ROC曲线：

* X轴=FP/(TP+FP)，即FP在所有预测的P中的比例
* Y轴=TP/(TP+FN)，即TP在所有实际的P中的比例
* 最好的点（0,100）--- 所有的P点都正确预测到，且没有N被错误预测为P

对于不平衡的数据：over-sampling少数类 + under-sampling多数类 –->两类数据的数据大体平衡

1. 随机抽样 + 填补/替代数据
2. Metacost算法、costSensive

实验显示：

1. underSampling多数类比overSampling少数类效果好
2. 两者结合的效果，并不比underSampling的效果好

OverSampling少数类技术：

1. 变换抽样 --- 容易引起overfit

随机变量y: g(y) 🡨 不易进行抽样

随机变量x: f(x) 🡨 容易进行抽样

如果能够找到x和y之间的一个一一对应的变换关系，y=y(x)，使得g(y)和f(x)满足关系

g(y) = f(x)| dx/dy|

则可先由f(x)分布抽取x的值x0， 再由变换得到y的对应值y0，满足g(y)

1. SMOTE：Synthetic Minority Over-sampling Technique

合成新的少数类样本，合成的策略是：对每个少数类样本a，搜索其k个最近邻样本；若向上采样的倍率是N，从k中选N个样本b0~bN，然后在a、b之间的连线上随机选一点作为新合成的少数类样本。一般k=5。例如，如果抽样倍率是200%，则从5个最近点中取2个点，每个方向上合成一个新的样本。

1. Under-sampling and SMOTE Combination

通过随机删除多数类中样本的方式underSampling多数类，直到少数类和多数类之间的比例达到指定值

underSampling多数类200%，意思是：在修改后的数据集包含的少数类的样本个数是多数类的2倍。

# 总结

1. SMOTE可以提高分类器少数类的精确性
2. SMOTE和underSampling结合的效果比单纯underSampling好