# CS5154/6054 Quiz 13 Key, 10/13/2022

**Exercise 13.15**

In the $\chi^2$ example on page 276 we have $|N_{11} - E_{11}| = |N_{10} - E_{10}| = |N_{01} - E_{01}| = |N_{00} - E_{00}|$. Show that this holds in general.

E11 = N ((N11 + N10)/N) ((N11 + N01)/N) = (N11 + N10)(N11 + N01)/N
N11 – E11 = (N11(N) - (N11 + N10)(N11 + N01))/N
= (N11(N11 + N01 + N10 + N00) - (N11 + N10)(N11 + N01))/N
= (N11N11 + N11N01 + N11N10 + N11N00 – N11N11 – N11N01 – N11N10 – N10N01) / N
= (N11N00 – N10N01)/N
N10 – E10 = (N10(N) - (N10 + N11)(N10 + N00))/N
= (N10(N11 + N01 + N10 + N00) - (N10 + N11)(N10 + N00))/N
= (N10N11 + N10N01 + N10N10 + N10N00 – N10N10 – N10N00 – N11N10 – N11N00) / N
= (N10N01 – N11N00)/N

Only two of the four are needed. Or even one if enough argument can be stated.

Question 2: Under what condition do we have E11 = (N11 + N01)(N11 + N10)/N = (N11 + N01)(N11 + N10)/(N11 + N01 + N10 + N00) = N11? (Hint: Try to multiply N to both sides and then cancel identical terms on both sides.)

E11 = N11 leads to (N11 + N01)(N11 + N10) = N11 N = N11(N11 + N01 + N10 + N00).

Multiplying out both sides: N11N11 + N11N10 + N01N11 + N01N10 = N11N11 + N11N01 + N11N10 + N11N00.

Cancelling three terms on either side that match terms on the other side, we get N01N10 = N11N00.

# CS5154/6054 Quiz 14 Key, 10/18/2022

▶ **Table 13.10** Data for parameter estimation exercise.

| | docID | words in document | in $c = China$? |
|---|---|---|---|
| training set | 1 | Taipei Taiwan | yes |
| | 2 | Macao Taiwan Shanghai | yes |
| | 3 | Japan Sapporo | no |
| | 4 | Sapporo Osaka Taiwan | no |
| test set | 5 | Taiwan Taiwan Sapporo | ? |

**Exercise 13.9**

Based on the data in Table 13.10, (i) estimate a multinomial Naive Bayes classifier, (ii) apply the classifier to the test document, (iii) estimate a Bernoulli NB classifier, (iv) apply the classifier to the test document. You need not estimate parameters that you don't need for classifying the test document.

Only do (iii) and (iv) following the algorithms below. You need N (= 4), $N_c$ (= 2), $N_{ct}$ (fill the table with different c and t), and P(t|c) (or condprob[t][c] in algorithms, fill the table). P(c|d5) is score[c] in algorithm when d is d5. You need $V_d$, too, for d = d5. You may consider P(t|c) as $p_t$ or $u_t$ in BIM or Quiz 10. But the data is different. You are not asked to sum the log score. You are asked to follow Example 13.2 (Slides 10/18/12-13) to multiply P(t|c) or 1 – P(t|c) for each of the two classes c.

```
TrainBernoulliNB(C, D)
V ← ExtractVocabulary(D)
N ← CountDocs(D)
for each c ∈ C
do N_c ← CountDocsInClass(D, c)
   prior[c] ← N_c / N
   for each t ∈ V
   do N_ct ← CountDocsInClassContainingTerm(D, c, t)
      condprob[t][c] ← (N_ct + 1) / (N_c + 2)
return V, prior, condprob
```

```
ApplyBernoulliNB(C, V, prior, condprob, d)
V_d ← ExtractTermsFromDoc(V, d)
for each c ∈ C
do score[c] ← log prior[c]
   for each t ∈ V
   do if t ∈ V_d
        then score[c] += log condprob[t][c]
        else score[c] += log(1 − condprob[t][c])
return arg max_{c∈C} score[c]
```

