线性回归器的详细步骤

- **第一步:初始化步骤**
  - 导入数据
    - 整理好数据
    - 预处理数据
    - 替换空数据
    - 分离出变量X 和 变量y
  - 建立模型后,需要部分数据去训练, 和检查是否达到满意度,就会把数据分成Traindata and Testdata
  - 分配好数据之后,skl 只能识别np中的数组,所以需要把X_train 通过reshape 变成numpy的数组
  - 导入 from sklearn import linear_model 的模块
  - 创建线性回归对象 , linear_regressor = linear_model.LinearRegression()
  - 利用训练数据集训练了线性回归器。向fit方法提供输入数据即可训练模型
    - 用训练数据集训练模型 ,linear_regressor.fit(X_train, y_train)
  - 开始制作图,看看训练拟合出来的模型是否客观
  - 在前面的代码中，我们用训练的模型预测了训练数据的输出结果，但这并不能说明模型对未知的数据也适用，因为我们只是在训练数据上运行模型。这只能体现模型对训练数据的拟合效果
    - import matplotlib.pyplot as plt
    - y_train_pred = linear_regressor.predict(X_train)
    - plt.figure()
    - plt.scatter(X_train, y_train, color='green')
    - plt.plot(X_train, y_train_pred, color='black', linewidth=4)
    - plt.title('Training data')
    - plt.show()
  - 接下来用模型对测试数据集进行预测，然后画出来看看
- **第二步:计算回归模型的准确性**
  - 误差（error）表示实际值与模型预测值之间的差值
  - 几个衡量回归器拟合效果的重要指标（metric）
    - 平均绝对误差（mean absolute error）：这是给定数据集的所有数据点的绝对误差平均值
      - round(sm.mean_absolute_error(y_test, y_test_pred)，2)
    - 均方误差（mean squared error）：这是给定数据集的所有数据点的误差的平方的平均值。这是最流行的指标之一。
      - round(sm.mean_squared_error(y_test, y_ test_pred)，2)
    - 中位数绝对误差（median absolute error）：这是给定数据集的所有数据点的误差的中位数。这个指标的主要优点是可以消除异常值（outlier）的干扰。测试数据集中的单个坏点不会影响整个误差指标，均值误差指标会受到异常点的影响
      - round(sm.median_absolute_error(y_ test, y_test_pred), 2)

    - 解释方差分（explained variance score）：这个分数用于衡量我们的模型对数据集波动的解释能力。如果得分1.0分，那么表明我们的模型是完美的
        - round(sm.explained_variance_ score(y_test, y_test_pred), 2)
    - R方得分（R2 score）：这个指标读作“R方”，是指确定性相关系数，用于衡量模型对未知样本预测的效果。最好的得分是1.0，值也可以是负数。
        - round(sm.r2_score(y_test, y_test_pred), 2)
    - 每个指标都描述得面面俱到是非常乏味的，因此只选择一两个指标来评估我们的模型。通常的做法是尽量保证均方误差最低，而且解释方差分最高。
- 保存模型数据
    - 模型训练结束之后，如果能够把模型保存成文件，那么下次再使用的时候，只要简单地加载就可以了