Natural Language Processing I

# 6. 评测方法

WU Xiaokun 吴晓堃

xkun.wu [at] gmail

2022/03/18

# 评测：精确率，召回率，F度量

# 回顾：两类误差

**假阳 False positives** (Type I errors)

- 查出错误信息
- 提高准确度 accuracy、精度 precision

**假阴 False negatives** (Type II errors)

- 没有查出正确信息
- 提高覆盖率 coverage、召回率 recall

# 回顾：两类误差

**假阳 False positives** (Type I errors)

- 查出错误信息
- 提高准确度 accuracy、精度 precision

**假阴 False negatives** (Type II errors)

- 没有查出正确信息
- 提高覆盖率 coverage、召回率 recall

**评测**：模型的预测性能，即预测正确率

# 预测与标注

首先考虑只有两个类的情况，如正负面评价

- **预测值**只有两种可能性：`+`, `-`
- **标注**也只有两种可能性：`T`, `F`

# 预测与标注

首先考虑只有两个类的情况，如正负面评价

- **预测值**只有两种可能性：`+`, `-`
- **标注**也只有两种可能性：`T`, `F`

评测时需要将预测与标注**两两组合**进行对比

- 假阳：错误的阳性；预测是`+`，但实际是`F`
- 假阴：错误的阴性；预测是`-`，但实际是`T`

# 混淆矩阵

**混淆 confusion 矩阵**：预测与标注对比的可视化

| | T | F |
|---|---|---|
| **P** | True Positive (TP) | False Positive (FP) |
| **N** | False Negative (FN) | True Negative (TN) |

- 两种正确的预测
  - TP：正确的阳性
  - TN：正确的阴性

# 混淆矩阵

**混淆 confusion 矩阵**：预测与标注对比的可视化

| | T | F |
|---|---|---|
| **P** | True Positive (TP) | False Positive (FP) |
| **N** | False Negative (FN) | True Negative (TN) |

- 两种正确的预测
  - TP：正确的阳性
  - TN：正确的阴性

评测：计算预测的正确率

- 一定与**正确的预测**（TP和TN）有关

# 准确度

**准确度 accuracy**：正确的预测所占百分比

| | T | F |
|---|---|---|
| **P** | TP | FP |
| **N** | FN | TN |

$$A = \frac{TP+TN}{TP+FP+TN+FN}$$

# 准确度

**准确度 accuracy**：正确的预测所占百分比

| | T | F | |
|---|---|---|---|
| **P** | TP | FP | |
| **N** | FN | TN | |
| | | | $A = \frac{TP+TN}{TP+FP+TN+FN}$ |

看起来像是很自然的度量

- 问题：当**类间数据不平衡**时会失效

# 不平衡的网页检索

假设网页检索返回$1 \cdot 10^6$条信息

- 但只有100条是相关的，比如**搜索

# 不平衡的网页检索

假设网页检索返回$1 \cdot 10^6$条信息

- 但只有100条是相关的，比如**搜索

构造一个假的分类器：对任何输入都输出“不相关”

- 准确度：$A = \frac{TP+TN}{TP+FP+TN+FN} = 99.99\%$

# 不平衡的网页检索

假设网页检索返回$1 \cdot 10^6$条信息

- 但只有100条是相关的，比如**搜索

构造一个假的分类器：对任何输入都输出“不相关”

- 准确度：$A = \frac{TP+TN}{TP+FP+TN+FN} = 99.99\%$

目标是发现**罕见情况**时，越罕见问题就越严重

- 例如：对罕见病检测都预测阴性

# 不平衡的网页检索

假设网页检索返回$1 \cdot 10^6$条信息

- 但只有100条是相关的，比如**搜索

构造一个假的分类器：对任何输入都输出“不相关”

- 准确度：$A = \frac{TP+TN}{TP+FP+TN+FN} = 99.99\%$

目标是发现**罕见情况**时，越罕见问题就越严重

- 例如：对罕见病检测都预测阴性

推论：**TP要更重要一些**，而TN会带来大量干扰信息

# 精确度、召回率

**精确度 precision**：预测为+的结果中，正确（实际确实是）的百分比

**召回率 recall**：标注为T的数据中，正确检测出来（预测阳性）的百分比

| | T | F | |
|---|---|---|---|
| **P** | TP | FP | $P = \frac{TP}{TP+FP}$ |
| **N** | FN | TN | |
| | $R = \frac{TP}{TP+FN}$ | | |

# 精确度、召回率

**精确度 precision**：预测为+的结果中，正确（实际确实是）的百分比

**召回率 recall**：标注为T的数据中，正确检测出来（预测阳性）的百分比

| | T | F | |
|---|---|---|---|
| **P** | TP | FP | $P = \frac{TP}{TP+FP}$ |
| **N** | FN | TN | |
| | $R = \frac{TP}{TP+FN}$ | | |

