Week14 参考資料

Week10：正規表現と評価計算（抜粋）

2. TF-IDF　文書中に含まれる単語の重要度を評価する手法

- 文書の特徴となる単語の抽出 (単語の重み付け)

TF(Term Frequency): $tf(t,d)$

各単語 t の文書内 d での出現頻度

頻出単語は特徴的

$$tf(t,d) = \frac{\text{文書}d\text{における単語}t\text{の出現頻度}}{\text{文書}d\text{における全単語の出現頻度}s\text{の和}} = \frac{n(t,d)}{\sum_{s\in d} n(s,d)}$$

IDF(Inverse Document Frequency): $idf(t)$

各単語 t が出現する文書数の逆数：逆文書頻度

$$idf(t) = \frac{\text{全文書数}N}{\text{単語}t\text{を含む文書数}} = \log\frac{N}{df(t)} + 1$$

他の文書にも出現するものは特徴的ではない:各単語のレア度

$df(t)$ ：索引語 t が出現する文書数

N ： 検索対象となる全文書数

2. TF-IDF

- 文章の特徴となる単語の抽出(単語の重み付け)

TF・IDF

$$tf \cdot idf = (\textbf{単語の出現頻度}) \times (\textbf{単語の逆文書頻度})$$
$$= tf(t,d) \times idf(t)$$

不要な語は前処理でストップワードリストを用いて除去しておく

ストップワードリスト

- 日本語：助詞（は，が，に，の，を…）や助動詞（する… ）
- 英語：冠詞(the, a…), 前置詞(at, on, with…)

情報検索後の重み付け手法でも利用されている

3. 代表的な評価指標

- 分類や検索結果などのモデル・システムの有効性を確認する

- 適合率：Precision (P)
- 再現率：Recall (R)
- F値：F-measure (F (調和平均))
- 正確度：Accuracy

▶ 0~1の値，もしくは，0～100%で表される.

どのモデルが1番よい精度か

3. 代表的な評価指標

Accuracy(正確度)

全てのデータのうち，正解，不正解が
正しく分類されたデータの割合

分類結果

Data		guessed positive	guessed negative
	positive	6 TP	1 FN
	negative	2 FP	5 TN

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

$$= \frac{6+5}{14}$$

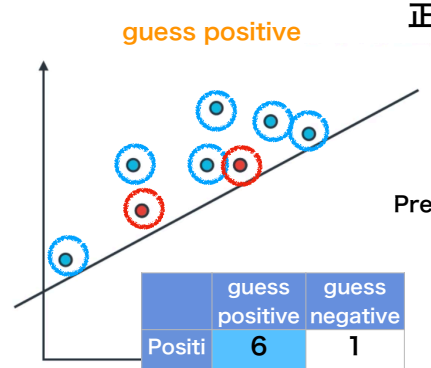
$$= 0.7857$$

Accuracyは78.57 %

3. 代表的な評価指標

Precision(適合率)

正しいと分類されたデータのうち，
正解データの割合．



	guess positive	guess negative
Positive	6 TP	1 FN
Negative	2 FP	5 TN

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$= \frac{6}{6 + 2}$$

$$= 0.75$$

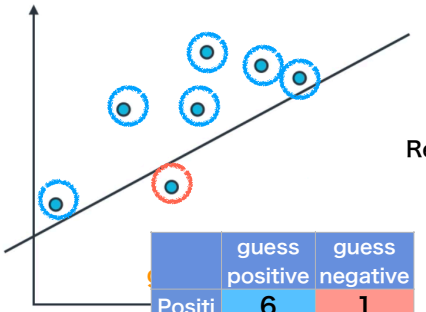
適合率は75 %

3. 代表的な評価指標

Recall(再現率)

正解データのうち，正しく分類された割合
どの程度，網羅されているか

coverall



	guess positive	guess negative
Positive	6 TP	1 FN
Negative	2 FP	5 TN

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

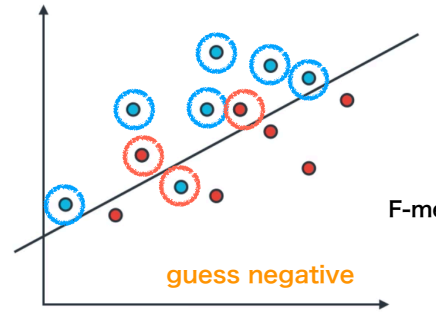
$$= \frac{6}{6 + 1} = 0.857$$

再現率は85.7 %

3. 代表的な評価指標

F-measure(F値)

PとRの調和平均
harmonic mean



Precision = 75 %
Recall = 85.7 %
Average = 80.35%

$$\text{F-measure} = \frac{2 \times P \times R}{P + R}$$

$$= \frac{2 \times 75 \times 85.7}{75 + 85.7}$$

$$= 0.8$$

F値は80 %