ELEC 4800/5800 Special Topics
Homework: Naïve Bayes

Name: Charles Bollig

[Problem 1] – Based on the data in Table below, (i) estimate a multinomial Naive Bayes classifier, (ii) apply the classifier to the test document.

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

Total words = 7

Training Set	Words	In China
1	Taipei Taiwan	Yes
2	Macao Taiwan Shanghai	Yes
3	Japan Sapporo	No
4	Sapporo Osaka Taiwan	No

Test Set	Words	In China
5	Taiwan Taiwan Sapporo	???

$P(Taipei|yes) = \frac{1+1}{5+7} = \frac{2}{12} = \frac{1}{6}$

$P(Taipei|no) = \frac{0+1}{5+7} = \frac{1}{12}$

$P(Taiwan|yes) = \frac{2+1}{5+7} = \frac{3}{12} = \frac{1}{4}$

$P(Taiwan|no) = \frac{1+1}{5+7} = \frac{1}{6}$

$P(Macao|yes) = \frac{1+1}{12} = \frac{1}{6}$

$P(Macao|no) = \frac{1}{12}$

$P(Shanghai|yes) = \frac{2}{12} = \frac{1}{6}$

$P(Shanghai|no) = \frac{1}{12}$

$P(Japan|yes) = \frac{1}{12}$

$P(Japan|no) = \frac{2}{12} = \frac{1}{6}$

$P(Sapporo|yes) = \frac{1}{12}$

$P(Sapporo|no) = \frac{3}{12} = \frac{1}{4}$

$P(Osaka|yes) = \frac{1}{12}$

$P(Osaka|no) = \frac{2}{12} = \frac{1}{6}$

$$P(yes) = \frac{1}{2}, \qquad P(no) = \frac{1}{2}$$

$$\boldsymbol{P(yes|test\_words) = \frac{1}{2} \times (2)\left(\frac{1}{4}\right) \times (1)\left(\frac{1}{12}\right) = \frac{1}{24}}$$

$$\boldsymbol{P(no|test\_words) = \frac{1}{2} \times (2)\left(\frac{1}{6}\right) \times (1)\left(\frac{1}{4}\right) = \frac{1}{24}}$$

[Problem 2] –

i. Hand-calculate the sufficient parameters for Naive Bayes using the data in Figure 1, that is, the prior class probabilities and the conditional probabilities for all of the symbols.
ii. Using these, calculate the most likely class (1 or -1) for the unlabeled example (currently labeled '???').

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

$$P(square|1) = \frac{8+1}{12+3} = \frac{9}{15} = \frac{3}{5}$$

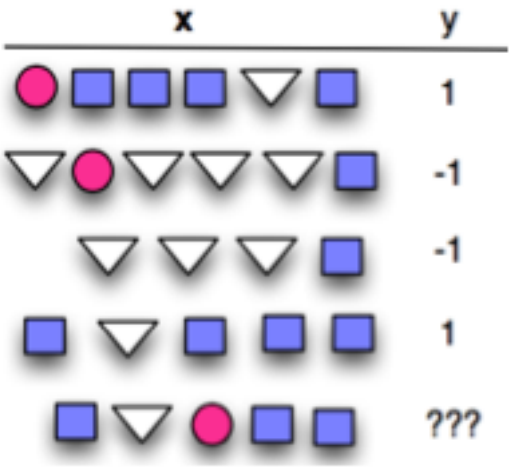
$$P(square|-1) = \frac{2+1}{10+3} = \frac{3}{13}$$

$$P(triangle|1) = \frac{2+1}{12+3} = \frac{3}{15} = \frac{1}{5}$$

$$P(triangle|-1) = \frac{7+1}{10+3} = \frac{7}{13}$$

$$P(circle|1) = \frac{1+1}{12+3} = \frac{2}{15} = \frac{2}{5}$$

$$P(circle|-1) = \frac{1+1}{10+3} = \frac{2}{13}$$



$$\boldsymbol{P(1|test\_obj) = (3)\left(\frac{3}{5}\right)\times(1)\left(\frac{1}{5}\right)\times(1)\left(\frac{2}{5}\right) = \frac{18}{125} = 14.4\%}$$ ***MORE LIKELY***

$$\boldsymbol{P(-1|test\_obj) = (3)\left(\frac{3}{13}\right)\times(1)\left(\frac{7}{13}\right)\times(1)\left(\frac{2}{13}\right) = \frac{126}{2197} = 5.7\%}$$

[Problem 3] – Similar to the problem of Decision tree, except develop the Naïve Bayes Solution

The problem is to predict whether it is a good day to play tennis given various factors and some initial data that provides information about whether previous days were good or bad days for tennis. The factors include (in the format "Attribute: List, of, Possible, Values"):

