WSM Assignment 3: Text classification.

何子安 111753229 資碩計一

1. Feature Selection

Following table, which there are four terms' frequency for the class Coffee in the first 100,000 documents of Return-RCV1: (select two of these four terms)

term	N_{00}	N_{01}	N_{10}	N_{11}
brazil	98,012	102	1835	51
council	96,322	133	3525	20
producers	98,524	119	1118	34
roasted	99,824	143	23	10

a) Chi-square Test

$$\text{brazil}: \chi^2(\mathbb{D}, t, c) = \frac{(N_{11}+N_{10}+N_{01}+N_{00}) \times (N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11}+N_{01}) \times (N_{11}+N_{10}) \times (N_{10}+N_{00}) \times (N_{01}+N_{00})}$$

$$= 818.93869$$

$$\text{council}: \chi^2(\mathbb{D}, t, c) = 40.674125$$

$$\text{producers}: \chi^2(\mathbb{D}, t, c) = 569.069999$$

$$\text{roasted}: \chi^2(\mathbb{D}, t, c) = 1964.29329$$

∴ it will be brazil & roasted.

b) Mutual Information

$$\text{brazil}: I(U;C) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1\cdot}N_{\cdot 1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0\cdot}N_{\cdot 1}} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1\cdot}N_{\cdot 0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0\cdot}N_{\cdot 0}}$$

$$= \frac{98012}{100000}\log_2\frac{100000\times 98012}{(102+98012)(1835+98012)} + \frac{102}{100000}\log_2\frac{100000\times 102}{(102+98012)(51+102)}$$

$$+ \frac{1835}{100000}\log_2\frac{100000\times 1835}{(51+1835)(1835+98012)} + \frac{51}{100000}\log_2\frac{100000\times 51}{(51+1835)(51+102)}$$

$$= 0.0015537$$

$$\text{council}: I(U;C) = 0.0001774$$

$$\text{producers}: I(U;C) = 0.00104799995$$

$$\text{roasted}: I(U;C) = 0.0006485$$

∴ it will be brazil & producers.

c) TF-IDF

$$\text{brazil}: tf(w,d) = \frac{\#\text{ of } w \text{ in } d}{\text{total \# of words in } d} = \frac{51}{100000} = 0.0051$$

$$tf(\text{council}, d) = \frac{20}{100000} = 0.0020$$

$$tf(\text{producers}, d) = \frac{34}{100000} = 0.0034$$

$$tf(\text{roasted}, d) = \frac{10}{100000} = 0.001$$

∴ it will be brazil & producers.

2. Evaluations of Text classification.

class1:	truth: yes	truth: no
call: yes	80	10
call: no	20	890

class2	truth: yes	truth: no
call: yes	20	40
call: no	60	880

a) Macro - averaged precision

$$\text{precision}_{\text{class1}} = \frac{80}{80+10} = 0.88889$$

$$\text{precision}_{\text{class2}} = \frac{20}{20+40} = 0.33333$$

$$\frac{\text{call \& truth: yes}}{\text{call: yes}}$$

$$\text{macro-averaged precision} = (\text{precision}_{\text{class1}} + \text{precision}_{\text{class2}}) / N_{\text{class}}$$

$$= (0.88889 + 0.33333) / 2$$

$$= 0.611111$$

b) Micro - averaged Precision

$$\text{Micro-averaged Precision} = \frac{TP_{\text{class1}} + TP_{\text{class2}}}{TP_{\text{class1}} + TP_{\text{class2}} + FP_{\text{class1}} + FP_{\text{class2}}}$$

$$= \frac{80+20}{80+20+10+40}$$

$$= \frac{100}{150}$$

$$= 0.66667$$

3. Vector Space Classification

a) Show the decision boundary / surface of Rocchio classification with two centroids (c1, c2). Specify the derivation in detail.

for Rocchio Classfier, while training documents in each category, compute a prototype vector by summing the vectors of the training documents in the category. (use cosine similarity to assign test documents to the category with the closest prototype vector.)

the centroid in Rocchio is defined as $\vec{\mu}(c) = \frac{1}{|D_c|} \sum_{d \in D_c} \vec{v}(d)$.

