CS57800 Statistical Machine Learning
Homework 4

Ting Zhang
School of Industiral Engineering
zhan1013@purdue.edu

December 10, 2015

1 Probability

1. From the probability table, we can get

$$P(X=1) = \frac{1}{15} + \frac{1}{10} + \frac{2}{15} + \frac{4}{45} = \frac{7}{18}$$
$$P(Y=1) = \frac{1}{10} + \frac{1}{10} + \frac{8}{45} + \frac{4}{45} = \frac{7}{15}$$

and

$$P(X=1, Y=1) = \frac{1}{10} + \frac{4}{45} = \frac{17}{90}$$

Then, we can observe,

$$P(X=1) \times P(Y=1) = \frac{7}{18} \times \frac{7}{15} = \frac{49}{270} \neq P(X=1, Y=1)$$

Therefore, X is not independent on Y.

Another way to prove X is not independt on Y is to prove $P(X|Y) \neq P(X)$.

$$P(X=1|Y=1) = \frac{P(X=1, Y=1)}{P(Y=1)} = \frac{\frac{1}{10} + \frac{4}{45}}{\frac{7}{15}} = \frac{17}{42}$$

which is not equal to $P(X=1) = \frac{7}{18}$.

2. To show if X is conditionally independnt of Y given Z, we need to do the following calculation:

Given $Z = 1$,

$$P(X=1|Z=1) = \frac{P(X=1,Z=1)}{P(Z=1)} = \frac{\frac{2}{15}+\frac{4}{35}}{\frac{2}{3}} = \frac{1}{3}$$
$$P(X=0|Z=1) = 1 - P(X=1|Z=1) = \frac{2}{3}$$
$$P(Y=1|Z=1) = \frac{P(Y=1,Z=1)}{P(Z=1)} = \frac{\frac{8}{45}+\frac{4}{35}}{\frac{2}{3}} = \frac{2}{5}$$
$$P(Y=0|Z=1) = 1 - P(Y=1|Z=1) = \frac{3}{5}$$
$$P(X=1,Y=1|Z=1) = \frac{P(X=1,Y=1,Z=1)}{P(Z=1)} = \frac{\frac{4}{45}}{\frac{30}{45}} = \frac{2}{15}$$
$$P(X=1,Y=0|Z=1) = \frac{P(X=1,Y=0,Z=1)}{P(Z=1)} = \frac{\frac{2}{15}}{\frac{30}{45}} = \frac{1}{5}$$
$$P(X=0,Y=1|Z=1) = \frac{P(X=0,Y=1,Z=1)}{P(Z=1)} = \frac{\frac{8}{45}}{\frac{30}{45}} = \frac{4}{15}$$
$$P(X=0,Y=0|Z=1) = \frac{P(X=0,Y=0,Z=1)}{P(Z=1)} = \frac{\frac{4}{15}}{\frac{30}{45}} = \frac{2}{5}$$

Here,

$$P(X=1,Y=1|Z=1) = P(X=1|Z=1) \times P(Y=1|Z=1),$$
$$P(X=1,Y=0|Z=1) = P(X=1|Z=1) \times P(Y=0|Z=1),$$
$$P(X=0,Y=1|Z=1) = P(X=0|Z=1) \times P(Y=1|Z=1),$$
$$P(X=0,Y=0|Z=1) = P(X=0|Z=1) \times P(Y=0|Z=1).$$

Given $Z = 0$,

$$P(X=1|Z=0) = \frac{P(X=1,Z=0)}{P(Z=0)} = \frac{\frac{1}{15}+\frac{1}{10}}{\frac{1}{3}} = \frac{1}{2}$$
$$P(X=0|Z=0) = 1 - P(X=1|Z=0) = \frac{1}{2}$$
$$P(Y=1|Z=0) = \frac{P(Y=1,Z=0)}{P(Z=0)} = \frac{\frac{1}{10}+\frac{1}{10}}{\frac{1}{3}} = \frac{3}{5}$$
$$P(Y=0|Z=0) = 1 - P(Y=1|Z=0) = \frac{2}{5}$$
$$P(X=1,Y=1|Z=0) = \frac{P(X=1,Y=1,Z=0)}{P(Z=0)} = \frac{\frac{1}{10}}{\frac{1}{3}} = \frac{3}{10}$$
$$P(X=1,Y=0|Z=0) = \frac{P(X=1,Y=0,Z=0)}{P(Z=0)} = \frac{\frac{1}{15}}{\frac{1}{3}} = \frac{1}{5}$$
$$P(X=0,Y=1|Z=0) = \frac{P(X=0,Y=1,Z=0)}{P(Z=0)} = \frac{\frac{1}{10}}{\frac{1}{3}} = \frac{3}{10}$$
$$P(X=0,Y=0|Z=0) = \frac{P(X=0,Y=0,Z=0)}{P(Z=0)} = \frac{\frac{1}{15}}{\frac{1}{3}} = \frac{1}{5}$$

