

第八课 机器学习简单介绍

第八节 数据集准备及划分

主 讲: Robin

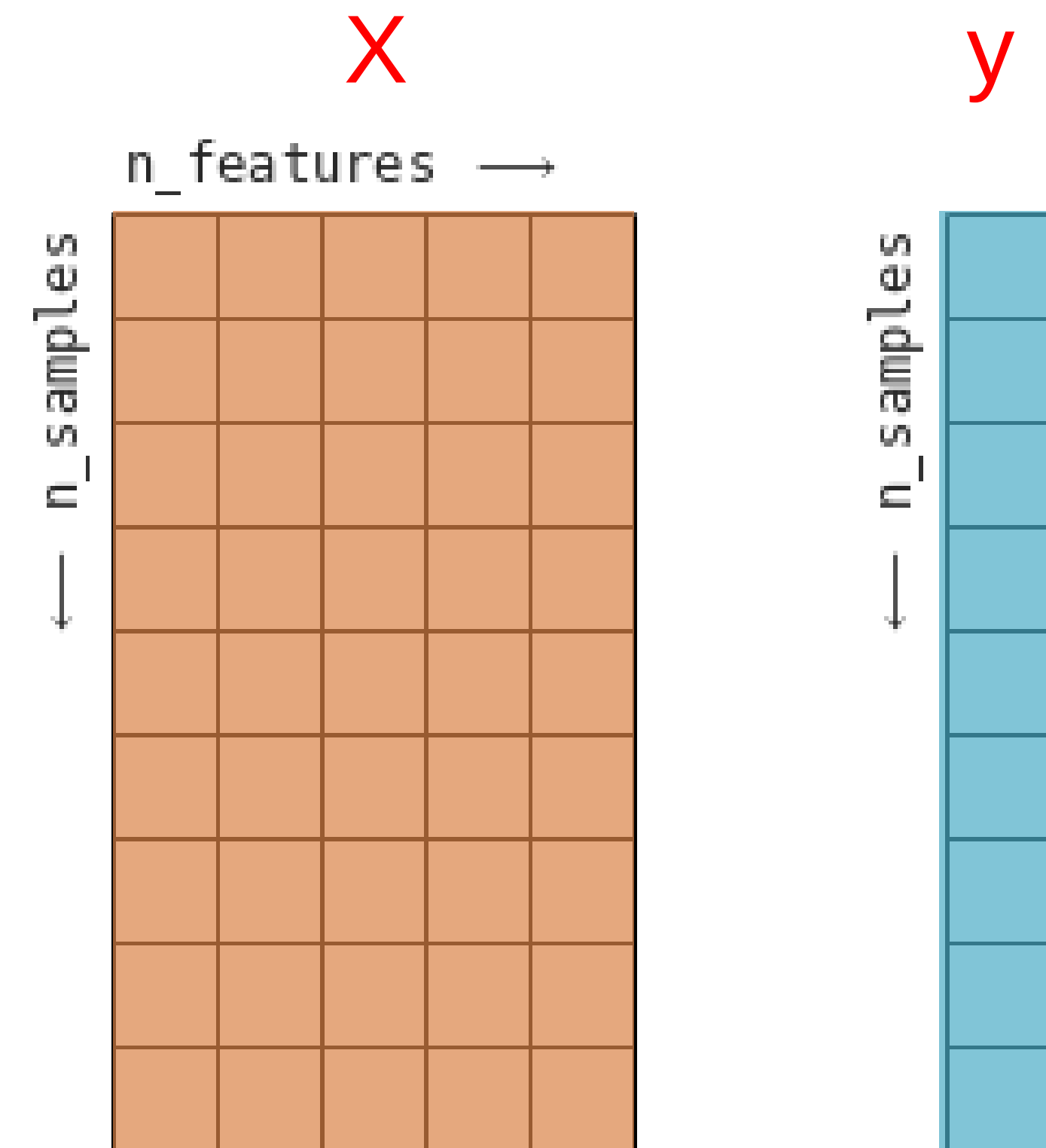
思考题：

- 在训练集上做测试是否合适？
 - 相当于考试使用了练习题。。。
- 现实中的情况是什么样的？
