Time: 120 minutes Max. Marks: 19

1. (4 points) A non-expert fishing enthusiast, curious to know what counts as a good day of fishing (F) at the lake and puzzled by the phenomenon that it can sometimes be a good day even when no fish are caught, decides to create a naive Bayes model with ($F \in \{true, false\}$) as the Boolean class variable and three features: whether it rained ($R \in \{true, false\}$), how many fish were caught ($C \in \{none, some, lots\}$), and whether it was windy ($W \in \{true, false\}$).
   (a) (1 point) Draw the Bayesian Network for the Naive Bayes model with F as the Boolean class (target) variable and W, C, R as the features.
   (b) (3 points) Following are the (plausible) CPTs given to you for this fictitious setting:

$$P(F = false) = 0.9$$
$$P(W = false|F = false) = 0.5, \quad P(W = false|F = true) = 0.8$$
$$P(C = none|F = false) = 0.7, \quad P(C = none|F = true) = 0.3$$
$$P(C = some|F = false) = 0.2, \quad P(C = some|F = true) = 0.4$$
$$P(R = false|F = fasle) = 0.6, \quad P(R = false|F = true) = 0.9$$

   Given the above model, calculate the following probabilities:
   i. (1 point) $P(F = true|R = true, C = none, W = true)$.
   ii. (1 point) $P(F = true|R = false, C = lots, W = false)$
   iii. (1 point) $P(F = true|R = true, C = some, W = false)$

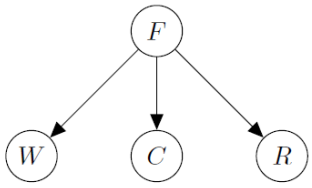

Figure 1: Bayesian Network for the Naive Bayes Classifier for predicting a *good* day for fishing

Solution:

(a) The Bayesian network for the Naive Bayes Classifier is given in Fig. 1

1 mark if the figure is correct

(b) We know that for a Naive Bayes classifier, the features are conditionally independent of each other, given the class. Therefore we have the following relationship:

$$P(R, C, W|F) = P(R|F)P(C|F)P(W|F) \tag{1}$$

From the product rule, we have

$$P(R, C, W, F) = P(R, C, W|F)P(F)$$

Then the joint probability $P(R, C, W)$ can be computed as:

$$\begin{aligned} P(R, C, W) &= P(R, C, W|F = t)P(F = t) + P(R, C, W|F = f)P(F = f) \\ &= P(R|F = t)P(C|F = t)P(W|F = t)P(F = t) \\ &\quad + P(R|F = f)P(C|F = f)P(W|F = f)P(F = f) \end{aligned} \tag{2}$$

where t denotes *true* and f denotes $false$. We also denote *none*, *some* and *lots* using n, s and l respectively.

This step is used for all the three parts. Therefore, the mark distribution of this step is done accordingly. If the equation to compute P(R, C, W) is correct, give 0.3 marks for each part. If the numerical values are correctly used, give 0.2 marks for each part. Therefore, if this step is done once for all the parts (like in the solution), give $0.3 \times 3 = 0.9$ marks for the correct equation. The numerical values are different for each part and therefore should be evaluated separately.

i. Therefore, we have the following relationships that can be simplified using eqn. (2)

$$
\begin{aligned}
P(F=t|R=t,C=n,W=t) &= \frac{P(R=t,C=n,W=t|F=t)P(F=t)}{P(R=t,C=n,W=t)} \\
&= \frac{P(R=t|F=t)P(C=n|F=t)P(W=t|F=t)P(F=t)}{P(R=t,C=n,W=t)} \\
&= \frac{P(R=t|F=t)P(C=n|F=t)P(W=t|F=t)P(F=t)}{P(R=t,C=n,W=t)} \\
&= \frac{0.1\times 0.3\times 0.2\times 0.1}{(0.1\times 0.3\times 0.2\times 0.1)+(0.4\times 0.7\times 0.5\times 0.9)} \\
&= \frac{6\times 10^{-4}}{(6\times 10^{-4})+(0.2\times 0.63)} = \frac{6\times 10^{-4}}{(6\times 10^{-4})+(0.126)} \\
&= \frac{6\times 10^{-4}}{0.1266} = \frac{6}{12.66}\times 10^{-2} \approx 5\times 10^{-3}
\end{aligned}
$$

where the denominator is computed using eqn. (2).

