Chapter 4 Naïve Bayes and Sentiment Classification

9/9/20 12:37 PM

function TRAIN NAIVE BAYES(D, C) **returns** log $P(c)$ and log $P(w|c)$

for each class $c \in C$ # Calculate $P(c)$ terms
  N_{doc} = number of documents in D
  N_c = number of documents from D in class c
  $logprior[c] \leftarrow \log \frac{N_c}{N_{doc}}$
  $V \leftarrow$ vocabulary of D
  $bigdoc[c] \leftarrow$ **append**(d) **for** d $\in$ D **with** class c
  for each word w in V # Calculate $P(w|c)$ terms
    $count(w,c) \leftarrow$ # of occurrences of w in $bigdoc[c]$
    $loglikelihood[w,c] \leftarrow \log \frac{count(w,c) + 1}{\sum_{w' \text{ in } V} (count(w',c) + 1)}$
return *logprior*, *loglikelihood*, V

function TEST NAIVE BAYES(*testdoc*, *logprior*, *loglikelihood*, C, V) **returns** best c

for each class $c \in C$
  $sum[c] \leftarrow logprior[c]$
  for each position i in *testdoc*
    $word \leftarrow testdoc[i]$
    if $word \in V$
      $sum[c] \leftarrow sum[c] + loglikelihood[word,c]$
return $\text{argmax}_c\ sum[c]$

Training on Naïve Bayes

Prior Probability:

- number of documents in certain class / number of all documents

$$\hat{P}(c) = \frac{N_c}{N_{doc}}$$

Conditional probability of a word (i) in bag:

- count of word_i in doc of class / count of all words in doc of class

$$\hat{P}(w_i|c) = \frac{count(w_i,c)}{\sum_{w \in V} count(w,c)}$$

- Easy to get 0 here, so use smoothing
- Ignore unknown words and stop words

Evaluation

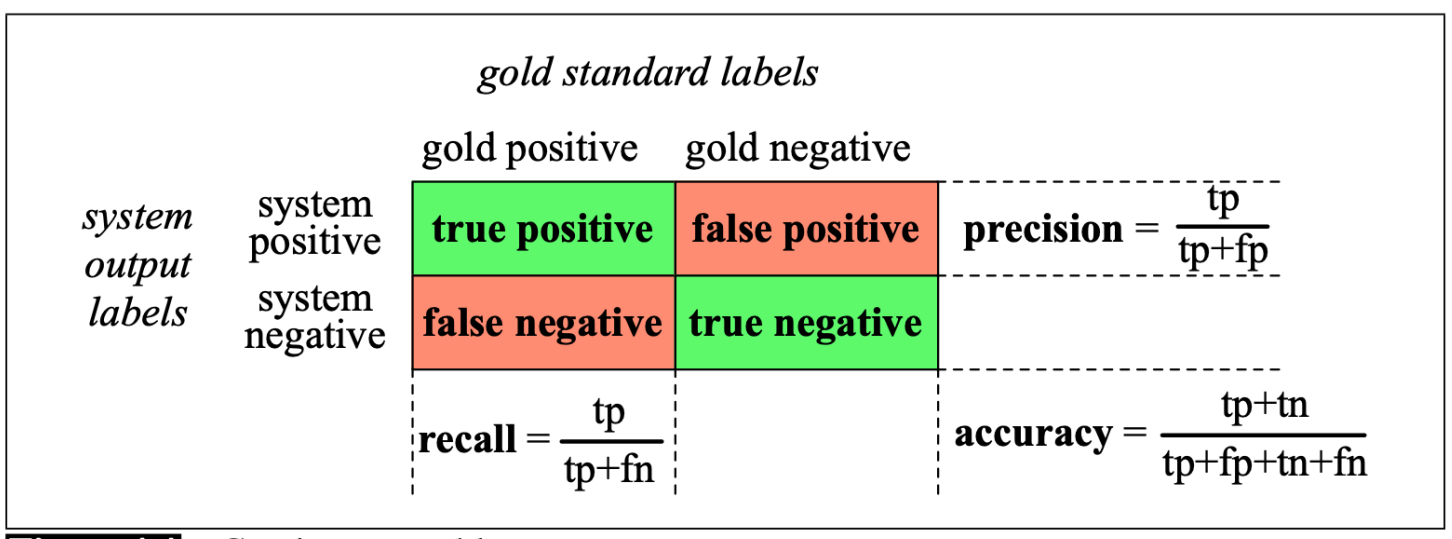

Figure 4.4 Contingency table

Precision

- Precision measures the percentage of the items that the system detected (i.e., the system labeled as positive) that are in fact positive (i.e., are positive according to the human gold labels)
- $\textbf{Precision} = \frac{\text{true positives}}{\text{true positives + false positives}}$

Recall

- Recall measures the percentage of items actually present in the input that were correctly identified by the system.
- $\textbf{Recall} = \frac{\text{true positives}}{\text{true positives + false negatives}}$

F-measure

- A single metric that incorporates both precision and recall
  - $F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}$
- Simplest: F-1 score:

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