 - 预测的样本都是从未见过 (unobserved) 的



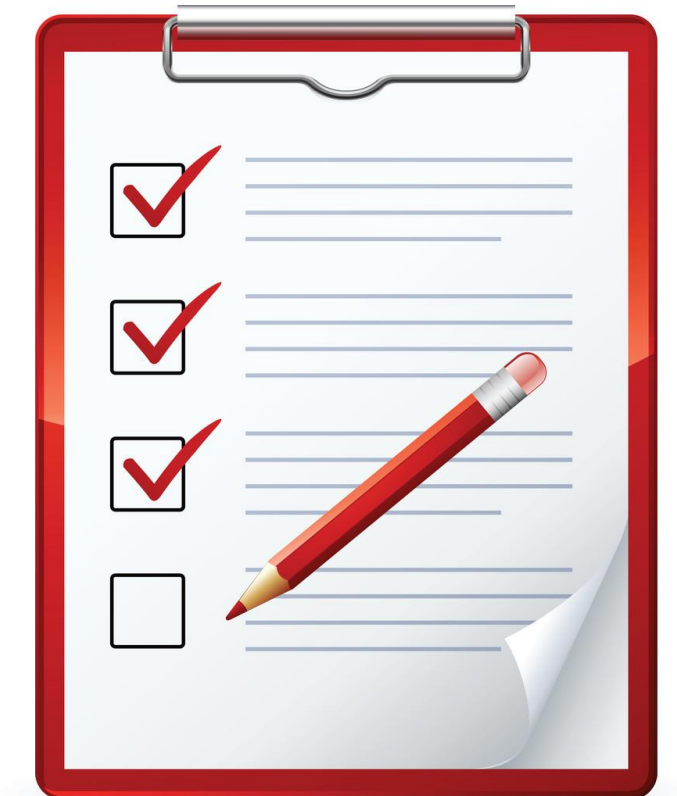
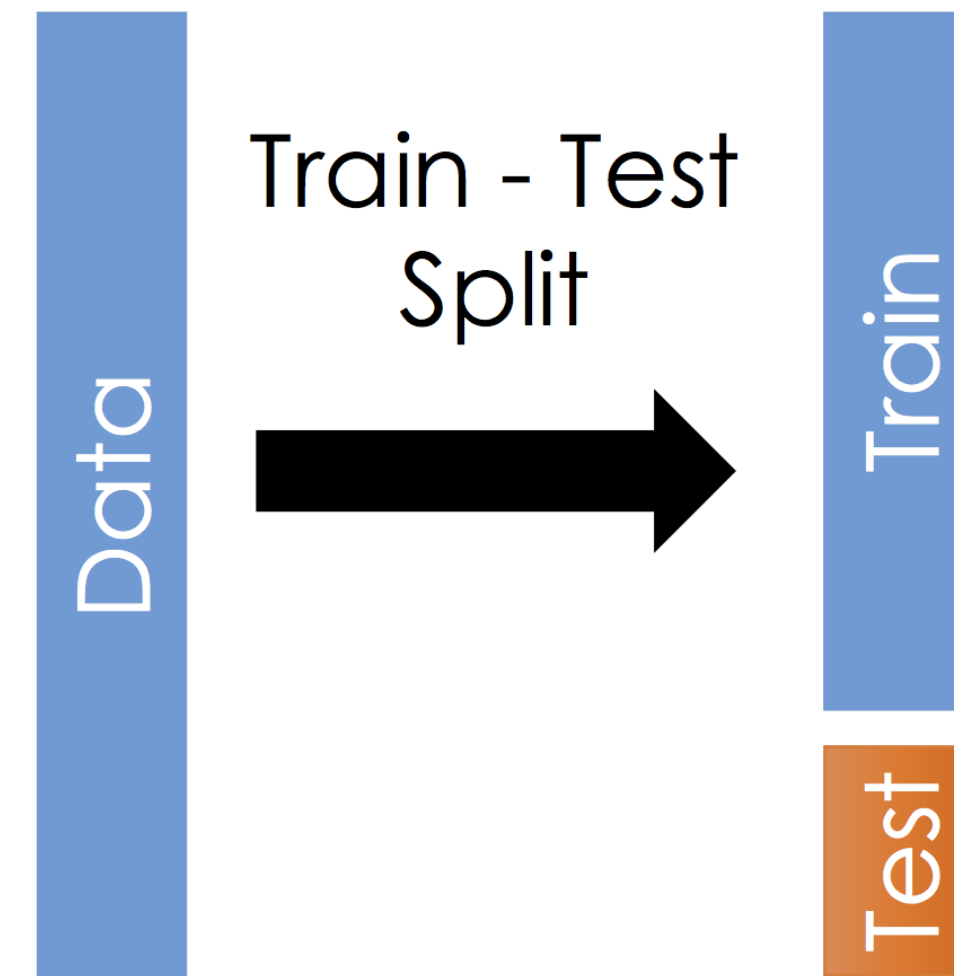
数据集准备及划分

