COMS 4030A

EXERCISES

(1) Consider the following dataset from Lecture 5:

$$S = \left[\begin{array}{c|c|c|c} F_1 & F_2 & F_3 & \text{target} \\ \hline T & 0 & B & Yes \\ F & 2 & A & No \\ F & 1 & C & Yes \\ T & 2 & B & Yes \\ F & 0 & A & No \\ F & 2 & C & No \\ F & 0 & A & Yes \\ T & 1 & B & No \\ T & 0 & B & Yes \\ F & 1 & C & Yes \\ T & 2 & A & No \\ T & 2 & C & No \end{array}\right]$$

Use the Naïve Bayes Classifier to classify input $(F, 1, A)$ as either Yes or No.

Do the same for inputs $(F, 0, A)$ and for $(T, 2, C)$.

Solutions:

Input $(F, 1, A)$:

$P(Yes \mid F_1 = F)\, P(Yes \mid F_2 = 1)\, P(Yes \mid F_3 = A)\, P(Yes) = \frac{1}{72}$

$P(No \mid F_1 = F)\, P(No \mid F_2 = 1)\, P(No \mid F_3 = A)\, P(No) = \frac{1}{48}$

Thus, for input $(F, 1, A)$, the Naïve Bayes Classifier predicts No.

Input $(F, 0, A)$:

$P(Yes \mid F_1 = F)\, P(Yes \mid F_2 = 0)\, P(Yes \mid F_3 = A)\, P(Yes) = \frac{1}{48}$

$P(No \mid F_1 = F)\, P(No \mid F_2 = 0)\, P(No \mid F_3 = A)\, P(No) = \frac{1}{48}$

Thus, for input $(F, 0, A)$, the Naïve Bayes Classifier has no preference between Yes and No.

Input $(T, 2, C)$:

$P(Yes \mid F_1 = T)\, P(Yes \mid F_2 = 2)\, P(Yes \mid F_3 = C)\, P(Yes) = \frac{1}{72}$

$P(No \mid F_1 = T)\, P(No \mid F_2 = 2)\, P(No \mid F_3 = C)\, P(No) = \frac{1}{18}$

Thus, for input $(T, 2, C)$, the Naïve Bayes Classifier predicts No.

(2) Consider the following dataset from Lecture 5:

$$S = \left[\begin{array}{c|c|c|c} F_1 & F_2 & F_3 & \text{target} \\ \hline a & 0 & Y & A \\ b & 0 & N & C \\ c & 1 & Y & B \\ b & 1 & Y & B \\ c & 0 & N & A \\ a & 0 & N & C \\ c & 1 & N & B \\ a & 1 & Y & A \\ b & 0 & Y & B \\ a & 1 & N & C \end{array}\right]$$

Use the Naïve Bayes Classifier to classify input $(b, 1, N)$ as either A, B or C.

Do the same with input $(c, 1, Y)$.

Solutions:

Input $(b, 1, N)$:

$P(F_1 = b \mid A)\, P(F_2 = 1 \mid A)\, P(F_3 = N \mid A)\, P(A) = 0$

$P(F_1 = b \mid B)\, P(F_2 = 1 \mid B)\, P(F_3 = N \mid B)\, P(B) = \frac{3}{80}$

$P(F_1 = b \mid C)\, P(F_2 = 1 \mid C)\, P(F_3 = N \mid C)\, P(C) = \frac{1}{30}$

Thus, for input $(b, 1, N)$, the Naïve Bayes Classifier predicts B.

Input $(c, 1, Y)$:

$P(F_1 = c \mid A)\, P(F_2 = 1 \mid A)\, P(F_3 = Y \mid A)\, P(A) = \frac{1}{45}$

$P(F_1 = c \mid B)\, P(F_2 = 1 \mid B)\, P(F_3 = Y \mid B)\, P(B) = \frac{9}{80}$

$P(F_1 = c \mid C)\, P(F_2 = 1 \mid C)\, P(F_3 = Y \mid C)\, P(C) = 0$

Thus, for input $(c, 1, Y)$, the Naïve Bayes Classifier predicts B.