Here,

$$P(X=1,Y=1|Z=0)=P(X=1|Z=0)\times P(Y=1|Z=0),$$
$$P(X=1,Y=0|Z=0)=P(X=1|Z=0)\times P(Y=0|Z=0),$$
$$P(X=0,Y=1|Z=0)=P(X=0|Z=0)\times P(Y=1|Z=0),$$
$$P(X=0,Y=0|Z=0)=P(X=0|Z=0)\times P(Y=0|Z=0).$$

Therefore, given Z, X and Y is independent.

3.

$$P(X=0|X+Y>0)=\frac{P(X=0,(X+Y)>0)}{P(X+Y>0)}=\frac{\frac{1}{10}+\frac{8}{45}}{\frac{1}{15}+\frac{2}{15}+\frac{1}{10}+\frac{1}{10}+\frac{8}{45}+\frac{4}{45}}=\frac{5}{12}$$

2 Hidden Markov Model

1. The HMM model can be described as:

   - States: $\{N,C\}$, where N means nice day, and C means cold day.
   - Observations: $\{O,I,H\}$, where O means groundhog outside his burrow, I means inside the burrow and H means only his head sticking out of the burrow.
   - Initial State Probabilities: $P(S_1=N)=P(S_1=C)=1/2$.
   - Transition Probabilities:

$$P(S_t=N|S_{t-1}=N)=\frac{2}{3}$$
$$P(S_t=N|S_{t-1}=C)=\frac{1}{2}$$
$$P(S_t=C|S_{t-1}=N)=\frac{1}{3}$$
$$P(S_t=C|S_{t-1}=C)=\frac{1}{2}$$

   - Observation Probabilities:

$$P(O_t=O|S_t=N)=\frac{2}{3}$$
$$P(O_t=I|S_t=N)=\frac{1}{6}$$
$$P(O_t=H|S_t=N)=\frac{1}{6}$$
$$P(O_t=O|S_t=C)=\frac{1}{4}$$
$$P(O_t=I|S_t=C)=\frac{1}{2}$$
$$P(O_t=H|S_t=C)=\frac{1}{4}$$

	t=1	t=2	t=3	t=4
Nice	V_{11}	V_{21}	V_{31}	V_{41}
Cold	V_{12}	V_{22}	V_{32}	V_{42}

2. To find out the most likely sequence of states, we can apply the viterbi algorithm using the initial state probabilities, transition and observation probabilities. We can build the table like this:

   The probabilities of the first column of the table as the initial probabilities, can be computed as:

$$V_{11} = P(O_1 = I|S_1 = N)P(S_1 = N) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$$
$$V_{12} = P(O_1 = I|S_1 = C)P(S_1 = C) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

   Then, the following probabilities can be calcualted following this equation:

$$V_{tk} = \max_{x \in S}(P(O_t|S_t = k) \times P(k|S_{t-1} = x) \times V_{t-1,x}),$$

   where k is the current state. Therefore, we can compute:

$$\begin{aligned}
V_{21} &= \max(P(O_2 = H|S_2 = N) \times P(S_2 = N|S_1 = N) \times V_{11}, \\
&\qquad P(O_2 = H|S_2 = N) \times P(S_2 = N|S_1 = C) \times V_{12}) \\
&= \max(\frac{1}{6} \times \frac{2}{3} \times \frac{1}{12}, \frac{1}{6} \times \frac{1}{2} \times \frac{1}{4}) \\
&= \max(\frac{1}{108}, \frac{1}{48}) = \frac{1}{48} \\
V_{22} &= \max(P(O_2 = H|S_2 = C) \times P(S_2 = C|S_1 = N) \times V_{11}, \\
&\qquad P(O_2 = H|S_2 = C) \times P(S_2 = C|S_1 = C) \times V_{12}) \\
&= \max(\frac{1}{4} \times \frac{1}{3} \times \frac{1}{12}, \frac{1}{4} \times \frac{1}{2} \times \frac{1}{4}) \\
&= \max(\frac{1}{144}, \frac{1}{32}) = \frac{1}{32} \\
V_{31} &= \max(P(O_3 = I|S_3 = N) \times P(S_3 = N|S_2 = N) \times V_{21}, \\
&\qquad P(O_3 = I|S_3 = N) \times P(S_3 = N|S_2 = C) \times V_{22}) \\
&= \max(\frac{1}{6} \times \frac{2}{3} \times \frac{1}{48}, \frac{1}{6} \times \frac{1}{2} \times \frac{1}{32}) \\
&= \max(\frac{1}{432}, \frac{1}{384}) = \frac{1}{384} \\
V_{32} &= \max(P(O_3 = I|S_3 = C) \times P(S_3 = C|S_2 = N) \times V_{21}, \\
&\qquad P(O_3 = I|S_3 = C) \times P(S_3 = C|S_2 = C) \times V_{22}) \\
&= \max(\frac{1}{2} \times \frac{1}{3} \times \frac{1}{48}, \frac{1}{2} \times \frac{1}{2} \times \frac{1}{32}) \\
&= \max(\frac{1}{288}, \frac{1}{128}) = \frac{1}{128}
\end{aligned}$$

$$
\begin{aligned}
V_{41} &= \max(P(O_4 = O|S_4 = N) \times P(S_4 = N|S_3 = N) \times V_{31}, \\
&\qquad P(O_4 = O|S_4 = N) \times P(S_4 = N|S_3 = C) \times V_{32}) \\
&= \max(\frac{2}{3} \times \frac{2}{3} \times \frac{1}{384}, \frac{2}{3} \times \frac{1}{2} \times \frac{1}{128}) \\
&= \max(\frac{1}{864}, \frac{1}{384}) = \frac{1}{384} \\
V_{42} &= \max(P(O_4 = O|S_4 = C) \times P(S_4 = C|S_3 = N) \times V_{31}, \\
&\qquad P(O_4 = O|S_4 = C) \times P(S_4 = C|S_3 = C) \times V_{32}) \\
&= \max(\frac{1}{4} \times \frac{1}{3} \times \frac{1}{384}, \frac{1}{4} \times \frac{1}{2} \times \frac{1}{128}) \\
&= \max(\frac{1}{4608}, \frac{1}{1024}) = \frac{1}{1024}
\end{aligned}
$$

Therefore, the table now looks like:

	t=1	**t=2**	**t=3**	**t=4**
Nice	$\frac{1}{12}$	$\frac{1}{48}$	$\frac{1}{384}$	$\frac{1}{384}$
Cold	$\frac{1}{4}$	$\frac{1}{32}$	$\frac{1}{128}$	$\frac{1}{1024}$

From this table, we will start from the last column, $t = 4$. We pick the largest value and check its condition in the equations above. First, we pick $P = 1/384$, and found it is the case of $S_3 = C$ and $S_4 = N$. Then, we go back to column of $t = 3$ in the table. Since, $S_3 = C$, $P = 1/128$ is selected and the case of $S_3 = C$ and $S_2 = C$ is found. Then, since $S_2 = C$, in column of $t = 2$, $P = 1/32$ is selected and found $S_2 = C$ and $S_1 = C$. Therefore, the most likely sequence of states is $\{C, C, C, N\}$.

3 Naive Bayes

1. The threshold function can be formulated as:

$$f_{TH(3,7)} = 1 \text{ if } \sum_{i=1}^{7} x_i \geq 3.$$

Therefore, it is a linear combination of all components with coefficient of 1. It is a linear decision surface over the 7 dimensionl Boolean cube.

