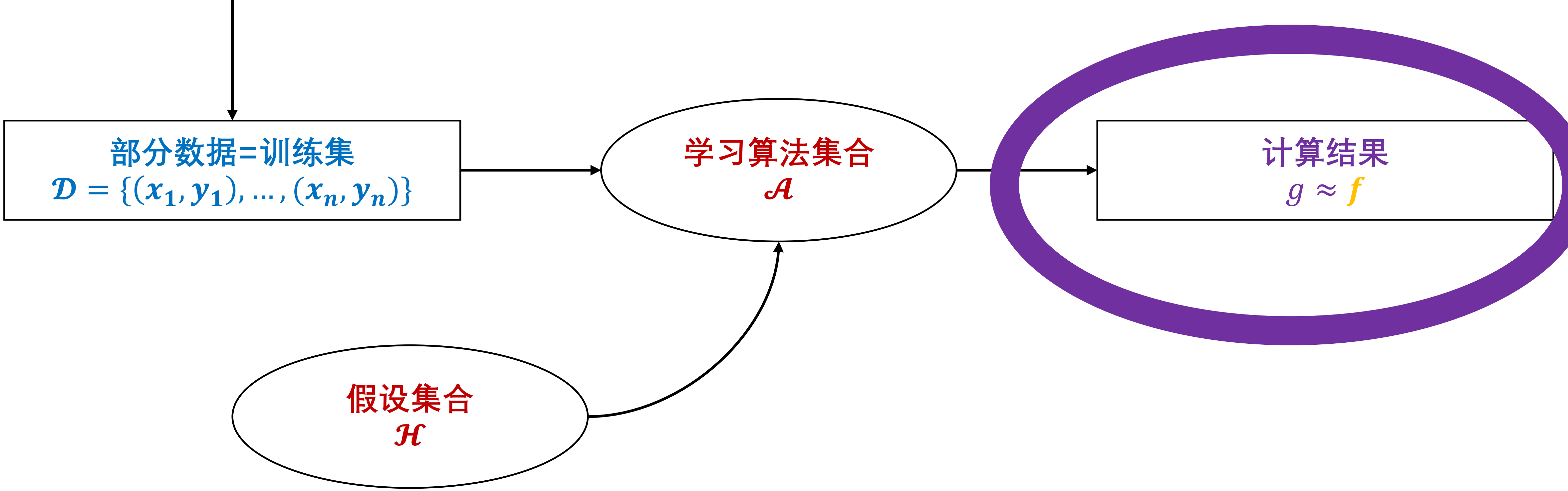
1.2 机器学习的抽象

机器学习：
使用**数据**，通过计算寻找**假设**g
使其接近目标**函数**f



2023

训练框架

目 录
CONTENT

Part 01

度量指标

- 二分类
- 多分类
- 回归问题

1.1

度量指标：二分类问题

最简单的问题：二分类问题如何度量?

