openGauss DB4AI特性实验

姓名： 王娇妹 学号： 2012679

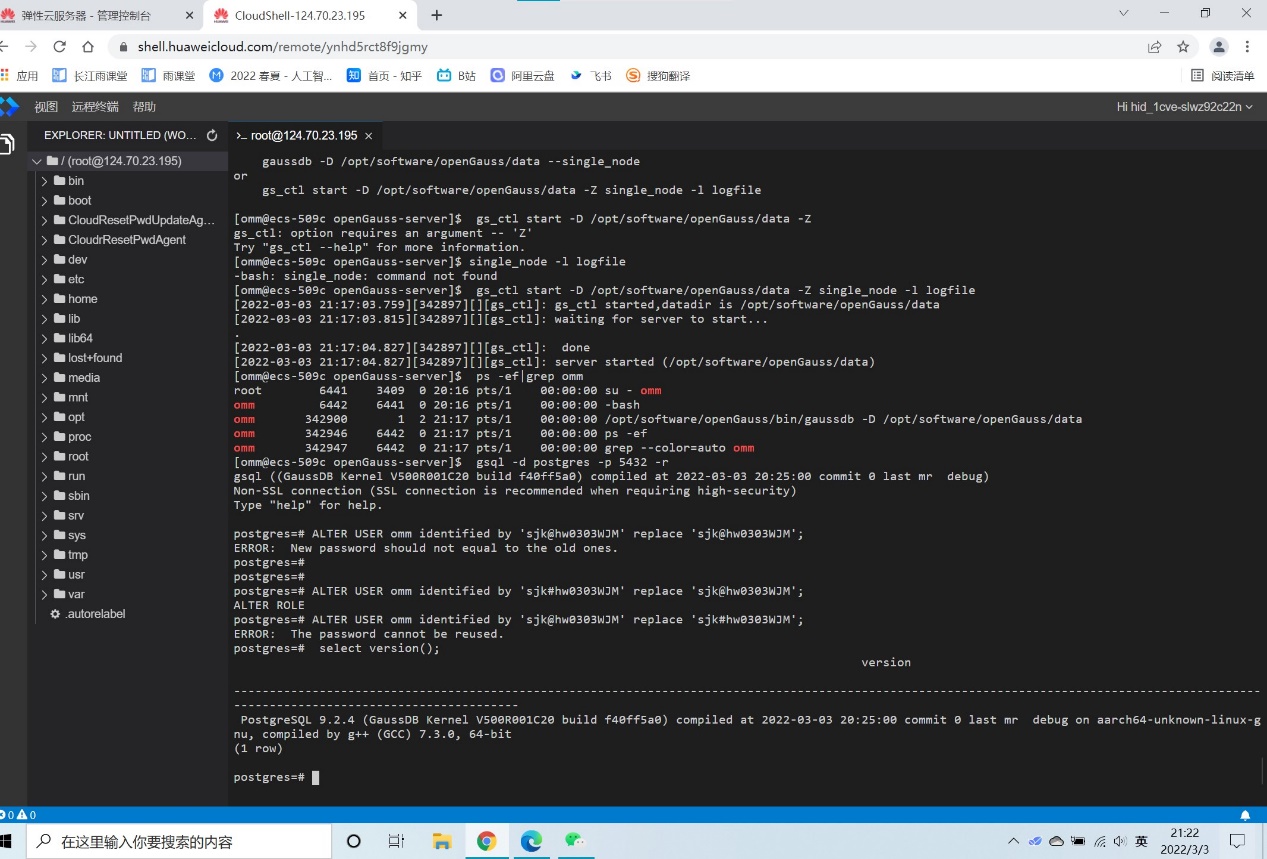
实验步骤：

* ECS弹性云服务器购买
* openGauss数据库安装