- 准备：
 - 样本特征 (X) 格式
 - 二维数组, 形状 $(n_samples, n_features)$
 - 标签 (y) 格式
 - 一维数组, 形状 $(n_samples)$



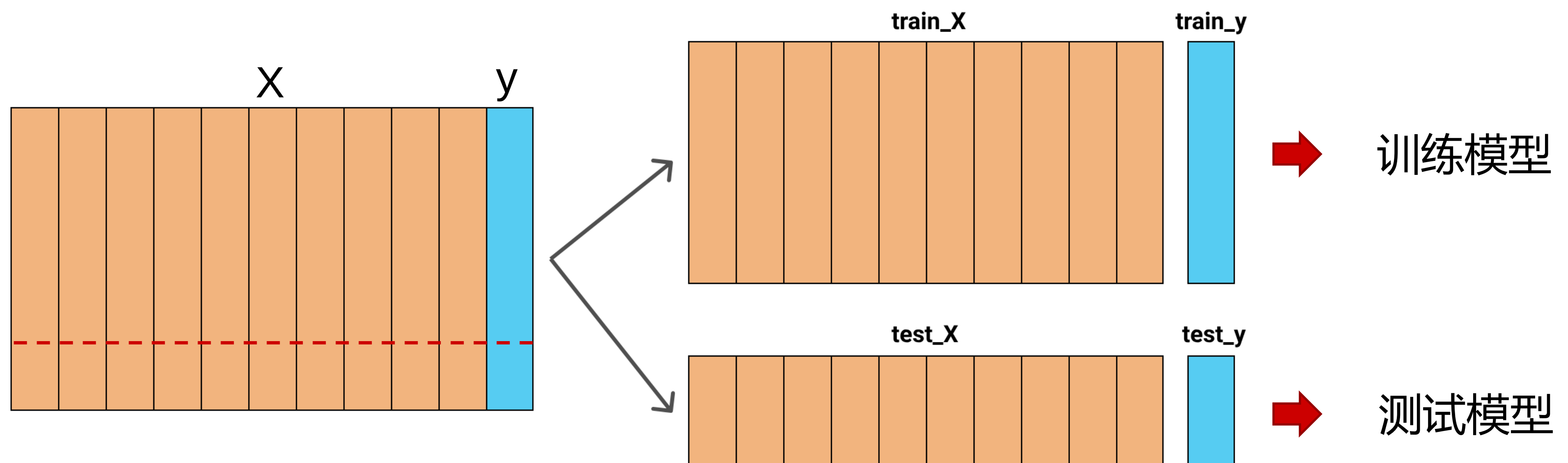
数据集准备及划分

- 将数据集划分为**训练集**和**测试集**
- 在训练集上训练模型
 - 输入训练数据的**特征**和**标签**
- 在测试集上测试模型
 - 输入测试数据的**特征**
 - 根据输出的**预测标签**和**真实标签**进行比较检验模型的性能



数据集准备及划分

- 如何划分?
 - $X_{\text{train}}, X_{\text{test}}, y_{\text{train}}, y_{\text{test}}$
`= train_test_split(X, y, test_size=1/4, random_state)`



- `random_state`: 随便赋值，为了保证程序每次运行都分割一样的训练集合测试集

只为遇见明天更优秀的你！