正确率？但往往不好用。不仅有正确的问题
也有召回率的问题（想一想百度）

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\hat{y}_i = y_i\right)$$

四个指标：数据-模型

True positive, False positive
False negative, True negative

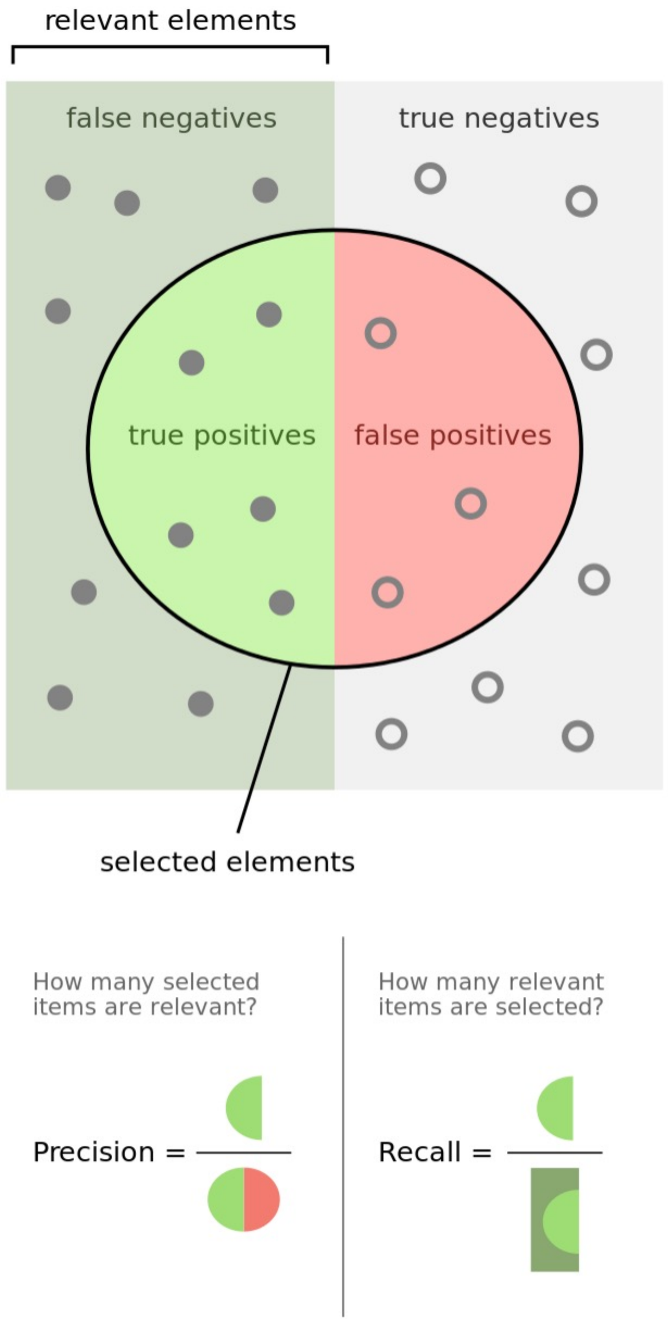
最简单的指标：f1_score

取值范围0-1，兼顾准确率与召回率

$$F1 = 2 * \frac{precision \ * recall}{precision + recall}$$

示例：贷款风险识别

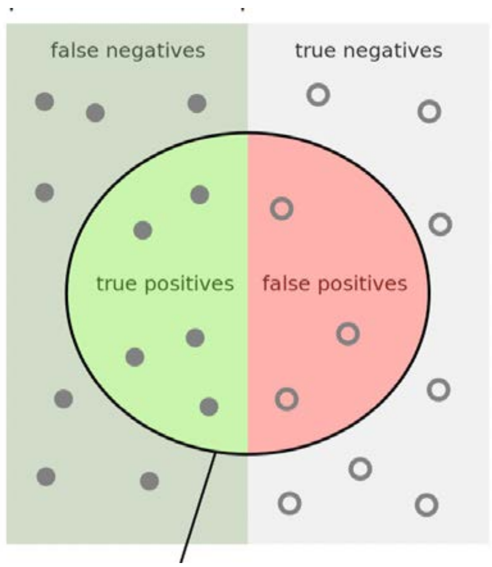
假如某种信贷的违约率大概在3%左右。
如果使用准确率，那么全部通过的话，“正确率”也是97%
但是f1 score，就是0



1.1 度量指标：二分类问题（续）

另外一种衡量模型能力的方法：ROC曲线与ROC面积（AUROC）

本质：衡量模型在不同宽紧度的识别力曲线



两个指标：TP ratio FP ratio

定义见左，我们来想象圆圈的大小与位置变动

真正例率 (**TPR**) 是召回率的同义词，因此定义如下：

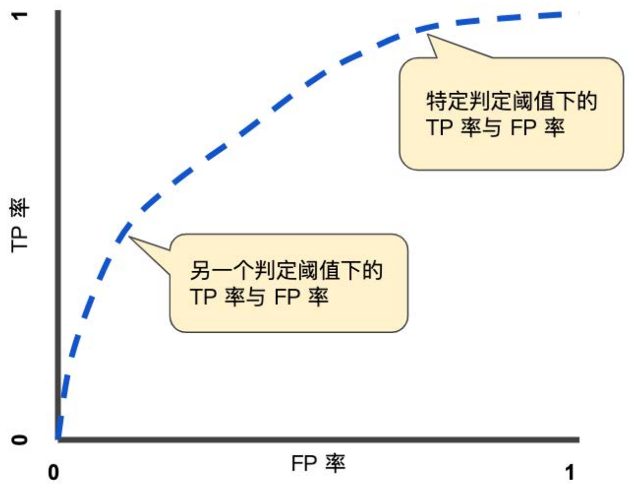
$$TPR = \frac{TP}{TP + FN}$$

假正例率 (**FPR**) 的定义如下：

$$FPR = \frac{FP}{FP + TN}$$

ROC曲线：将圆放大的思想实验

TPR是我们的获益，FPR是我们的代价
ROC曲线实际上是模型的投入-产出曲线
曲线一定连接0，0与1，1



ROC面积：量化指标

这条曲线与x轴与x=1围成的面积
取值范围是[0.5,1)
为什么下界是0.5?