* openGauss DB4AI特性实验
* 云服务器资源释放/云服务器关机

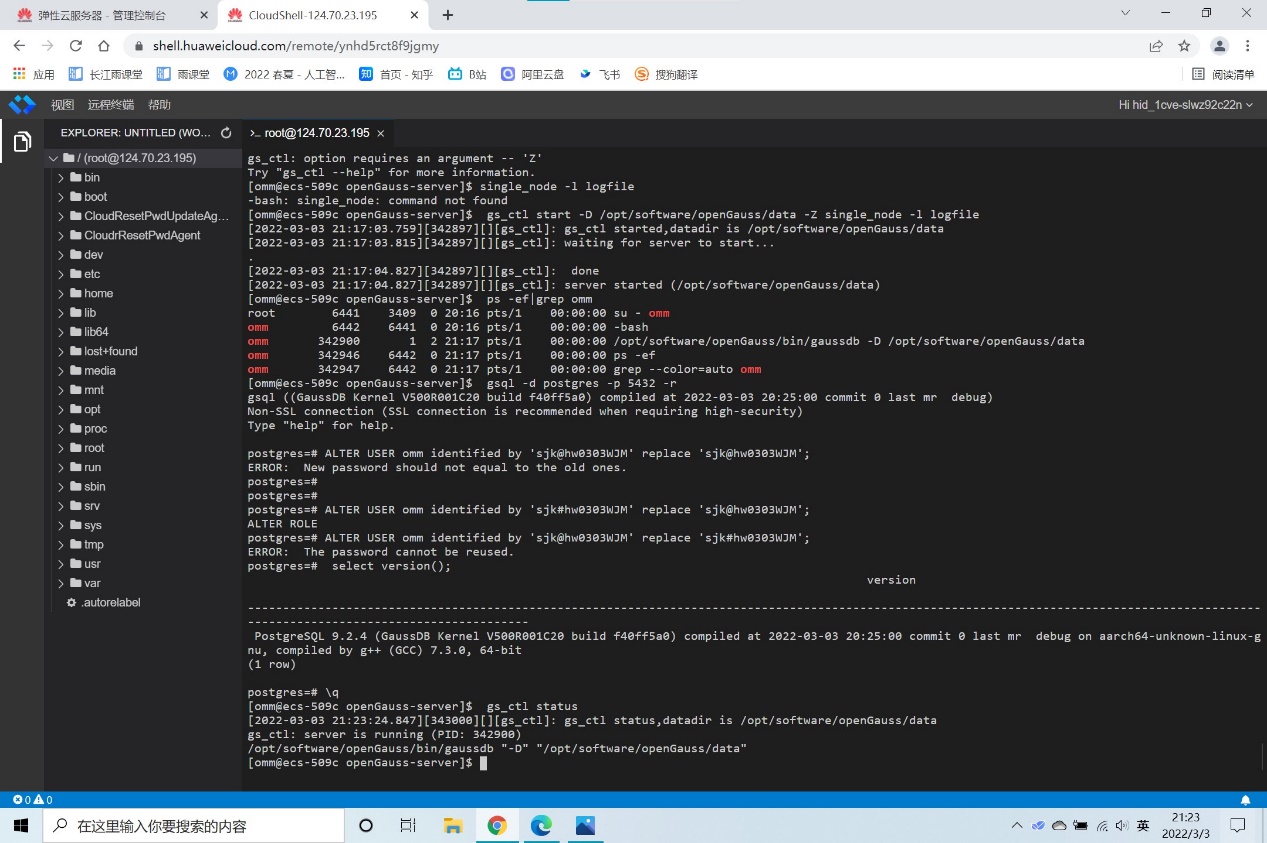
实验报告

实验步骤截图：

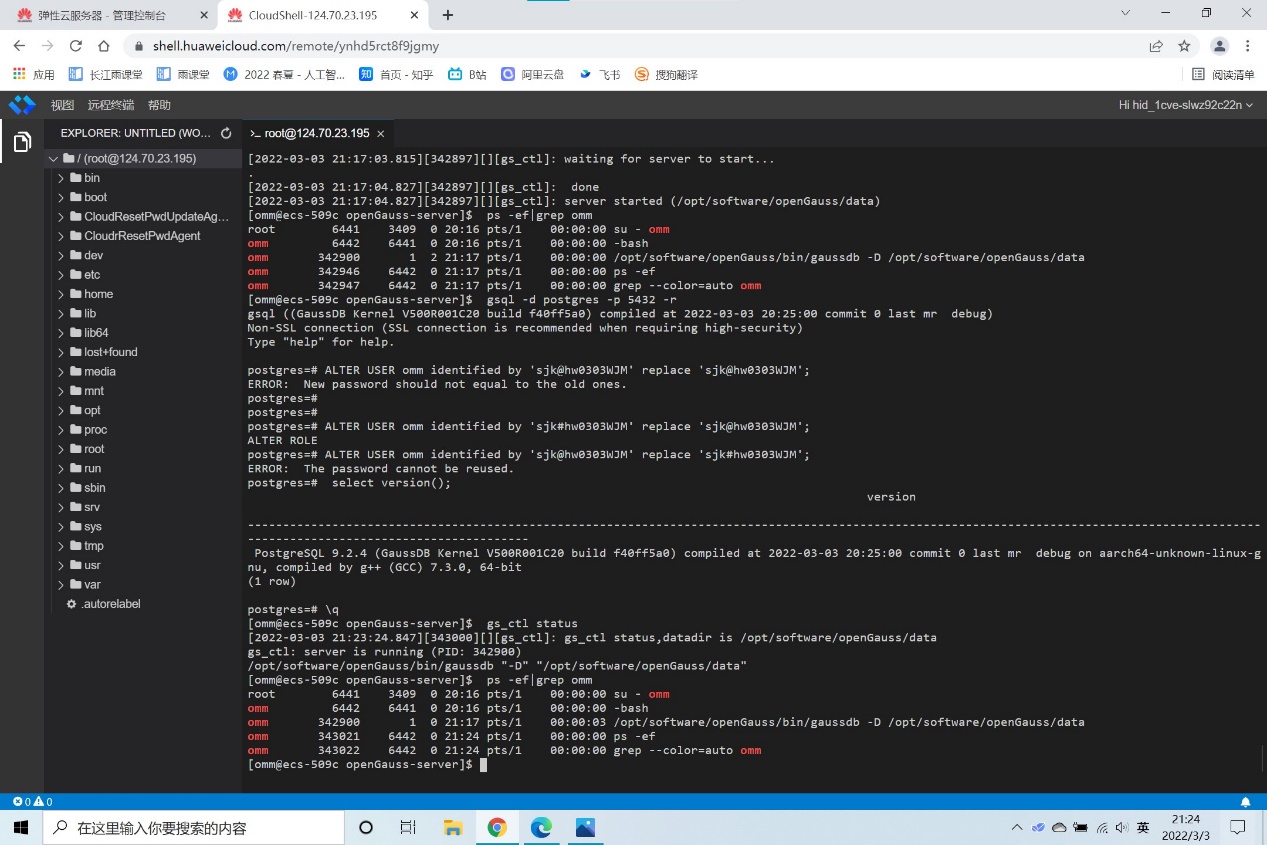
