Midterm 1 Resources

October 10th

1 Naive Bayes

$c_{NB} = \text{argmax}_{c \in C} P(c_j) \prod_{x \in X} P(x|c)$ where C is the set of classes and X is the set of features.

Prior: $p(c_j)$

Likelihood: $p(x|c)$

Add-1 Smoothed Naive Bayes: $\hat{P}(w_i|c) = \frac{count(w_i,c)+1}{(\sum_{w \in V} count(w,c))+|V|}$ where V is the vocabulary.

2 N-gram Models

$P(w_1 w_2 ... w_i) \approx \Pi_i P(w_i|w_{i-k} ... w_{i-1})$ where k is the context window.

Bigram Maximum Likelihood Estimates: $P(w_i|w_{i-1}) = \frac{count(w_{i-1},w_i)}{count(w_{i-1})}$

3 Metrics

Precision: $\frac{TP}{TP+FP}$ Recall: $\frac{TP}{TP+FN}$

where TP= true positives, FP = false positives, and FN = false negatives

4 Byte Pair Encoding

function BYTE-PAIR ENCODING(strings *C*, number of merges *k*) **returns** vocab *V*

V ← all unique characters in *C* # initial set of tokens is characters
for *i* = 1 **to** *k* **do** # merge tokens *k* times
  t_L, t_R ← Most frequent pair of adjacent tokens in *C*
  $t_{NEW} \leftarrow t_L + t_R$ # make new token by concatenating
  $V \leftarrow V + t_{NEW}$ # update the vocabulary
  Replace each occurrence of t_L, t_R in *C* with t_{NEW} # and update the corpus
return *V*