Hw3

Highlight

- Q1: run the NB learner in Mallet
- Q2: build a Multi-variate Bernoulli NB learner
- Q3: build a Multinomial NB learner

Q2

- build_NB1.sh training_data test_data prior_delta cond_prob_delta model_file sys_output > acc

- prior_delta: delta for calculating P(c).

- cond_prob_delta: delta for calculating P(f|c).

Model file

c1 P(c1) lg P(c1)   ## lg is 10-based

….

f1 c1   P(f1|c1) lg P(f1|c1)

f2 c1   P(f2|c1) lg P(f2|c1)

…

f1   c2   P(f1|c2) lg P(f1|c2)

f2   c2   P(f2|c2) lg P(f2|c2)

…

Sys_output

instanceName trueClass c_1 p_1 c_2 p_2 …

instanceName will be array:0, array:1, etc.

The (c_i, p_i) pairs should be sorted by the value of p_i.

$$p_i = P(c_i|x) = \frac{P(x|c_i)P(c_i)}{P(x)}$$

$$P(x) = \sum_i P(c_i, x) = \sum_i P(x|c_i)P(c_i)$$

The issue of underflow

$$p_i = P(c_i|x) = \frac{P(x|c_i)P(c_i)}{P(x)} = \frac{P(x,c_i)}{\sum_{c_i} P(x,c_i)}$$

$lgP(x,c_1)$ is -200, $lgP(x,c_2)$ is -201, $lgP(x,c_3)$ is -202.

What is p_i?

$$p_1 = \frac{10^{-200}}{10^{-200}+10^{-201}+10^{-202}} = \frac{1}{1+10^{-1}+10^{-2}} = 100/111 = 0.901$$

$$p_2 = \frac{10^{-1}}{1+10^{-1}+10^{-2}} = 10/111 = 0.09$$

$$p_3 = \frac{10^{-2}}{1+10^{-1}+10^{-2}} = 1/111 = 0.009$$

Efficiency issue: Ex 1

$$lg\ P(c)\prod_{k=1}^{|V|} P(w_k|c)^{N_{ik}}$$

$$= lgP(c) + \sum_{k=1}^{|V|} lg(P(w_k|c)^{N_{ik}})$$

$$= lgP(c) + \sum_{k=1}^{|V|} N_{ik}\ lg\ P(w_k|c)$$

Efficiency: Ex #2

$$P(d_i, c)$$

$$= P(c) \ \left(\prod_{w_k \in d_i} P(w_k|c)\right) \ \left(\prod_{w_k \notin d_i} (1 - P(w_k|c))\right)$$

$$= P(c) \ \left(\prod_{w_k \in d_i} P(w_k|c)\right) \ \frac{\prod_{w_k} (1-P(w_k|c))}{\prod_{w_k \in d_i} (1-P(w_k|c))}$$

$$= P(c) \prod_{w_k \in d_i} \frac{P(w_k|c)}{1-P(w_k|c)} \ \prod_{w_k} (1 - P(w_k|c))$$

Efficiency: Ex #3

Multinomial model:

Let $P(c_j \mid d_i) = 1$ if d_i has the label c_j
$= 0$ otherwise

$$P(w_t|c_j) = \frac{1 + \sum_{i=1}^{|D|} N_{it} P(c_j|d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N_{is} P(c_j|d_i)}$$

$$\text{Complexity: } O(|V| * |D| * |V| * |C|) = O(|V|^2 * |C| * |D|)$$

How to make it faster?

$Z(c_j) = 0$ for every c_j;

for each d_i

    Let c_j be the class label of d_i

    for each w_t that is present in d_i

        Let N_{it} be the number of times w_t appears in d_i

        $cnt(w_t, c_j) \ += N_{it}$

        $Z(c_j) \ += N_{it}$

for each c_j

    for each w_t

        $P(w_t|c_j) = \frac{1+cnt(w_t,c_j)}{|V|+Z(c_j)}$

Complexity: $O(|V| * |C| + |D| * avg(feat/doc))$