截图1：数据库编译安装版本截图



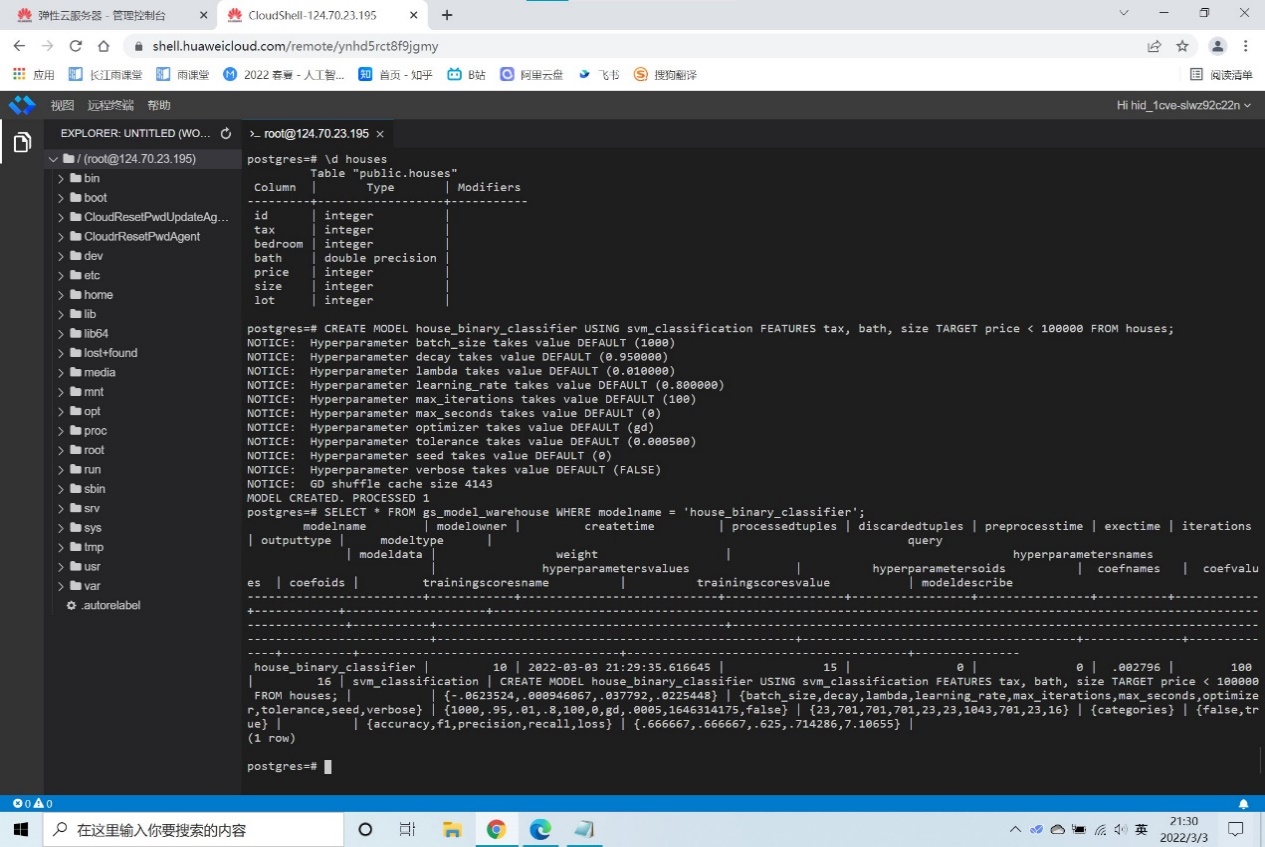
截图2：数据库状态截图



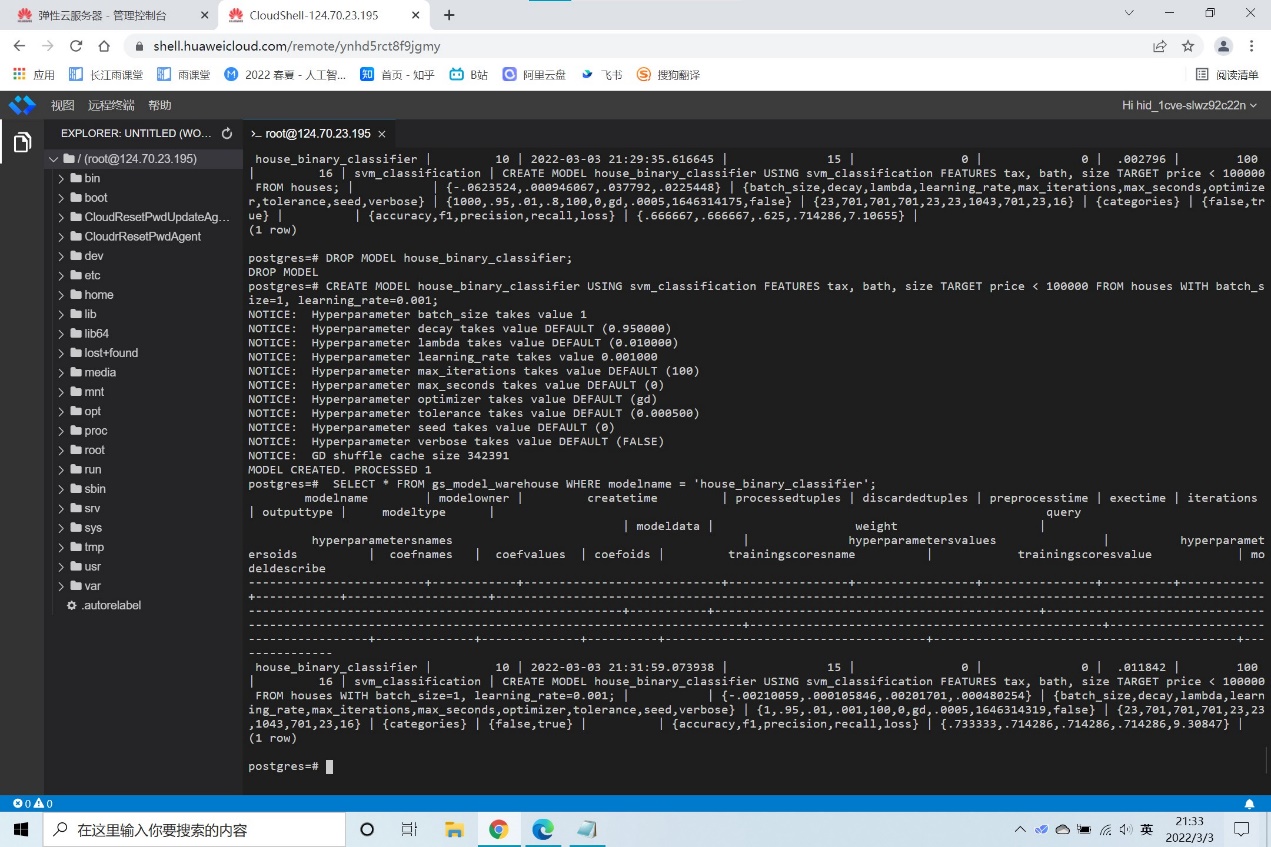