1.1

ROC vs F1_score

$$F1 = 2 * \frac{precision \ * recall}{precision + recall} = \frac{\frac{TP}{FP+TP} * \frac{TP}{TP+FN}}{\frac{TP}{FP+TP} + \frac{TP}{TP+FN}}$$

真正例率 (TPR) 是召回率的同义词，因此定义如下：

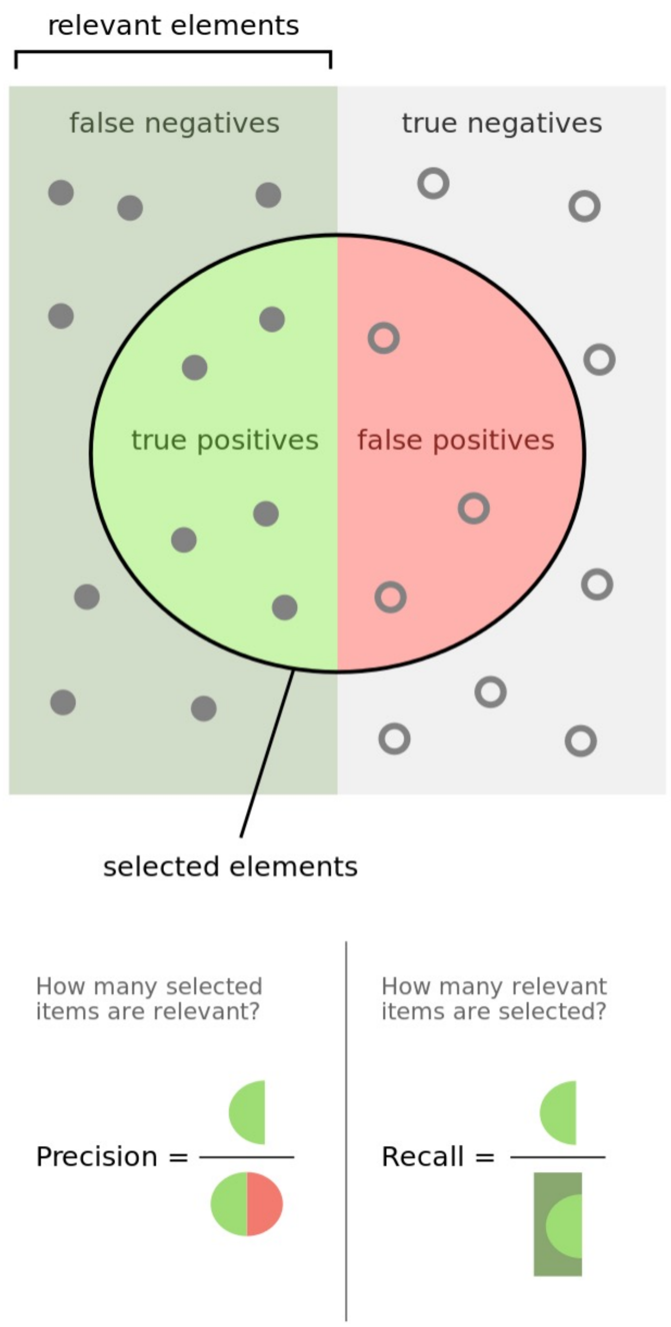
$$TPR = \frac{TP}{TP+FN}$$

假正例率 (FPR) 的定义如下：

$$FPR = \frac{FP}{FP+TN}$$

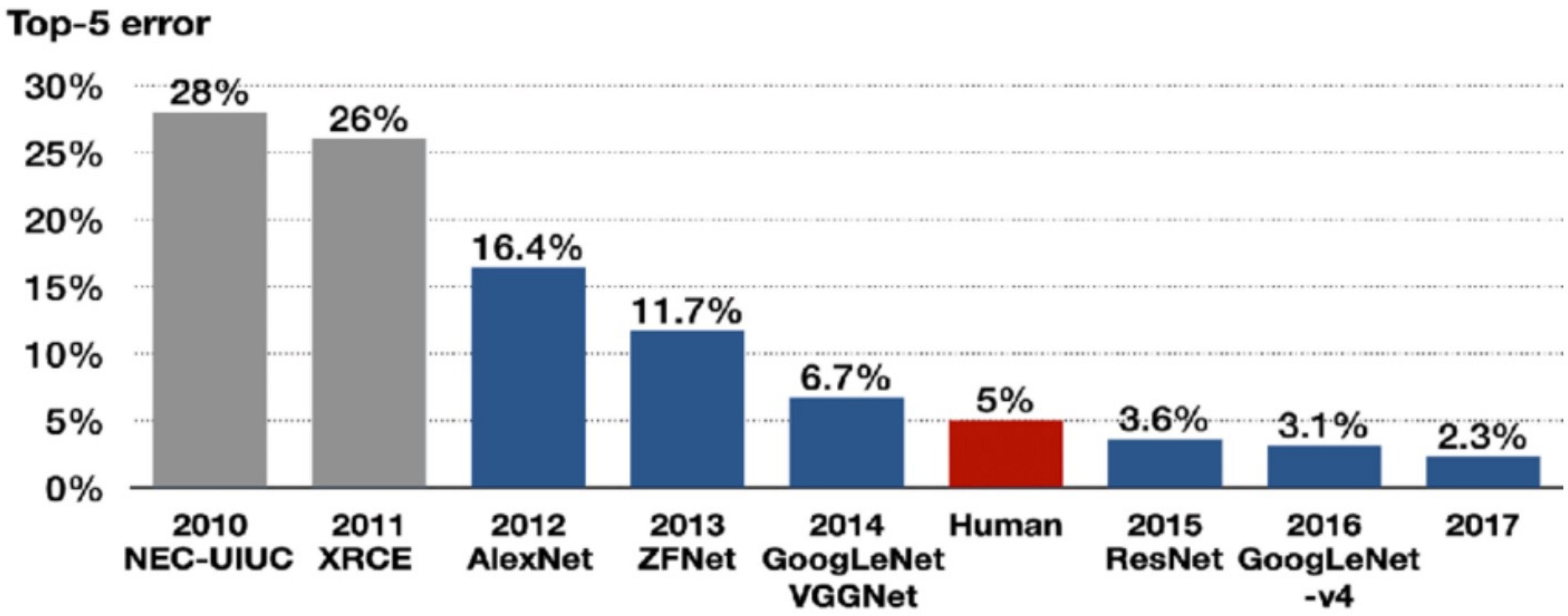
ROC vs F1

ROC上每一个点都有一个F1的值
F1衡量的是某个阈值下的表现，ROC是不同阈值下模型的综合表现

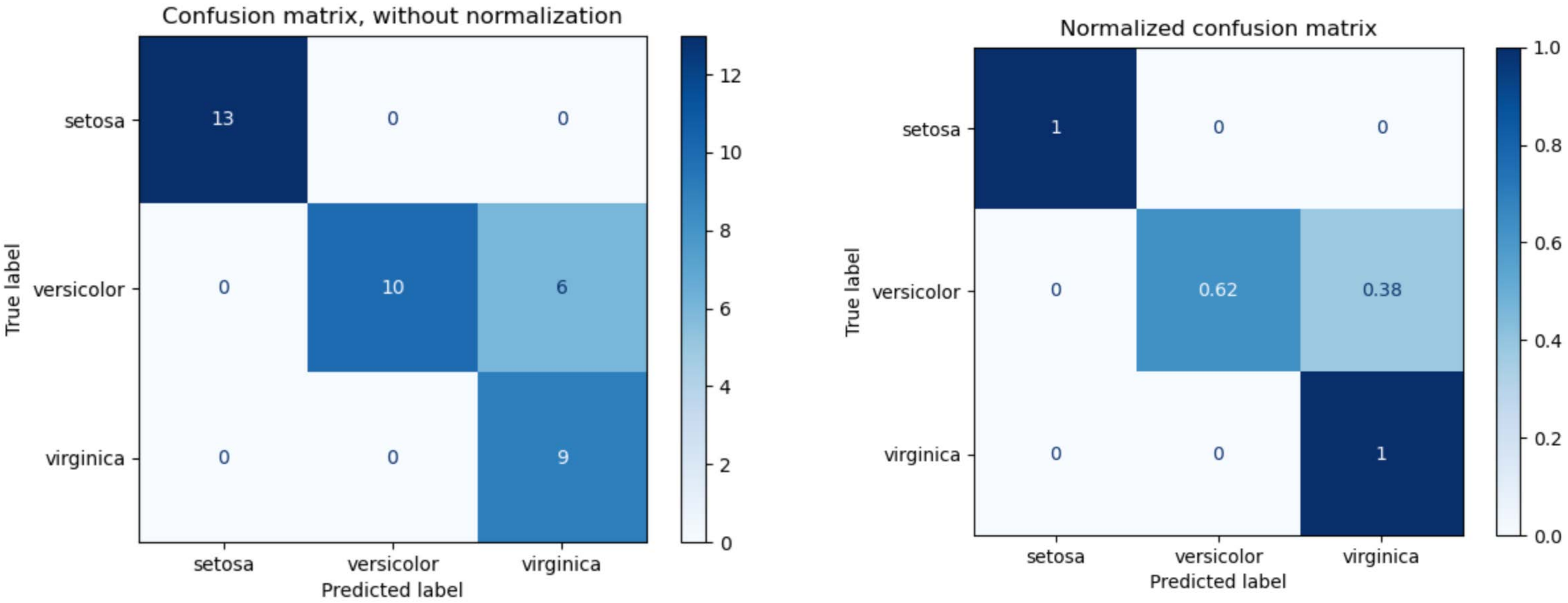


多分类问题： top k选择