| t | Nct (c = China) | P(t\|c = China) | Nct (c != China) | P(t\|c != China) |
|---|---|---|---|---|
| Taiwan | 2 | (2+1)/(2+2)=3/4 | 1 | (1+1)/(2+2)=2/4 |
| Sapporo | 0 | 1/4 | 2 | 3/4 |
| Taipei | 1 | 2/4 | 0 | 1/4 |
| Macao | 1 | 2/4 | 0 | 1/4 |
| Shanghai | 1 | 2/4 | 0 | 1/4 |
| Japan | 0 | 1/4 | 1 | 2/4 |
| Osaka | 0 | 1/4 | 1 | 2/4 |
| P(c = China\|d5) is proportional to | | (1/2)(3/4)(1/4)(1-2/4)(1-2/4)(1-2/4)(1-1/4)(1-1/4) =3*1*2*2*2*3*3/(2*$4^7$) | | |
| P(c != China\|d5) is proportional to | | (1/2)(2/4)(3/4)(1-1/4)(1-1/4)(1-1/4)(1-2/4)(1-2/4) =2*3*3*3*3*2*2/(2*$4^7$) | | |
| d5's classification is | | P(c=China\|d5)/P(c!=China\|d5)=1/3 < 1 (not China) | | |

# CS5154/6054 Quiz 15 Key, 10/20/2022

▶ **Table 13.10** Data for parameter estimation exercise.

| | docID | words in document | in $c = China$? |
|---|---|---|---|
| training set | 1 | Taipei Taiwan | yes |
| | 2 | Macao Taiwan Shanghai | yes |
| | 3 | Japan Sapporo | no |
| | 4 | Sapporo Osaka Taiwan | no |
| test set | 5 | Taiwan Taiwan Sapporo | ? |

**Exercise 13.9**

Based on the data in Table 13.10, (i) estimate a multinomial Naive Bayes classifier, (ii) apply the classifier to the test document, (iii) estimate a Bernoulli NB classifier, (iv) apply the classifier to the test document. You need not estimate parameters that you don't need for classifying the test document.

Only do (i) and (ii) following the algorithms below. You need N (= 4), $N_c$ (= 2), $T_{ct}$ (fill the table with different c and t), and P(t|c) (or condprob[t][c] in algorithms, fill the table). P(c|d5) is proportional to exp(score[c]) in algorithm when d is d5. You need W, too, for d = d5. Mimic Example 13.1. A number of tokens in the table below may not be used by ApplyMultinomialNB on d5. Also (13.7) gives an easier formula for line 10 of TrainNultinomialNB that requires B.

```
TRAINMULTINOMIALNB(C, D)
 1  V ← EXTRACTVOCABULARY(D)
 2  N ← COUNTDOCS(D)
 3  for each c ∈ C
 4  do N_c ← COUNTDOCSINCLASS(D, c)
 5     prior[c] ← N_c/N
 6     text_c ← CONCATENATETEXTOFALLDOCSINCLASS(D, c)
 7     for each t ∈ V
 8     do T_ct ← COUNTTOKENSOFTERM(text_c, t)
 9     for each t ∈ V
10     do condprob[t][c] ← (T_ct + 1) / Σ_t'(T_ct' + 1)
11  return V, prior, condprob
```

```
APPLYMULTINOMIALNB(C, V, prior, condprob, d)
1  W ← EXTRACTTOKENSFROMDOC(V, d)
2  for each c ∈ C
3  do score[c] ← log prior[c]
4     for each t ∈ W
5     do score[c] += log condprob[t][c]
6  return arg max_{c∈C} score[c]
```

▶ **Figure 13.2** Naive Bayes algorithm (multinomial model): Training and testing.

| t | $T_{ct}$ (c = China) | P(t\|c = China) | $T_{ct}$ (c = noChina) | P(t\|c = noChina) |
|---|---|---|---|---|
| Taiwan | 2 | (2+1)/(5+7)=1/4 | 1 | (1+1)/(5+7)=1/6 |
| Sapporo | 0 | (0+1)/(5+7)=1/12 | 2 | (2+1)/(5+7)=1/4 |
| Taipei | | | | |
| Macao | | | | |
| Shanghai | | | | |
| Japan | | | | |
| Osaka | | | | |
| P(c = China\|d5) is proportional to | | $(1/2)(3/12)^2(1/12)$ proportional to 9 (0.0026) | | |
| P(c = noChina\|d5) is proportional to | | $(1/2)(2/12)^2(3/12)$ proportional to 12 (0.0035) | | |
| d5's classification is | | c = noChina | | |