注意：两者都只关注TP，而**把容易带来干扰的TN排除**了

- 例如医疗检测中的TN就是健康人

# 精确度、召回率

**精确度 precision**：预测为+的结果中，正确（实际确实是）的百分比

**召回率 recall**：标注为T的数据中，正确检测出来（预测阳性）的百分比

| | T | F | |
|---|---|---|---|
| **P** | TP | FP | $P = \frac{TP}{TP+FP}$ |
| **N** | FN | TN | |
| | $R = \frac{TP}{TP+FN}$ | | |

注意：两者都只关注TP，而**把容易带来干扰的TN排除**了

- 例如医疗检测中的TN就是健康人

前面例子：$R = \frac{TP}{TP+FN} = \frac{0}{100} = 0\%$

# 分类问题应用

精确度高：阳性**置信度**高；召回率高：阳性**辨识力**强。

- 如何确定辨识力？选取**阈值**

# 分类问题应用

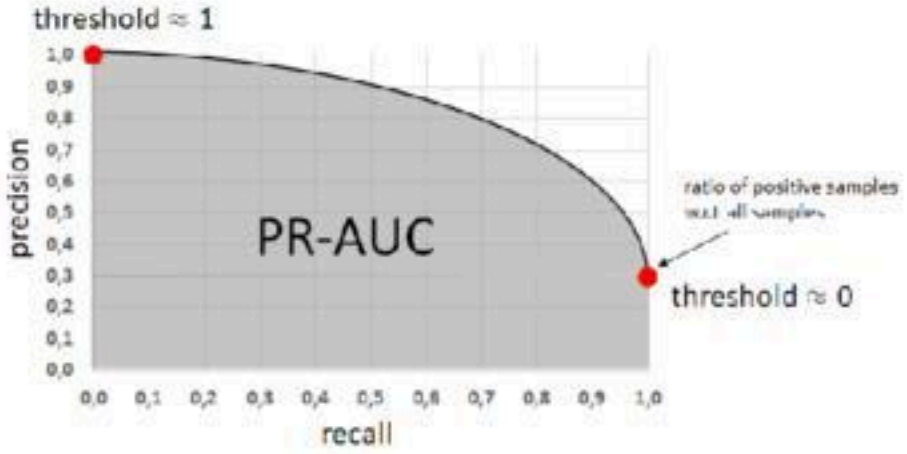
精确度高：阳性**置信度**高；召回率高：阳性**辨识力**强。

- 如何确定辨识力？选取**阈值**

例如分类真值：$[T, F, F]$；预测值：$[0.7, 0.3, 0.5]$

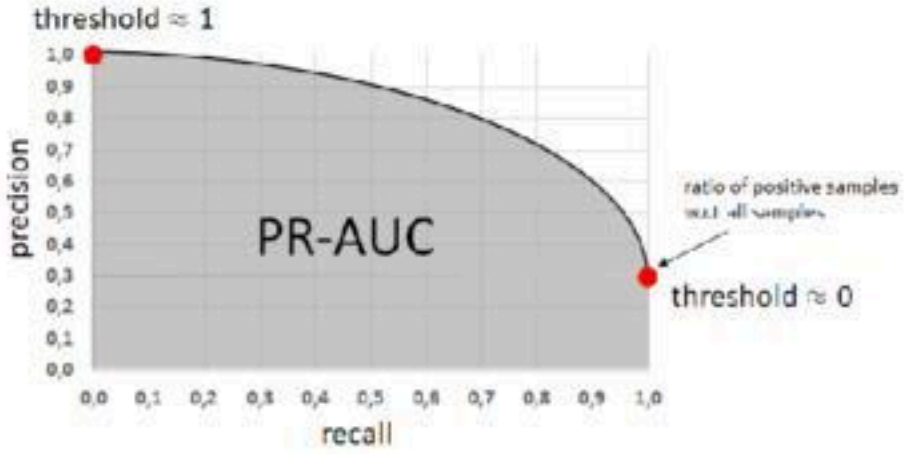
- 阈值$0.5$：$P = \frac{TP}{TP+FP} = 1/2$；$R = \frac{TP}{TP+FN} = 1$

