# 2016－2017年第一学期

# 《大数据统计基础》试题

**说明：**

1. **试题共含五道大题，总计100分。请仔细阅读题目，按照要求完成，并在指定时间内提交。**
2. **所有分析应填写在答题纸内的相应位置上。**
3. **在分析正文中，要求用统计图、统计表、截图等展示分析结果，并通过文字加以阐述；整个分析过程中的程序语句请加上注释后附后。答卷以word方式提交。**

一、抽样

在loan data的2013-2014年数据集中，对变量loan\_amnt，使用简单随机抽样分别抽取样本容量为100、1000、5000、10000的样本，给出相应的样本质量，并且计算最优样本容量。（20分）

二、数据预处理

对Cell segmentation数据进行变量选择

数据获取的方法：

> library(AppliedPredictiveModeling)

> data(segmentationOriginal)

1、参考《Applied Predictive Modeling》中的第19章，使用该章中提到的过滤法（filter）和封装法（wrapper），来选择一个最优的预测变量子集。（10分）

2、对于linear discriminant analysis和logistic regression，使用嵌套（build-in）的特征选择方法（如R包glmnet和稀疏的LDA）选择变量。对比这些方法，从预测变量个数、预测性能、训练时间来对比不同方法。（10分）

参考书籍：Max Kuhn, Kjell Johnson. Applied Predictive Modeling. Springer, 2013.

三、数据探索性分析

“手机问卷数据.xls”是对手机用户进行满意度调查所得数据（问卷请见“手机问卷.doc”）。根据相关理论，Q1－Q22中可能隐含了三个潜在因素，分别是：用户体验、性价比、象征价值，而且这三个因素会对手机用户的满意度（Q23-25）和忠诚度（Q26-Q28）产生影响。请运用所学过的知识，对该数据开展分析。要求：

1、提取出用户体验、性价比、象征价值三个潜在因素，说明它们对数据的代表性，并予以适当分析；（10分）

2、分析上述三个潜在因素对满意度、忠诚度的影响。（7分）

3、从数据中是否还能获得了其他有价值的信息？请说明分析思路和你的结论。（3分）

四、数据可视化

在loan data中，自选合适的变量，绘制以下图形：

1、分面的风玫瑰图，玫瑰叶片的颜色至少四种；（4分）

2、某一个连续型变量的分布直方图，并加入拟合分布线，直方图的组距和组数自己设定（不要使用默认的），并且每个柱子里面填上相应的组的频数，整个图片加上一个黑色的外框，并且图的底色为深色，柱子为浅色，在密度最高的部分加上文字标注“此处密度最大”；（6分）

3、某两个连续型变量的密度图，并且在图中找出一个部分加上一个方框与其他部分区别开来（比如说，密度最高或者最低的区域，用一个矩形的方框加以标示）；（6分）

4、挑选多个连续型变量，进行聚类，并且绘制相应的热图并进行美化（可以不用全部样本）。（4分）

要求：变量选择方面尽量使整个分析具有一定的逻辑性，而不是为了画图而画图。以上所有图表都要求有图表名称、图例、行标题、列标题等要素。

五、空间统计

基于“动态数据2012.xlsx”，完成以下任务：

1、2012年长三角地区四个季度某疾病发病率动态地图展示（10分）

2、2012年长三角地区四个季度某疾病的发病热点区域探测（10分）

# 2016－2017年第一学期

# 《大数据统计基础》试题 答题纸

**学校\_\_\_\_\_\_\_\_\_\_\_\_\_\_学号\_\_\_\_\_\_\_\_\_\_\_\_\_\_姓名\_\_\_\_\_\_\_\_\_\_\_\_\_\_成绩\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**一、抽样**

**二、数据预处理**

**三、数据探索性分析**

**四、数据可视化**

**五、空间统计**