Outlook: Sunny, Rain, Overcast

Temperature: Hot, Mild, Cool

Humidity: High, Normal

Wind: String, Weak

Outlook=Sunny,Temperature=Hot,Humidity=High,Wind=Weak,No

Outlook=Sunny,Temperature=Hot,Humidity=High,Wind=Strong,No

Outlook=Overcast,Temperature=Hot,Humidity=High,Wind=Weak,Yes

Outlook=Rain,Temperature=Mild,Humidity=High,Wind=Weak,Yes

Outlook=Rain,Temperature=Cool,Humidity=Normal,Wind=Weak,Yes

Outlook=Rain,Temperature=Cool,Humidity=Normal,Wind=Strong,No

Outlook=Overcast,Temperature=Cool,Humidity=Normal,Wind=Strong,Yes

Outlook=Sunny,Temperature=Mild,Humidity=High,Wind=Weak,No

Outlook=Sunny,Temperature=Cool,Humidity=Normal,Wind=Weak,Yes

Outlook=Rain,Temperature=Mild,Humidity=Normal,Wind=Weak,Yes

Outlook=Sunny,Temperature=Mild,Humidity=Normal,Wind=Strong,Yes

Outlook=Overcast,Temperature=Mild,Humidity=High,Wind=Strong,Yes

Outlook=Overcast,Temperature=Hot,Humidity=Normal,Wind=Weak,Yes

Outlook=Rain,Temperature=Mild,Humidity=High,Wind=Strong,No

i. *Written answer. Show explicitly how the last line below is derived from the first line using Bayes rule, the chain rule, independence assumptions, etc.*

$$p(\textbf{yes} \mid o,t,h,w) > p(\textbf{no} \mid o,t,h,w)$$

$$i\ goes\ from\ 0 \rightarrow n$$

$$P(yes|x_i) = \frac{(P(x_i\ |no) \times P(no))}{P(x_i)}$$

$$P(no|x_i) = \frac{(P(x_i\ |no) \times P(no))}{P(x_i)}$$

$$\boldsymbol{X = o, t, h, w}$$

$$P(yes|X) = \frac{(P(x_1\ |yes) \times P(yes))}{P(x_1)} \times \ldots\ldots \times \frac{(P(x_n\ |yes) \times P(yes))}{P(x_n)} \times P(yes)$$

and

$$P(no|X) = \frac{(P(x_1\ |no) \times P(no))}{P(x_1)} \times \ldots\ldots \times \frac{(P(x_n\ |no) \times P(no))}{P(x_n)} \times P(no)$$

compare $\boldsymbol{P(yes|X)}$ and $\boldsymbol{P(no|X)}$

ii. Written answer. Using the training set to determine the relevant parameters, what is the most probable label for:

```
Outlook=Sunny,Temperature=Hot,Humidity=Normal,Wind=Weak
```

Make sure to show your work, including the values you obtained for each label.

$$\boldsymbol{P(yes|X) = (0.4) \times (0.5) \times \left(\frac{6}{7}\right) \times (0.75) = \frac{9}{70} \times \left(\frac{9}{14}\right) = 8.3\%}\ \textbf{\textit{MOST PROBABLE}}$$

$$\boldsymbol{P(yes|X) = (0.6) \times (0.5) \times \left(\frac{1}{7}\right) \times (0.25) \times \left(\frac{5}{14}\right) = 0.04\%}$$

Outlook

$$P(yes|Sunny) = \frac{(P(Sunny\ |yes) \times P(yes))}{P(sunny)} = \frac{\frac{2}{9} \times \frac{9}{14}}{\frac{5}{14}} = 0.4$$

$$P(no|Sunny) = \frac{(P(Sunny\ |no) \times P(no))}{P(sunny)} = \frac{\frac{3}{5} \times \frac{5}{14}}{\frac{5}{14}} = 0.6$$

$$P(yes|Overcast) = \frac{(P(Overcast\ |yes) \times P(yes))}{P(Overcast)} = \frac{\frac{4}{9} \times \frac{9}{14}}{\frac{4}{14}} = 1$$

$$P(no|Overcast) = \frac{(P(Overcast\ |no) \times P(no))}{P(Overcast)} = \frac{\frac{0}{5} \times \frac{5}{14}}{\frac{4}{14}} = 0$$

$$P(yes|Rainy) = \frac{(P(Rainy\ |yes) \times P(yes))}{P(Rainy)} = \frac{\frac{3}{9} \times \frac{9}{14}}{\frac{5}{14}} = 0.6$$

$$P(no|Rainy) = \frac{(P(Rainy\ |no) \times P(no))}{P(Rainy)} = \frac{\frac{2}{5} \times \frac{5}{14}}{\frac{5}{14}} = 0.4$$