# CS5154/6054 Quiz 16 Key, 10/25/2022

TRAINROCCHIO($\mathbb{C}, \mathbb{D}$)
**for each** $c_j \in \mathbb{C}$
**do** $D_j \leftarrow \{d : \langle d, c_j \rangle \in \mathbb{D}\}$
$\quad \vec{\mu}_j \leftarrow \frac{1}{|D_j|} \sum_{d \in D_j} \vec{v}(d)$
**return** $\{\vec{\mu}_1, \ldots, \vec{\mu}_J\}$

APPLYROCCHIO($\{\vec{\mu}_1, \ldots, \vec{\mu}_J\}, d$)
**return** $\arg\min_j |\vec{\mu}_j - \vec{v}(d)|$

APPLYLINEARCLASSIFIER($\vec{w}, b, \vec{x}$)
$score \leftarrow \sum_{i=1}^{M} w_i x_i$
**if** $score > b$
$\quad$ **then return** 1
$\quad$ **else return** 0

APPLY-KNN($\mathbb{C}, \mathbb{D}', k, d$)
$S_k \leftarrow$ COMPUTENEARESTNEIGHBORS($\mathbb{D}', k, d$)
**for each** $c_j \in \mathbb{C}$
**do** $p_j \leftarrow |S_k \cap c_j| / k$
**return** $\arg\max_j p_j$

Some basic arithmetic shows that this corresponds to a linear classifier with normal vector $\vec{w} = \vec{\mu}(c_1) - \vec{\mu}(c_2)$ and $b = 0.5 * (|\vec{\mu}(c_1)|^2 - |\vec{\mu}(c_2)|^2)$ (Exer-

► **Table 13.10** Data for parameter estimation exercise.

| | docID | words in document | in $c = China$? |
|---|---|---|---|
| training set | 1 | Taipei Taiwan | yes |
| | 2 | Macao Taiwan Shanghai | yes |
| | 3 | Japan Sapporo | no |
| | 4 | Sapporo Osaka Taiwan | no |
| test set | 5 | Taiwan Taiwan Sapporo | ? |

The following are the normalized tf-idf vectors representing the documents. Compute the centroids µ1 and µ2 for the classes "China" and "not China", and w

| | Japan | Macao | Osaka | Sapporo | Shanghai | Taipei | Taiwan |
|---|---|---|---|---|---|---|---|
| D1 | 0 | 0 | 0 | 0 | 0 | 0.98 | 0.20 |
| D2 | 0 | 0.70 | 0 | 0 | 0.70 | 0 | 0.15 |
| D3 | 0.89 | 0 | 0 | 0.45 | 0 | 0 | 0 |
| D4 | 0 | 0 | 0.88 | 0.44 | 0 | 0 | 0.18 |
| D5 | 0 | 0 | 0 | 0.88 | 0 | 0 | 0.48 |
| µ1 | 0 | 0.35 | 0 | 0 | 0.35 | 0.49 | 0.18 |
| µ2 | 0.45 | 0 | 0.44 | 0.45 | 0 | 0 | 0.09 |
| w=µ1-µ2 | -0.45 | 0.35 | -0.44 | -0.45 | 0.35 | 0.49 | 0.09 |

| µ1 – D5 | = 1.161   | µ2 – D5 | = 0.856
Dot product between w and D5 = -0.353
b = ½ (|µ1|^2 - |µ2|^2) = (0.5174 – 0.6067)/2 = -0.044
classification of D5: not China

Fill out the following table of cosine similarities between D5 and documents in the training set.

| | D1 | D2 | D3 | D4 |
|---|---|---|---|---|
| cos(D5, Di) | 0.48*0.20=0.096 | 0.48*0.15=0.072 | 0.88*0.45=0.396 | 0.482 |

What is the class assignment for D5 when 1NN is used with cosine similarity?
D4 is the nearest sample and the class assignment is not China
What is the class assignment for D5 when 3NN is used? Why?
D1, D3, and D4 are 3NN and 2/3 is p for not China and 1/3 is for China. class assignment is not China.