截图3：数据库进程截图（包含数据库服务器主机名）



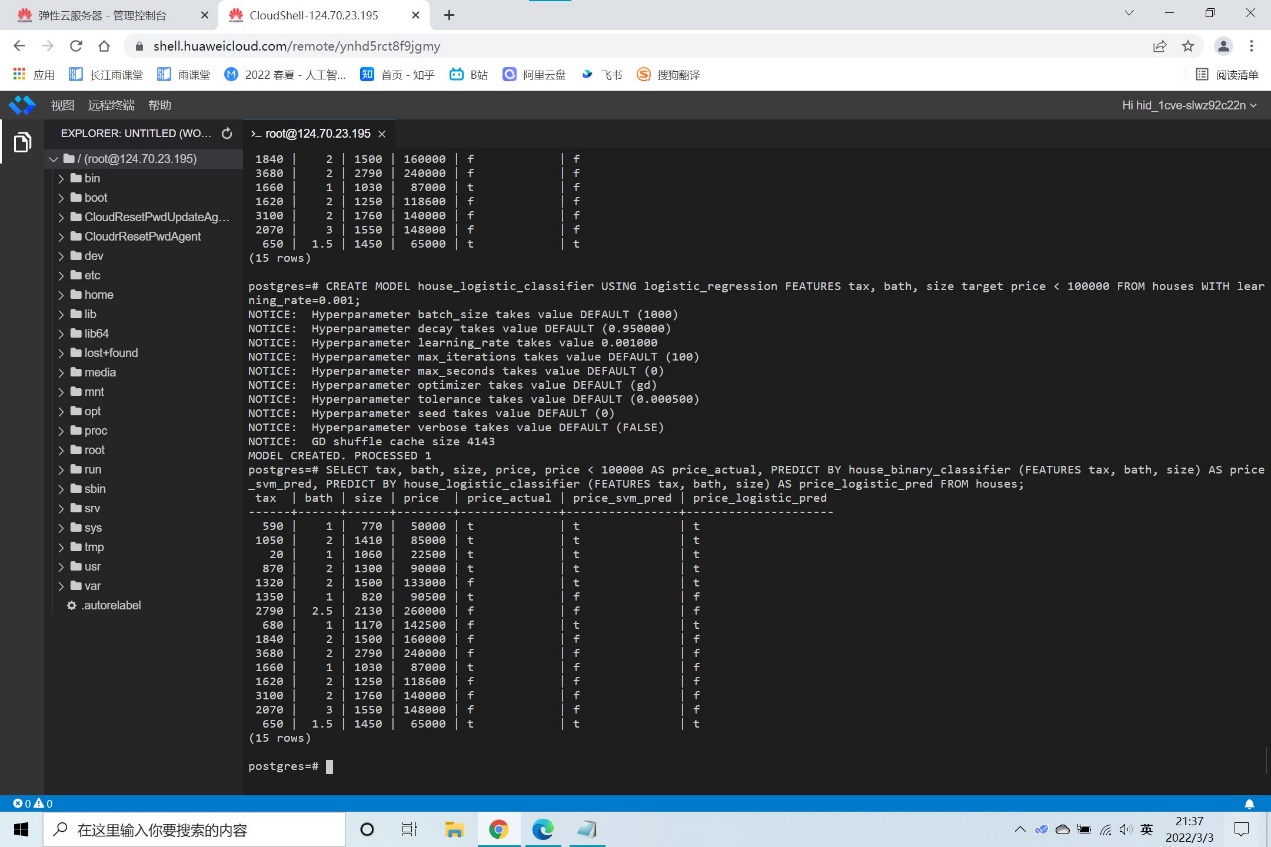
截图4：SVM二分类模型执行结果截图



截图5：超参设置新模型结果截图



截图6：SVM二分类模型与逻辑回归模型执行结果比较截图



实验思考题：

1. 为什么需要通过源码编译，完成数据库的安装？简述安装数据库所需步骤

在源码编译安装过程，我们可以自定义相关参数，按照需求进行安装，并且可以自行选择安装版本，更加灵活方便。此外，源码编译安装数据库还能够满足不同的运行平台， 兼容性更强。