Temperature

$$P(yes|Hot) = \frac{(P(Hot\ |yes) \times P(yes))}{P(Hot)} = \frac{\frac{2}{9} \times \frac{9}{14}}{\frac{4}{14}} = 0.5$$

$$P(no|Hot) = \frac{(P(Hot\ |no) \times P(no))}{P(Hot)} = \frac{\frac{2}{5} \times \frac{5}{14}}{\frac{4}{14}} = 0.5$$

$$P(yes|Mild) = \frac{(P(Mild\ |yes) \times P(yes))}{P(Mild)} = \frac{\frac{4}{9} \times \frac{9}{14}}{\frac{6}{14}} = \frac{2}{3}$$

$$P(no|Mild) = \frac{(P(Mild\ |no) \times P(no))}{P(Mild)} = \frac{\frac{2}{5} \times \frac{5}{14}}{\frac{6}{14}} = \frac{1}{3}$$

$$P(yes|Cool) = \frac{(P(Cool\ |yes) \times P(yes))}{P(Cool)} = \frac{\frac{3}{9} \times \frac{9}{14}}{\frac{4}{14}} = 0.75$$

$$P(no|Cool) = \frac{(P(Cool\ |no) \times P(no))}{P(Cool)} = \frac{\frac{1}{5} \times \frac{5}{14}}{\frac{4}{14}} = 0.25$$

Humidity

$$P(yes|High) = \frac{(P(High\ |yes) \times P(yes))}{P(High)} = \frac{\frac{3}{9} \times \frac{9}{14}}{\frac{7}{14}} = \frac{3}{7}$$

$$P(no|High) = \frac{(P(High\ |no) \times P(no))}{P(High)} = \frac{\frac{4}{5} \times \frac{5}{14}}{\frac{7}{14}} = \frac{4}{7}$$

$$P(yes|Normal) = \frac{(P(Normal\ |yes) \times P(yes))}{P(Normal)} = \frac{\frac{6}{9} \times \frac{9}{14}}{\frac{7}{14}} = \frac{6}{7}$$

$$P(no|Normal) = \frac{(P(Normal\ |no)\times P(no))}{P(Normal)} = \frac{\frac{1}{5}\times\frac{5}{14}}{\frac{7}{14}} = \frac{1}{7}$$

Wind

$$P(yes|Strong) = \frac{(P(Strong\ |yes)\times P(yes))}{P(Strong)} = \frac{\frac{3}{9}\times\frac{9}{14}}{\frac{6}{14}} = 0.5$$

$$P(no|High) = \frac{(P(Strong\ |no)\times P(no))}{P(Strong)} = \frac{\frac{3}{5}\times\frac{5}{14}}{\frac{6}{14}} = 0.5$$

$$P(yes|Normal) = \frac{(P(Weak\ |yes)\times P(yes))}{P(Weak)} = \frac{\frac{6}{9}\times\frac{9}{14}}{\frac{8}{14}} = 0.75$$

$$P(no|Normal) = \frac{(P(Weak\ |no)\times P(no))}{P(Weak)} = \frac{\frac{2}{5}\times\frac{5}{14}}{\frac{8}{14}} = 0.25$$

i. Graduate Student. Written answer. Derive the general formula for calculating p(x|o,t,h,w) and calculate p(yes|overcast,cool,normal,weak) based on parameters estimated from the training set.

(repeated from above)

$$goes\ from\ 0 \rightarrow n$$

$$P(c|x_i) = \frac{(P(x_i\ |c)\times P(c))}{P(x_i)}$$

$$\boldsymbol{X = o, t, h, w}$$

$$P(yes|X) = \frac{(P(x_1\ |c)\times P(c))}{P(x_1)} \times \ldots\ldots \times \frac{(P(x_n\ |c)\times P(c))}{P(x_n)} \times P(c)$$

$$\boldsymbol{P(yes|overcast, cool, normal, weak) = (1)\times(0.75)\times\left(\frac{6}{7}\right)\times(0.75)\times\left(\frac{9}{14}\right) = 30.1\%}$$

ii. Graduate Student. Written answer. Provide a set of attribute values o, t, h, and w for which the probability of either yes or no is zero.

X =(o = Overcast, t = Cool, h = High, w = Strong) will guarantee that the probability of no is zero because:

$$P(no|Overcast) = \frac{(P(Overcast\ |no)\times P(no))}{P(Overcast)} = \frac{\frac{0}{5}\times\frac{5}{14}}{\frac{4}{14}} = 0$$

$$\therefore P(no\ |X) = (0)\times(0.25)\times\left(\frac{4}{7}\right)\times(0.5) = 0\%$$