大家好，这是今天的课堂纪要，可以帮助大家快速复习。

第一部分：作业反馈

（1）如果离散型变量中因子水平个数过多，可以合并分组，比如可以将城市重新分成一线/二线...或者南方/北方

（2）如果连续型变量出现大量的单一值（比如0）， 可以进行分组离散化再分析。也可以替换成两个变量，第一个变量为二值变量，反映取值是否等于该单一值，第二个变量对于原始取值为单一值的记录记为0，对于取值不是该单一值的

记录保持原值， 其系数反应的是当变量不等于该单一值时，每增加1个单位，所引起的因变量的变化

（3）图表要有标号、标题，要在正文中予以引用

第二部分：逻辑回归（ST案例）

在我国的资本市场中，上市公司是否被ST是一个备受投资者关注的指标。该案例的研究目的是选用第t年不亏损公司的财务数据预测第t+3年是否被ST，因此因变量为是否被ST, 这是一个0-1变量。自变量包括应收账款占比（反映盈利质量）、资产规模（反映公司规模）、资产周转率（反映资产利用效率，注意要同行业进行比较）、资产回报率、销售增长速度、杠杆水平等、第一大股东持股比率。

第三部分：逻辑回归（理论基础）

（1）因变量y为0-1变量，如何将它变为连续取值呢？考虑y=1的概率，P(y=1)就是连续的变量，但概率的取值在0～1之间。对于，它的取值有可能为负数，此时采用指数变换可以将数据变化为大于0，再调整取值到0～1之间，最终得到，称为广义线性模型（Generalized Linear Model）。经过变换可以得到log(p(y=1)/p(y=0))=，左边称为对数几率（Log Odds Ratio），是对概率的logit变换，右边为解释性变量的线性组合。

（2）与线性回归相比，逻辑回归的不同在于：

a. 用极大似然方法进行估计，大致思想是存在的就是合理的

b. 将线性回归中的残差平方和更加一般化为deviance（离差），利用离差进行假设检验，DEV=-2（对数似然函数）

c. AIC与BIC准则将原本的残差平方和改为现在的离差平方和

d. 模型整体的显著性检验由F-检验变为卡方检验

第四部分：逻辑回归（预测评估）

（1）对数据切分为训练集和验证集，在训练集上估计参数，在验证集上预测

（2）两种评判标准：

a. 错分率（Misclassification Rate），如果y=1的数据占比很小时例如占1%，此时模型如果全部预测都为0，正确率都可以有99%，此时错分率只有1%，但这显然不是一个好模型。

b. 加权错分率（Weighted Misclassification Rate）

（3）TPR（True Positive Rate），好人中多少比例被找出来；FPR（False Positive Rate）：坏人中多少被认为是好人

（4）ROC曲线可以看做是在不同阈值下，对应的FPR（横坐标）和TPR（纵坐标）连成的曲线。45度线上任何一点对应一种胡蒙乱猜的策略，ROC曲线偏离45度线越远，说明模型付出了较少的成本（FPR）得到了较高的收益（TPR）。可以使用AUC（Area Under Curve）来衡量这种偏离的水平，因为模型的预测结果与阈值的选取有关，模型的综合表现应该考虑阈值的所有可能取值，所以一般使用AUC作为评价模型预测能力 的指标，取值越大说明模型的预测能力越好。问题：在业务上究竟选择ROC曲线上的哪个点呢？答：在运营中调整。

（5）为大家介绍了两个行业内的实际案例，首先是支付刷卡数据，基于这种数据，我们提出了RFMS模型，并在这个方法论的指导下，产生大量的X标签，并最终变成一个可被产品化的模型。其次是APP数据。通过这个案例，希望大家了解APP数据采集行业的大概技术方法，从这个案例可以看到什么样的APP是最能表达一个人的特指的。