安装数据库所需步骤：

1. 在编译前，用 root 用户名，使用之前设置的密码登录 ECS。创建数据库的安装用户 omm 及其属组 dbgrp，并修改密码。然后创建 openGauss 源码存放及 openGauss 安装路径。下载第三方编译库。解压下载好的第三方编译库，并重命名为 binarylibs。
2. 下载 openGauss 源码。上传 cmake 包，将 cmake-3.20.0-linux-aarch64.tar.gz 包，上传至服务器/usr/local/src/下（可使用winscp 工具）。然后进行解压。使用 yum 安装依赖包
3. 替换 python 版本。将 python3 的链接设置为 python 的，然后检查确定 python 的版本是否为3.7.4。最后修改/opt/software 路径的用户所属组及权限。
4. 数据库安装编译：切换 omm 用户并设定 omm 用户的环境变量。然后添加环境变量，使环境变量生效。进入 openGauss 源码下，生成配置文件。执行 make 命令进行编译。执行 make install 安装。
5. 初始化数据库。在初始化数据库时，需要设置数据库密码，并且要使用复杂密码。之后启动数据库并登录数据库，查询数据库版本状态等。
6. 分类模型与回归模型有何不同？什么是SVM算法？

分类模型与回归模型的不同之处：

两者的的预测目标变量类型不同，回归问题是连续变量，分类问题离散变量。

回归与分类的根本区别在于输出空间是否为一个度量空间。回归问题是定量问题，分类问题是定性问题。对于回归问题，其输出空间B是一个度量空间，即所谓“定量”，回归问题的输出空间定义了一个度量去衡量输出值与真实值之间的“误差大小”；对于分类问题，其输出空间B不是度量空间，即所谓“定性”，在分类问题中，只有分类“正确”与“错误”之分，相当于将事物打上一个标签。

SVM算法：

SVM是支持向量机（support vector machines）的简称，它是一种二分类模型。SVM将实例的特征向量映射为空间中的一些点，目的就是想要画出一条线，以 “最好地” 区分这两类点，以至如果以后有了新的点，这条线也能做出很好的分类。SVM 适合中小型数据样本、非线性、高维的分类问题。

1. 分类问题有哪些指标，其含义是什么？
2. 准确率

Accuracy = =

Accuracy 能够清晰的判断模型的表现，准确率可以反映模型的整体情况。

2、精确率

Precision = =

Precision是正确分类的正例个数占分类为正例的实例个数的比例，能够很好的体现模型对于负样本的区分能力，精确率越高，则模型对负样本区分能力越强。

3、召回率

Recall = =

Recall是正确分类的正例个数占实际正例个数的比例，能够很好的体现模型对于正样本的区分能力，召回率越高，则模型对正样本的区分能力越强。

4、F1值

F1 = =

F1值是对Precision和Recall进行的整体评价。

5、F-Measure

F-Measure =

Precision和Recall指标有时候会出现矛盾的情况，这样就需要综合考虑它们，最常见的方法就是F-Measure（又称为F-Score）。F-Score是Precision和Recall的加权调和平均。

6、ROC曲线和AUC

ROC曲线：接收者操作特征（receiver operating characteristic），ROC曲线上每个点反映着对同一信号刺激的感受性。

横轴：负正类率（FPR，特异度）；纵轴：真正类率（TPR，灵敏度）。

AUC：ROC曲线下的面积，介于0.1和1之间。AUC作为数值可以直观的评价分类器的好坏，值越大越好。AUC值是一个概率值。假设分类器的输出是样本属于正类的Score（置信度），则AUC的物理意义为，任取一对（正、负）样本，正样本的Score大于负样本的Score的概率。

1. 回归问题有哪些评价指标，其含义是什么？
2. 均方误差MSE

MSE =

1. 均方根误差RMSE

RMSE =

1. 平均绝对误差MAE

MAE =

1. R2

R2 = 1-

R2是线性回归的衡量标准。在R2中，分子是真实值和预测值之差的差值，是模型没有捕获到的信息总量，分母是真实标签所带的信息量。