# Precision-Recall Curve




- 阈值0：所有真值都检出；阈值接近1：所有真值都正确
  - 曲线越向右上角越好

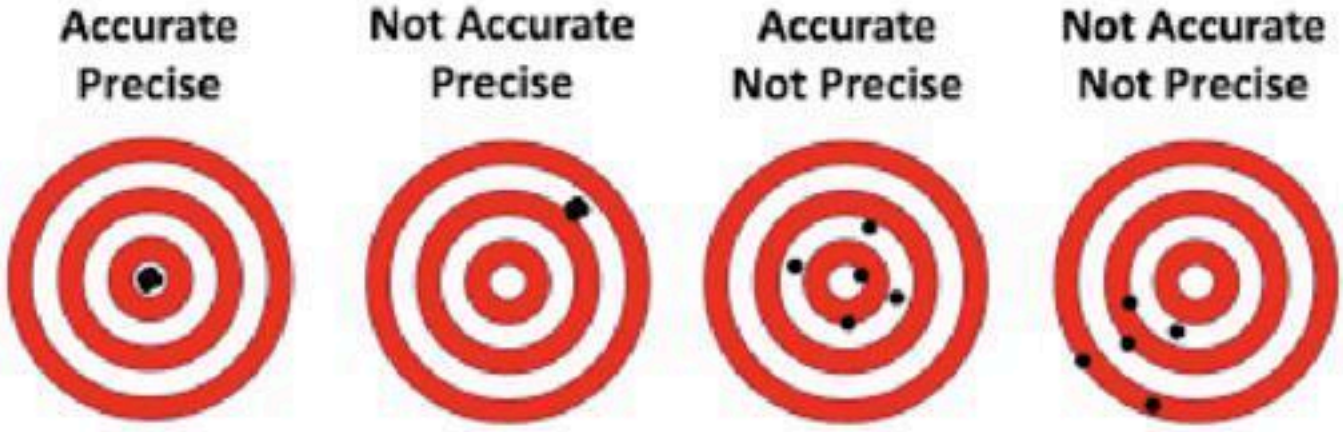
# Precision-Recall Curve




- 阈值0：所有真值都检出；阈值接近1：所有真值都正确
  - 曲线越向右上角越好

**Area Under Curve (AUC)**：面积越大越好

# 准确度、精确度




- TP：靶心附近；FP：偏离靶心

| | T | F | |
|---|---|---|---|
| P | TP | FP | $P = \frac{TP}{TP+FP}$ |
| N | FN | TN | |
| | | | $A = \frac{TP+TN}{TP+FP+TN+FN}$ |

# 单一度量

实际应用通常希望将两种度量（精确度、召回率）综合考虑

- 最常用的是加权**调和平均 harmonic mean**

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

# 单一度量

实际应用通常希望将两种度量（精确度、召回率）综合考虑

- 最常用的是加权**调和平均 harmonic mean**

$$F_\alpha = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

- 调和平均是**保守 conservative 度量**：（相比算数平均）较低的值权重更高

# 单一度量

实际应用通常希望将两种度量（精确度、召回率）综合考虑

- 最常用的是加权**调和平均 harmonic mean**

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

- 调和平均是**保守 conservative 度量**：（相比算数平均）较低的值权重更高

常用变换形式：$\beta^2 = \frac{1}{\alpha} - 1 \Rightarrow \beta^2 + 1 = \frac{1}{\alpha}$

- $\beta$的定义可以保证$\frac{1}{\alpha} - 1$非负

# 单一度量

实际应用通常希望将两种度量（精确度、召回率）综合考虑

- 最常用的是加权**调和平均 harmonic mean**

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

- 调和平均是**保守 conservative 度量**：（相比算数平均）较低的值权重更高

常用变换形式：$\beta^2 = \frac{1}{\alpha} - 1 \Rightarrow \beta^2 + 1 = \frac{1}{\alpha}$

- $\beta$的定义可以保证$\frac{1}{\alpha} - 1$非负

$$F_\alpha = \frac{\frac{1}{\alpha}}{\frac{1}{P} + (\frac{1}{\alpha} - 1)\frac{1}{R}} = \frac{\beta^2 + 1}{\frac{1}{P} + \beta^2\frac{1}{R}}$$

# F度量

**F度量 F-measure**：精确度、召回率的加权调和平均

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{\beta^2+1}{\frac{1}{P} + \beta^2\frac{1}{R}}$$

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

# F度量

**F度量 F-measure**：精确度、召回率的加权调和平均

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{\beta^2+1}{\frac{1}{P} + \beta^2\frac{1}{R}}$$

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

- $\beta > 1$：强调召回率；$\beta < 1$：强调精确度

# F度量

**F度量 F-measure**：精确度、召回率的加权调和平均

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{\beta^2+1}{\frac{1}{P} + \beta^2\frac{1}{R}}$$

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

- $\beta > 1$：强调召回率；$\beta < 1$：强调精确度

最常用的是$F_1$**度量**：$\beta = 1$

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

# Average Precision (AP)

精确度的加权和；权重：召回率增量

$$AP = \sum_{k=0}^{n-1} [R(k) - R(k+1)] * P(k)$$

- $R(n) = 0, P(n) = 1$

# Average Precision (AP)

精确度的加权和；权重：召回率增量

$$AP = \sum_{k=0}^{n-1} [R(k) - R(k+1)] * P(k)$$

- $R(n) = 0, P(n) = 1$

Mean Average Precision (mAP)：多类AP均值

$$AP = \frac{1}{n} \sum_{k=1}^{n} AP_k$$