$$\texttt{top-k accuracy}(y, \hat{f}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \sum_{j=1}^{k} 1(\hat{f}_{i,j} = y_i)$$



1.2 多分类问题：混淆矩阵



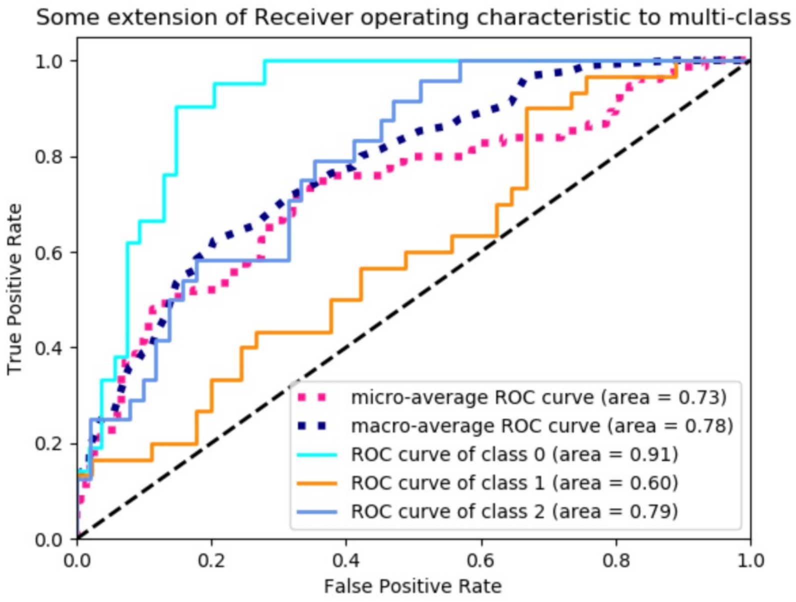
https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html#sphx-glr-auto-examples-model-selection-plot-confusion-matrix-py

多分类问题

F1-micro, F1-macro, F1-weighted

不同的f1组合方式，micro是计算整体，macro是计算不同类的简单平均，weighted以样本数为权重

	1类	2类	3类	4类	总数
TP	3	2	2	1	8
FP	0	0	3	1	4
FN	2	2	1	1	6



1.3 回归的评价方法

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2$$

最常用的一种

易于解释，RMSE则可以放缩至数据维度（标准化）
方便计算，显示解

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \widehat{y}_i|}{\max(\varepsilon, y_i)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$$

$$= 1 - MSE/Var(y)$$

最熟悉的一种

易于解释，不必理会数据原本特征，
易于计算 、转换

$$MedAE = median(y_1 - \widehat{y_1}, y_2 - \widehat{y_2}, \ldots, y_N - \widehat{y_N})$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y}_i|$$

MAE 好解释 好性质 不好计算

如果发现算不动，就不用

$$MaxAE = Max(|y_i - \widehat{y}_i|)$$

特别的模型用特别的Loss

计数模型用half Poisson
分位数回归用Pinball Loss
特殊模型一般会说明适用的Loss function