Machine Learning 2019

袁欣

2019 年 3 月 3 日

1 模型评估与选择

1.1 构造数据集

构造一个包含 1000 个样本的数据集，按照某种模型对样本排序，前 500 个样本中正例（取值 1）占 90%，后 500 个样本中反例（取值 0）占 80%。

- 代码如下：

```
pred <- c(round(runif(500) / 2 + 0.45),
          round(runif(500) / 2 + 0.10))
```

- 数据展示：

```
head(pred)
```

```
## [1] 1 1 1 0 1 1
```

```
tail(pred)
```

```
## [1] 1 1 1 0 0 0
```

```
mean(pred)
```

```
## [1] 0.544
```

- 真实均值：

$$\bar{pred} = (500 \times 0.9 + 500 \times 0.2) \div 1000 = 0.55$$

1.2 绘制曲线

试给出该模型的 P-R 曲线和 ROC 曲线的代码。

1.2.1 P-R 曲线

1.2.2 ROC 曲线

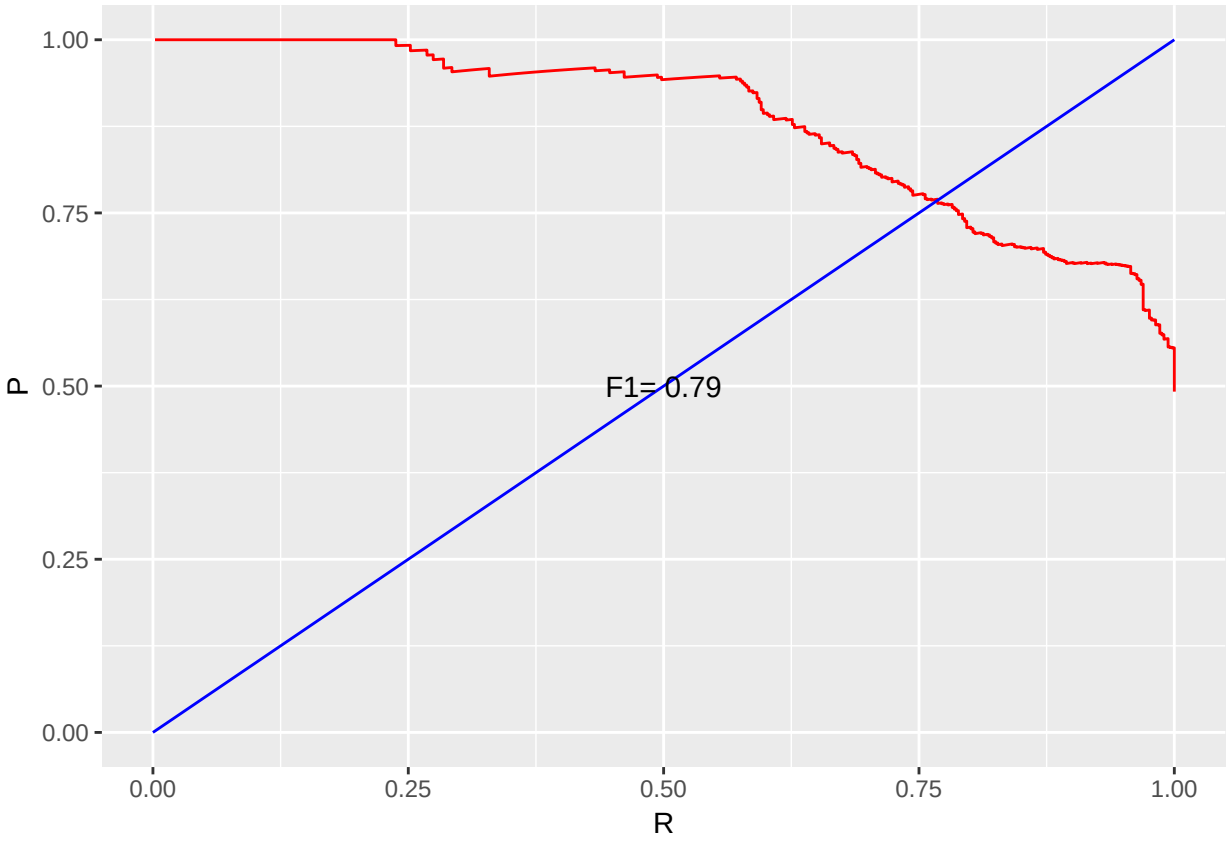

图 1: P-R 曲线

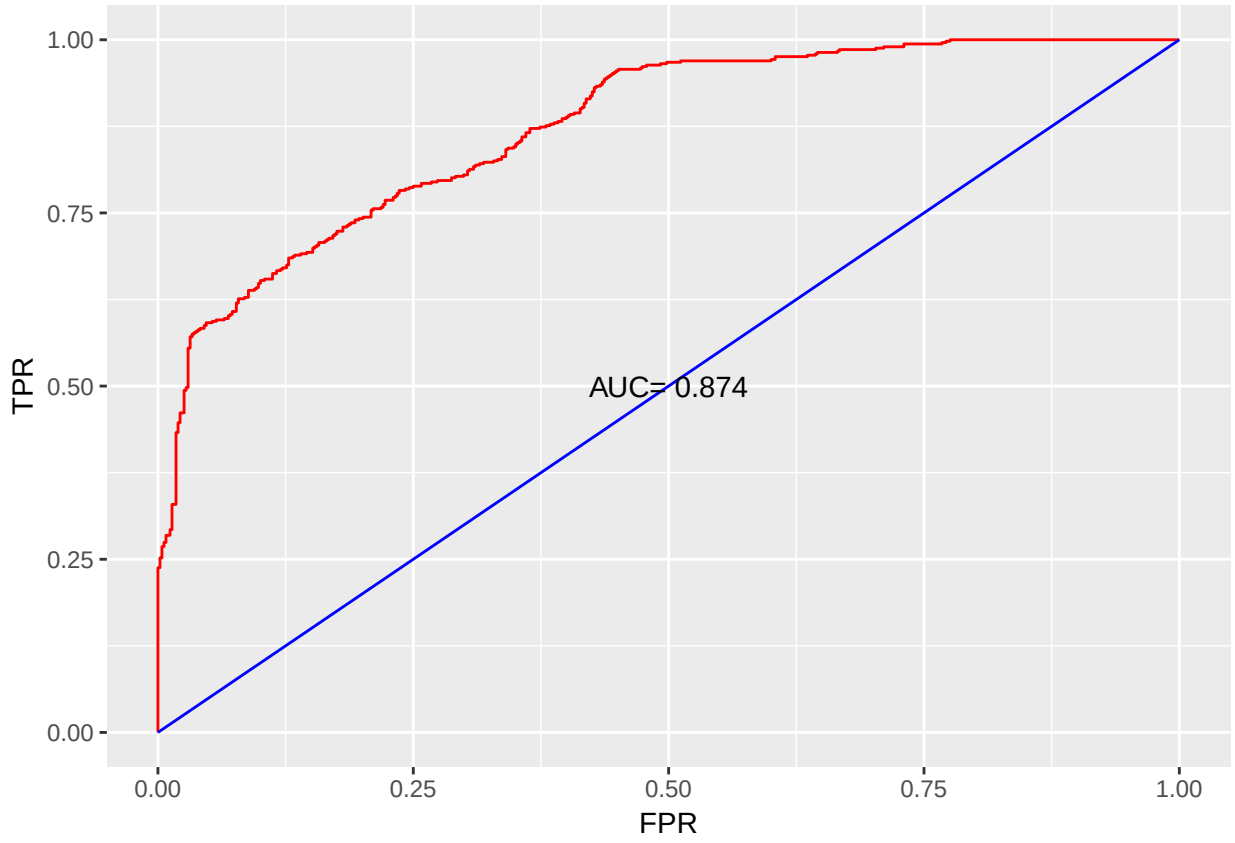

图 2: ROC 曲线