(3) Consider an urn problem similar to the example in the notes, but where Urn 1 has 1 *Red* ball and 3 *Blue* balls, and Urn 2 has 7 *Red* balls and 3 *Blue* balls.

(a) A friend uses a fair coin to choose an urn. Give the prior probability distribution on the urns.

(b) Your friend secretly takes a ball out of the chosen urn (so you can't see which urn) and tells you the ball is *Blue* and then returns it to the urn. Use Bayes' rule to work out $P(\text{Urn } 1 | \mathit{Blue})$ and $P(\text{Urn } 2 | \mathit{Blue})$.

(c) Your friend then draws another ball out of the chosen urn, and tells you it is *Red*. Now work out $P(\text{Urn 1}|Red)$ and $P(\text{Urn 2}|Red)$.

Solutions:

(a) $P(\text{Urn 1}) = \frac{1}{2}$, $P(\text{Urn 2}) = \frac{1}{2}$

(b) $P(\text{Urn 1}|Blue) = \frac{(\frac{3}{4})(\frac{1}{2})}{\frac{21}{40}} = \frac{5}{7}$ $\quad P(\text{Urn 2}|Blue) = \frac{(\frac{3}{10})(\frac{1}{2})}{\frac{21}{40}} = \frac{2}{7}$

(c) $P(\text{Urn 1}|Blue) = \frac{(\frac{1}{4})(\frac{5}{7})}{\frac{53}{140}} = \frac{25}{53}$ $\quad P(\text{Urn 2}|Blue) = \frac{(\frac{7}{10})(\frac{2}{7})}{\frac{53}{140}} = \frac{28}{53}$

(4) Suppose you have a dataset compiled from 200 randomly chosen people. For each person in the dataset you have some personal information and you also know if the person prefers cricket or soccer. There are 80 people in the dataset that prefer cricket and 120 people that prefer soccer.

(a) Suppose you want to know if a new random person prefers cricket or soccer. Based on the sample dataset, what is your prior probability that the person prefers cricket and the prior probability that the person prefers soccer?

(b) You are now informed that the new person is female. In your dataset, you know that 45 of the 80 people that prefer cricket are female, and 50 of the 120 people that prefer soccer are female. Use Bayes' rule to update your probability that the person prefers cricket and the probability that the person prefers soccer.

(c) You now receive some more information - the person is under 25 years old. In your dataset, of the 80 that prefer cricket, 30 are under 25 and 15 of them are female, and of the 120 that prefer soccer, 50 are under 25 and 30 of them are female. Use Bayes' rule to determine the probability that the person prefers cricket and the probability that the person prefers soccer.

Solutions:

(a) $P(\text{cricket}) = \frac{80}{200}$, $\quad P(\text{soccer}) = \frac{120}{200}$

(b) $P(\text{cricket}|\text{female}) = \frac{(\frac{45}{80})(\frac{80}{200})}{\frac{95}{200}} = \frac{9}{19}$, $\quad P(\text{soccer}|\text{female}) = \frac{(\frac{50}{120})(\frac{120}{200})}{\frac{95}{200}} = \frac{10}{19}$

(c) $P(\text{cricket}|\text{female and under 25}) = 0.403$, $\quad P(\text{soccer}|\text{female and under 25}) = 0.597$

(5) Suppose you are given a dataset S with N datapoints. Suppose S has attributes x_1, x_2, x_3 and target t, where attribute x_1 can take values in $\{T, F\}$, attribute x_2 can take values in $\{a, b, c\}$ and attribute x_3 can take values in $\{0, 1, 2, 3\}$. The target t can take values in $\{C_1, C_2, C_3\}$. Describe how to implement the Naïve Bayes Classifier for any input $\boldsymbol{x} = (x_1, x_2, x_3)$, where $x_1 \in \{T, F\}$, $x_2 \in \{a, b, c\}$ and $x_3 \in \{0, 1, 2, 3\}$.

Give pseudocode for an algorithm that would efficiently implement the Naïve Bayes Classifier for a given input on this dataset ....