D_c is the set of all documents that belong to class c.

$v(d)$ is the vector space representation of d.

* the boundary between two classes in Rocchio classification is the set of points with equal distance from the two centroids.

the line / hyperplane can be defined as: $\sum_{i=1}^{M} w_i d_i = b$.

b will be the distance from the origin.

w_i will be the vector orthogonal to the hyperplane.

so, for Rocchio, we may set: (for two centroids c_1, c_2)

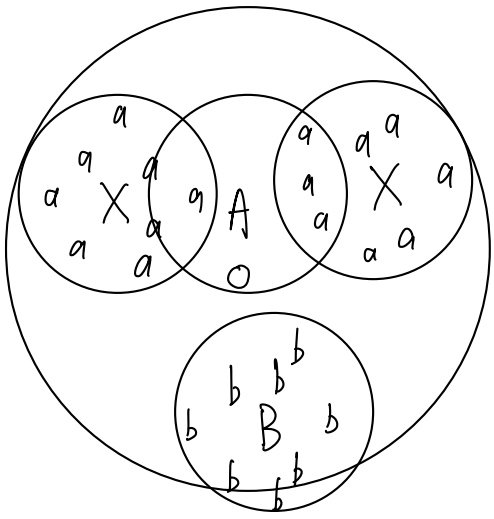
$$\vec{w} = \vec{\mu}(c_1) - \vec{\mu}(c_2)$$

$$b = 0.5 \times \left(|\vec{\mu}(c_1)|^2 - |\vec{\mu}(c_2)|^2 \right)$$

b) Describe what would happen when Rocchio classification method encounters the situation of multimodel classes.

Rocchio cannot handle nonconvex, multimodel classes.

Suppose we have a dataset that show as below:



A is the centroid of a's, B is the centroid of b's.

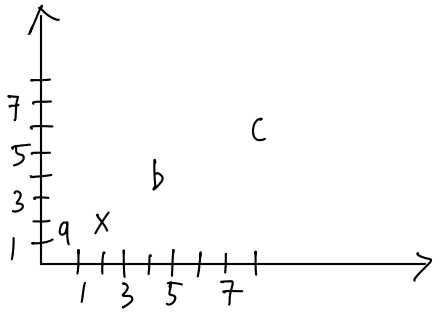
There is a point O closer to A than B.

But O is a better fit for the b class.

But Rocchio only have one prototype, which in this case, A is a multimodal class with two prototype

4. Vector Space classification

There are three points: $a=(0.5, 1.5)$, $b=(4,4)$, $c=(8,6)$, they representing a specific class, and a testing point $x=(2,2)$.



a) inner product similarity

$dot(a,x) = [0.5, 1.5][2,2] = 4$

$dot(b,x) = [4,4][2,2] = 16$ $\therefore \vec{c}$ will be the most similar.

$dot(c,x) = [8,6][2,2] = 28$

b) cosine similarity

$$\cos(a, x) = (a \times x) / (\|a\| \times \|x\|)$$

$$= [(0.5 \times 2) + (1.5 \times 2)] / (\sqrt{0.5^2 + 1.5^2}\ \sqrt{2^2 + 2^2})$$

$$= 0.8944$$

$$\cos(b, x) = 0.9999$$

$$\cos(c, x) = 0.9899$$

$\therefore \vec{b}$ will be the most similar.

c) Euclidean distance

$$d(a, x) = \sqrt{(x_1 - a_1)^2 + (x_2 - a_2)^2}$$

$$= 1.581139$$

$$d(b, x) = 2.82843$$

$$d(c, x) = 7.21111$$

$\therefore \vec{a}$ will be the most similar.