README

问题

安德森鸢尾花卉数据集$(Anderson'sIrisdataset)$，其中包含150个样本，对应数据集的每行数据。每行数据包含每个样本的4个特征和样本的1个类别信息。每个样本包含了花萼长度、花萼宽度、花瓣长度、花瓣宽度4个特征 (前4列)、1个品种信息，即目标属性 (第5列，也叫$target$)。

请将测试集与训练集按照$1:4$划分，建立一个分类器，分类器可以通过样本的4个特征来进行样本的分类，判断样本属于山鸢尾、变色鸢尾还是维吉尼亚鸢尾 (三个品种名称 : 分别对应0、1、2)中的哪种。

数据特征内容

	数据特征
1	$sepal\ \ length(cm)$: 花萼长度
2	$sepal\ \ width(cm)$: 花萼宽度
3	$petal\ \ length(cm)$: 花瓣长度
4	$petal\ \ width(cm)$: 花瓣宽度
5	$target$: 品种信息 $0,1,2$

关键指标

通过一个混淆矩阵阐述指标:

混淆矩阵

实际 \ 预测	预测为好的样本数	预测为坏的样本数
实际为“好”的样本数	*TP*	*FN*
实际为“坏”的样本数	*FP*	*TN*

1、查准率: $P=\frac{TP}{TP+FP}$ ，基于预测数据，看实际“好”样本的占比

2、查全率: $R=\frac{TP}{TP+FN}$ ，基于实际数据，看实际“好”样本的占比

但我们能够发现，查准率与查全率是一对矛盾的度量（如：查全率越高查准率就会越低）

3、想同时关注查准率和查全率，可考虑对两个指标进行加权，这里我们引入F_1值:

$$F_1=\frac{2RP}{P+R}$$

此时R,P的权重各取 $\frac{1}{2}$ ，将二者同等看待。

注：数据来源为*Creator:R.A. Fisher, Donor:Michael Marshall (MARSHALL%PLU '@' http://io.arc.nasa.gov)*