1 试分析在什么情况下，式不必考虑偏置项b。

参考答案：在线性回归中，所有参数的确定都是为了让残差项的均值为0且残差项的平方和最小。在所有其他参数项确定后，偏置项b（或者说是常数项）的变化体现出来的就是拟合曲线的上下整体浮动，可以看做是其他各个解释变量留下的bias的线性修正。因此在线性拟合过程中是需要考虑偏置项的。

但若需要做的是比较不同自变量对因变量的影响，那么不需要考虑常数项，这样得到的回归系数是标准化的回归系数。

3.2 试证明，对于参数w，对率回归（logistics回归）的目标函数是非凸的，但其对数似然函数是凸的。

参考答案：如果一个多元函数是凸的，那么它的Hessian矩阵是半正定的。

, ,

合同于单位矩阵，所以是半正定矩阵。的值域为,当时， ,导致半负定，所以 是非凸的。

同理 , , 所以是凸的。

3.3 EOOC编码能起到理想纠错作用的重要条件是：在每一位编码上出错的概率相当且独立。试析多分类任务经ECOC编码后产生的二类分类器满足该条件的可能性及由此产生的影响。

参考答案：理论上的ECOC码能理想纠错的重要条件是每个码位出错的概率相当，因为如果某个码位的错误率很高，会导致这位始终保持相同的结果，不再有分类作用，这就相当于全0或者全 1的分类器，这点和NFL的前提很像。但由于事实的样本并不一定满足这些条件，所以书中提到了有多种问题依赖的ECOC被提出。

3.4 使用OvR和MvM将多分类任务分解为二分类任务求解时，试述为何无需专门针对类别不平衡性进行处理。

参考答案：对于OvR，MvM来说，由于对每个类进行了相同的处理，其拆解出的二分类任务中类别不平衡的影响会相互抵消，因此通常不需要专门处理。以ECOC编码为例，每个生成的二分类器会将所有样本分成较为均衡的二类，使类别不平衡的影响减小。当然拆解后仍然可能出现明显的类别不平衡现象，比如一个超级大类和一群小类。

3.5 LR与线性回归的区别与联系。

  参考答案：逻辑回归和线性回归首先都是广义的线性回归，其次经典线性模型的优化目标函数是最小二乘，而逻辑回归则是似然函数，另外线性回归在整个实数域范围内进行预测，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，因而对于这类问题来说，逻辑回归的鲁棒性比线性回归的要好。

  逻辑回归的模型本质上是一个线性回归模型，逻辑回归都是以线性回归为理论支持的。但线性回归模型无法做到sigmoid的非线性形式，sigmoid可以轻松处理0/1分类问题

3.6 线性分类器与非线性分类器的区别以及优劣

  参考答案：如果模型是参数的线性函数，并且存在线性分类面，那么就是线性分类器，否则不是。

  常见的线性分类器有：LR,贝叶斯分类，单层感知机、线性回归

  常见的非线性分类器：决策树、RF、GBDT、多层感知机

  SVM两种都有(看线性核还是高斯核)

  线性分类器速度快、编程方便，但是可能拟合效果不会很好

  非线性分类器编程复杂，但是效果拟合能力强

3.7 梯度下降法找到的一定是下降最快的方向么？

  参考答案：梯度下降法并不是下降最快的方向，它只是目标函数在当前的点的切平面（当然高维问题不能叫平面）上下降最快的方向。

3.8 在分类问题中,我们经常会遇到正负样本数据量不等的情况,比如正样本为10w条数据,负样本只有1w条数据,以下最合适的处理方法是()

　　A.将负样本重复10次,生成10w样本量,打乱顺序参与分类

　　B.直接进行分类,可以最大限度利用数据

　　C.从10w正样本中随机抽取1w参与分类

　　D.将负样本每个权重设置为10,正样本权重为1,参与训练过程

　　参考答案: ACD

解析：

(1) 重采样。 A可视作重采样的变形。改变数据分布消除不平衡，可能导致过拟合。

　　(2)欠采样。 C的方案 提高少数类的分类性能，可能丢失多数类的重要信息。

　　如果1：10算是均匀的话，可以将多数类分割成为1000份。然后将每一份跟少数类的样本组合进行训练得到分类器。而后将这1000个分类器用assemble的方法组合位一个分类器。A选项可以看作此方式，因而相对比较合理。

　　另：如果目标是 预测的分布 跟训练的分布一致，那就加大对分布不一致的惩罚系数。

　　(3)权值调整。 D方案也是其中一种方式。

当然，这只是在数据集上进行相应的处理，在算法上也有相应的处理方法。

3.9 特征比数据量还大时，选择什么样的分类器？

参考答案：线性分类器，因为维度高的时候，数据一般在维度空间里面会比较稀疏，很有可能线性可分。

3.10 关于线性回归的描述,以下正确的有:

　　A.基本假设包括随机干扰项是均值为0,方差为1的标准正态分布

　　B.基本假设包括随机干扰下是均值为0的同方差正态分布

　　C.在违背基本假设时,普通最小二乘法估计量不再是最佳线性无偏估计量

　　D.在违背基本假设时,模型不再可以估计

　　E.可以用DW检验残差是否存在序列相关性

　　F.多重共线性会使得参数估计值方差减小

参考答案：ACEF

解析1、AB一元线性回归的基本假设有：

　　（1）随机误差项是一个期望值或平均值为0的随机变量；

　　（2）对于解释变量的所有观测值，随机误差项有相同的方差；

　　（3）随机误差项彼此不相关；

　　（4）解释变量是确定性变量，不是随机变量，与随机误差项彼此之间相互独立；

　　（5）解释变量之间不存在精确的（完全的）线性关系，即解释变量的样本观测值矩阵是满秩矩阵；

　　（6）随机误差项服从正态分布

2、CD 违背基本假设的计量经济学模型还是可以估计的，只是不能使用普通最小二乘法进行估计。

　　当存在异方差时，普通最小二乘法估计存在以下问题： 参数估计值虽然是无偏的，但不是最小方差线性无偏估计。

3、E杜宾-瓦特森（DW）检验，计量经济，统计分析中常用的一种检验序列一阶 自相关 最常用的方法。

4、F所谓多重共线性（Multicollinearity）是指线性回归模型中的解释变量之间由于存在精确相关关系或高度相关关系而使模型估计失真或难以估计准确。影响

　　（1）完全共线性下参数估计量不存在

　　（2）近似共线性下OLS估计量非有效

多重共线性使参数估计值的方差增大，1/(1-r2)为方差膨胀因子(Variance Inflation Factor, VIF)

　　（3）参数估计量经济含义不合理

　　（4）变量的显著性检验失去意义，可能将重要的解释变量排除在模型之外

　　（5）模型的预测功能失效。变大的方差容易使区间预测的“区间”变大，使预测失去意义。

　　对于线性回归模型,当响应变量服从正态分布,误差项满足高斯–马尔科夫条件（零均值、等方差、不相关）时,回归参数的最小二乘估计是一致最小方差无偏估计。

　　当然，该条件只是理想化的假定，为的是数学上有相应的较为成熟的结论。其实大多数实际问题都不完全满足这些理想化的假定。

　　线性回归模型理论的发展正是在不断克服理想化条件不被满足时得到许多新方法。如加权LSE、岭估计、压缩估计、BOX\_COX变换等一系列段。做实际工作时一定是要超越书本上的理想化条件的。

3.11 逻辑回归常用的优化方法

参考答案：逻辑回归本身是可以用公式求解的，但是因为需要求逆的复杂度太高，所以才引入了梯度下降算法。

一阶方法：梯度下降、随机梯度下降、mini 随机梯度下降降法。随机梯度下降不但速度上比原始梯度下降要快，局部最优化问题时可以一定程度上抑制局部最优解的发生。

二阶方法：牛顿法、拟牛顿法.

3.12 列举几种解决数据不平衡问题的方法。

参考答案：

这主要是由于数据分布不平衡造成的。解决方法如下：

  1）采样，对小样本加噪声采样，对大样本进行下采样

  2）进行特殊的加权，如在Adaboost中或者SVM中

  3）采用对不平衡数据集不敏感的算法

  4）改变评价标准：用AUC/ROC来进行评价

  5）采用Bagging/Boosting/ensemble等方法

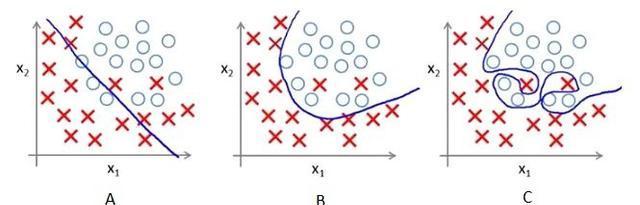
  6）考虑数据的先验分布

3.13 逻辑回归中可以用以下哪种方法来调整数据？

A.最小二乘法 B最大似然估计 C.杰卡德距离 C.A和B

参考答案：B.逻辑回归使用最大似然估计来训练回归模型。

3.14 下面3幅散点图是包含决策边界的逻辑回归模型，哪一幅图的决策边界过拟合了训练数据？哪一个决策边界的正则化值最大？



参考答案: C。不平滑的决策边界过拟合了数据。A。更大的正则化的值意味着惩罚更大，模型更简单。

3.15 判断下列说法是否正确，并说明理由。

（1）逻辑回归是监督机器学习的算法.

（2）逻辑回归主要用来做回归。

（3）在训练逻辑回归模型之前，对特征进行标准化是必须的。

参考答案：

（1）对，逻辑回归之所以是监督机器学习的算法，原因在于它使用了真值对数据进行训练。监督机器学习训练模型时会有输入变量X和目标变量Y。

（2）错。逻辑回归是分类算法。

（3）非必须。特征标准化的主要目的是实现模型的最优化。