1 Naive Bayes for Text Categorization (10 points)

Given the following short documents (1-8), each labeled with a class:

1. banana, carrot, cucumber, pea : **vegetable**
2. school, pea, rose, lily, basket : **flower**
3. banana, pea, potato, lotus : **vegetable**
4. hibiscus, grape, potato, mango, apple : **fruit**
5. hibiscus, lotus, lily, apple, banana : **fruit**
6. rose, hibiscus, banana, rose : **flower**
7. rose, rose, rose, cucumber : **flower**
8. carrot, mango, grape, rose : **fruit**

And documents:

1. D1 : rose, lily, apple, carrot
2. D2 : pea, pea, lotus, grape

Compute the most likely class for D1 and D2. Assume a Naive Bayes classifier and use add-lambda smoothing (with lambda = 0.1) for the likelihood.

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

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Conditional Probabilities:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

Choosing a class:

$P(c|d5) \propto 3/4 * (3/7)^3 * 1/14 * 1/14 \approx 0.0003$

$P(j|d5) \propto 1/4 * (2/9)^3 * 2/9 * 2/9 \approx 0.0001$

20 Sep 2018

Naive-Bayes

$$\hat{P}(w|c) = \frac{count(w,c) + \lambda}{count(c) + \lambda|V|}$$ with add-λ smoothing

$$\hat{P}(c) = \frac{N_c}{N} \qquad P(c|D) = P(c) \prod_i P(w_i|c)$$

V = vocabulary = { w : w ∈ some document }

N = # documents

N_c = # documents classified c

V	vegetable	flower	fruit	
banana	11	1	1	4
carrot	1		1	2
cucumber	1	1		2
pea	11	1		3
school		1		1
rose		1111	1	7
lily		1	1	2
basket		1		1
potato	1		1	2
lotus	1		1	2
hibiscus		1	11	3
grape			11	2
mango			11	2
apple			11	2
	8	13	14	35 / 35

∩ ALT - 239

$P(A|B) > P(A)$

$> \frac{P(A|B)\,P(B)}{P(B|A)}$

$15 > \frac{10+2}{3}$

$3 \cdot 15 > 10+2$

$3 > \frac{10+2}{15}$

median and quartiles in Python

$\hat{P}$ (c): vegetable $= \frac{2}{8}$ flower $= \frac{3}{8}$ fruit $= \frac{3}{8}$

$\hat{P}$ (w/ vegetable)

rose	$\frac{0+0.1}{8+0.1(14)}$	$= \frac{0.1}{9.4}$	$= \frac{1}{94}$
lily	$\frac{0+0.1}{8+0.1(14)}$	$= \frac{0.1}{9.4}$	$= \frac{1}{94}$
apple	$\frac{0+0.1}{8+0.1(14)}$	$= \frac{0.1}{9.4}$	$= \frac{1}{94}$
carrot	$\frac{1+0.1}{8+0.1(14)}$	$= \frac{1.1}{9.4}$	$= \frac{11}{94}$
pea	$\frac{2+0.1}{8+0.1(14)}$	$= \frac{2.1}{9.4}$	$= \frac{21}{94}$
lotus	$\frac{1+0.1}{8+0.1(14)}$	$\frac{1.1}{9.4}$	$= \frac{11}{94}$
grape	$\frac{0+0.1}{8+0.1(14)}$	$= \frac{0.1}{9.4}$	$= \frac{1}{94}$

D1 = rose lily apple carrot

D2 = pea pea lotus grape

P(w|flower)

rose $\frac{6+0.1}{13+0.1(14)} = \frac{6.1}{14.4} = \frac{61}{144}$

lily $\frac{1+0.1}{13+0.1(14)} = \frac{1.1}{14.4} = \frac{11}{144}$

apple $\frac{0+0.1}{13+0.1(14)} = \frac{0.1}{14.4} = \frac{1}{144}$

carrot $\frac{0+0.1}{13+0.1(14)} = \frac{0.1}{14.4} = \frac{1}{144}$

pea $\frac{1+0.1}{13+0.1(14)} = \frac{1.1}{144} = \frac{11}{144}$

lotus $\frac{0+0.1}{13+0.1(14)} = \frac{0.1}{14.4} = \frac{1}{144}$

grape $\frac{0+0.1}{13+0.1(14)} = \frac{0.1}{14.4} = \frac{1}{144}$

P(w|fruit)

rose $\frac{1+0.1}{14+0.1(14)} = \frac{1.1}{15.4} = \frac{11}{154}$

lily $\frac{1+0.1}{14+0.1(14)} = \frac{1.1}{15.4} = \frac{11}{154}$

apple $\frac{2+0.1}{14+0.1(14)} = \frac{2.1}{15.4} = \frac{21}{154}$

carrot $\frac{1+0.1}{14+0.1(14)} = \frac{1.1}{15.4} = \frac{11}{154}$

pea $\frac{0+0.1}{14+0.1(14)} = \frac{0.1}{15.4} = \frac{1}{154}$

lotus $\frac{1+0.1}{14+0.1(14)} = \frac{1.1}{15.4} = \frac{11}{154}$

grape $\frac{2+0.1}{14+0.1(14)} = \frac{2.1}{15.4} = \frac{21}{154}$

$p(\text{vegetable}|D1) = \frac{1}{4} \cdot \frac{1}{94} \cdot \frac{1}{94} \cdot \frac{1}{94} \cdot \frac{11}{94} = 3.52 \times 10^{-8}$

* $p(\text{vegetable}|D2) = \frac{1}{4} \cdot \left(\frac{21}{94}\right)^2 \cdot \frac{11}{94} \cdot \frac{1}{94} = 1.55 \times 10^{-5}$

$p(\text{flower}|D1) = \frac{3}{8} \cdot \frac{61}{144} \cdot \frac{11}{144} \cdot \frac{1}{144} \cdot \frac{1}{144} = 5.85 \times 10^{-7}$

$p(\text{flower}|D2) = \frac{3}{8} \cdot \left(\frac{11}{144}\right)^2 \cdot \frac{1}{144} \cdot \frac{1}{144} = 1.06 \times 10^{-7}$

* $p(\text{fruit}|D1) = \frac{3}{8} \cdot \frac{11}{154} \cdot \frac{11}{154} \cdot \frac{21}{154} \cdot \frac{11}{154} = 1.86 \times 10^{-5}$

$p(\text{fruit}|D2) = \frac{3}{8} \cdot \left(\frac{1}{154}\right)^2 \cdot \frac{11}{154} \cdot \frac{21}{154} = 1.54 \times 10^{-7}$

D1 is fruit 1.86×10^{-5}

D2 is vegetable 1.55×10^{-5}

(1/4)*(1/94)*(1/94)*(1/94)*(11/94)	3.522259E-08
(1/4)*((21/94)^2)*(11/94)*(1/94)	1.553316E-05
(3/8)*(61/144)*(11/144)*(1/144)*(1/144)	5.851993E-07
(3/8)*((11/144)^2)*(1/144)*(1/144)	1.055277E-07
(3/8)*(11/154)*(11/154)*(21/154)*(11/154)	1.86357E-05
(3/8)*((1/154)^2)*(11/154)*(21/154)	1.540141E-07