2. Since the data sampled are uniformly distributed, we can calcualte the probabilities as following:
To compute $P(f_{TH(3,7)} = 0)$, we need to count the number of combinations that has 0, 1, or 2 components are 0, which is ${}^7C_0 = 1, {}^7C_1 = 7, {}^7C_2 = 21$, repectively. Therefore,

$$
\begin{aligned}
P(f_{TH(3,7)} = 0) &= \frac{1}{128} + \frac{7}{128} + \frac{21}{128} = \frac{29}{128} \\
P(f_{TH(3,7)} = 1) &= 1 - P(f_{TH(3,7)} = 0) = 1 - \frac{29}{128} = \frac{99}{128}
\end{aligned}
$$

Here, we denote the result of $f_{TH(3,7)}$ as label. Then, similar method is applied to compute the conditional probabilites $p_i = P(x_i = 1|label = 1)$ and $q_i = P(x_i = 1|label = 0)$. Therefore, to compute q_i, given label equals 0, we need to count the number of combinations that has x_i equals 1. When the label is 0, and x_i is 1, it means there can be 0 or 1more compents are 1, which is ${}^6C_0 = 1$ and ${}^6C_1 = 6$, repectively. Therefore,

$$q_i = P(x_i = 1|label = 0) = \frac{1}{1+7+21} + \frac{6}{1+7+21} = \frac{7}{29}.$$

Then, to compute p_i, given label equals 1, we need to count the number of combinations that has x_i equals 1. When the label is 1 and x_i is 1, it means there must be 2, 3, 4, 5 or 6 more components are 1, which is ${}^6C_2 = 15$, ${}^6C_3 = 20$, ${}^6C_4 = 15$, ${}^6C_5 = 6$ and ${}^6C_6 = 1$, respectively. Therefore,

$$p_i = P(x_i = 1|label = 1) = \frac{15}{99} + \frac{20}{99} + \frac{15}{99} + \frac{6}{99} + \frac{1}{99} = \frac{57}{99}.$$

Then, the hypothesis can be formulated as, we predict $label = 1$ iff:

$$\begin{aligned}
\log\frac{P(label=1)}{P(label=0)} + \sum_i \log\frac{1-p_i}{1-q_i} + \sum_i(\log\frac{p_i}{1-p_i} - \log\frac{q_i}{1-q_i})x_i \geq 0 \\
\log\frac{99}{29} + \sum_i \log(\frac{42}{99}/\frac{22}{29}) + \sum_i(\log\frac{57}{42} - \log\frac{7}{22})x_i \geq 0 \\
0.533 + 7\times(-0.252) + \sum_i(0.133 - (-0.497))x_i \geq 0 \\
0.63\sum_i x_i \geq 1.231 \\
\sum_i x_i \geq 1.954
\end{aligned}$$

3. The hypothesis generated by Naive Bayes can't represent this function, $f_{TH(3,7)}$. If there are only two components of the Cube point is 1, the Naive Bayes hypothesis would classify the label as 1; however, the funtion, $f_{TH(3,7)}$, will generate a label of 0, instead. The Naive Bayes hypothesis can correctly classify the points who has 3, or more than 3 components of 1s, or has 0 or 1 component of 1s.

4. The assumption of Naive Bayes is that feature values are independent given the target value, which means

$$P(x_1 = b_1, x_2 = b_2, ..., x_n = b_n|v = v_j) = \prod_i P(x_i = b_i|v = v_j).$$

The function, $f_{TH(3,7)}$, does not satisfy the naive Bayes assumption. First, we can compute,

$$P(x_1 = 1, x_2 = 1, ..., x_7 = 1|label = 1) = \frac{P(x_1 = 1, x_2 = 1, ..., x_7 = 1, label = 1)}{P(label = 1)} = \frac{1}{99}$$

Since the data is uniformally sampled, as computed in last sub-question,

$$P(x_1 = 1|label = 1) = P(x_2 = 1|label = 1) = ... = P(x_7 = 1|label = 1) = \frac{57}{99}.$$

Then, we can observe

$$P(x_1 = 1, x_2 = 1, ..., x_7 = 1|label = 1) = \frac{1}{99} \neq \prod_{i=1}^{7} P(x_i = 1|label = 1) = (\frac{57}{99})^7,$$

which means the feature values are not independent given the target value.