sklearn中的模型

该课程主要诶大家讲授如下的内容：

- 模型训练（数据集划分、sklearn中的模型、常用API）

1. 模型训练
    1. 数据集划分
    进入模型训练前，还需要将数据进行划分：训练集和测试集，测试集是为了验证模型的性能而准备的。
    化代码样式为：

    ```
    X_train,X_test,y_train,y_test=model_selection.train_test_split
    (test_size=None, random_state=None, shuffle=True, stratify=None)
    ```

    该方法中需要主要的参数有：
    test_size：float或int
    如果为 float，则应介于 0.0 和 1.0 之间，表示测试集的在整个数据集中所占比例
    如果为 int，则表示测试样本的绝对数。
    Shufflebool：拆分前是否对数据打乱。
    random_state ： 打乱时的随机数种子。
    2. Sklearn中的模型
    klearn内置了很多模型，包括监督学习的线性回归、高斯回归、MLP神经网络，以及无监督的主成分分析、聚类分析、因子分析等。具体的模型和相关内容大家可以去官网上查看。我们这里以逻辑回归模型为例，介绍模型类中常用的、通用的参数以及API。
    3. 模型参数（以逻辑回归为例）
    化代码样式为：

    ```
    logistic_regression =linear_model.LogisticRegression(
    penalty='l2',
    ```

```
tol=0.0001,
fit_intercept=True,
random_state=None,
solver='lbfgs',
max_iter=100
)
```

该方法中需要主要的参数有：
penalty：正则化惩罚项，str类型，可选参数为l1和l2，默认为l2。用于指定惩罚项中使用的规范。newton-cg、sag和lbfgs求解算法只支持L2规范。L1规范假设的是模型的参数满足拉普拉斯分布，L2假设的模型参数满足高斯分布。
tol：停止求解的标准，float类型。即求解到多少时认为已经求出最优解
fit_intercept：是否存在截距或偏差，bool类型。
random_state：随机数种子，int类型，仅在正则化优化算法为sag,liblinear时有用。
max_iter：算法收敛最大迭代次数，int类型，默认为10。仅在正则化优化算法为newton-cg, sag和lbfgs才有用，算法收敛的最大迭代次数
solver：优化算法选择参数，`{‘newton-cg’,‘lbfgs’,‘liblinear‘,’sag‘,’saga‘}`。默认为liblinear。
liblinear：使用了开源的liblinear库来实现，内部使用了坐标轴下降法来迭代优化损失函数，适用于小数据集。
lbfgs：拟牛顿法的一种，利用损失函数二阶导数矩阵即海森矩阵来迭代优化损失函数。
newton-cg：也是牛顿法的一种，利用损失函数二阶导数矩阵即海森矩阵来迭代优化损失函数。
sag：即随机平均梯度下降，是梯度下降法的变种，和普通梯度下降法的区别是每次迭代仅仅用一部分的样本来计算梯度，适合于样本数据多的时候。
saga：线性收敛的随机优化算法的变种。适用于大数据集。

4. 模型常用API
fit()方法，即模型的训练。放入训练集就可以对模型进行训练
predict()方法，即数据的预测，训练好的模型调用改方法，放入想

要预测的特征数据就能得到预测的结果。
predict_proba()方法，分类问题特有的方法，放入想要预测的特征数据，得到的这条数据属于各个类别的估计值。
score(X, y)方法，放入预先准备好的特征数据和标签数据，计算模型的得分，来评估模型。面对不同问题的模型会有不同的评估指标。线性回归的默认指标是结果的R^2值、Kmeans聚类的默认指标是聚类误差、逻辑回归的默认指标是平均精确度。

5. 评价指标
除了模型默认评价指标外，sklearn还单独提供了其他评价指标。如图所示，第一列为该指标的名称，在某些模型的API中可以指定想要使用的评价。第二列为指标对应函数，放入预测值和真实值即可得出对应指标。
评价指标1.png
评价指标2.png