ii.

$$
\begin{aligned}
P(F=t|R=f,C=l,W=f) &= \frac{P(R=f,C=l,W=f|F=t)P(F=t)}{P(R=f,C=l,W=f)} \\
&= \frac{P(R=f|F=t)P(C=l|F=t)P(W=f|F=t)P(F=t)}{P(R=f,C=l,W=f)} \\
&= \frac{0.9\times 0.3\times 0.8\times 0.1}{(0.9\times 0.3\times 0.8\times 0.1)+(0.6\times 0.1\times 0.5\times 0.9)} \\
&= \frac{0.72\times 0.03}{(0.0216)+(0.03\times 0.9)} = \frac{0.0216}{0.0216+0.027} \\
&= \frac{216}{486} = \frac{72}{162} = \frac{8}{18} = \frac{4}{9} \approx 0.44
\end{aligned}
$$

iii.

$$
\begin{aligned}
P(F=t|R=t,C=s,W=f) &= \frac{P(R=t,C=s,W=f|F=t)P(F=t)}{P(R=t,C=s,W=f)} \\
&= \frac{P(R=t|F=t)P(C=s|F=t)P(W=f|F=t)P(F=t)}{P(R=t,C=s,W=f)} \\
&= \frac{0.1\times 0.4\times 0.8\times 0.1}{(0.1\times 0.4\times 0.8\times 0.1)+(0.4\times 0.2\times 0.5\times 0.9)} \\
&= \frac{0.01\times 0.32}{(0.0032)+(0.36\times 0.1)} = \frac{0.0032}{0.0032+0.036} \\
&= \frac{0.0032}{0.0392} = \frac{32}{392} = \frac{4}{49} \approx 0.08
\end{aligned}
$$

For each part, if the equations are correct, give 0.3 marks and if the numerical values are correctly used, give 0.2 marks. Please note that the students should not be penalized for calculation mistakes. As long as the equation and the numerical values are correctly used, the students should be given full marks.

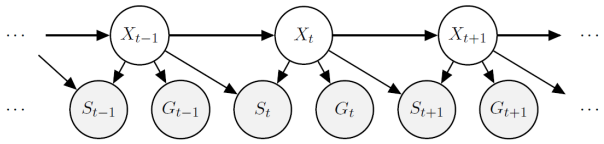

Figure 2: HMM for Q2

2. (5 points) Transportation researchers are trying to improve traffic in the city but, in order to do that, they first need to estimate the location of each of the cars in the city. They need our help to model this problem as an inference problem of an HMM. Assume that only one car is being modeled.

   We define the variables as follows: X, the location of the car; S, the noisy location of the car from the signal strength at a nearby cell phone tower; and G, the noisy location of the car from GPS. In our HMM model Fig. 2, the signal-dependent location S_t not only depends on the current state X_t but also depends on the previous state X_{t-1}. We want to compute the belief $P(x_t|s_{1:t}, g_{1:t})$. In this part we consider an update that combines the dynamics and observation update in a single update. Complete the forward update expression by filling in the expression of the form:

   $P(x_t|s_{1:t}, g_{1:t}) =$ _____ $P(x_{t-1}|s_{1:t-1}, g_{1:t-1})$.

   Over the steps of the derivation, identify where you apply the different assumptions about the transition and the sensor model.

Solution: For this modified HMM, we have the dynamics and observation update in a single update because one of the previous independence assumptions does not longer holds.

$$P(x_t|s_{1:t}, g_{1:t}) = \sum_{x_{t-1}} P(x_{t-1}, x_t|s_t, g_t s_{1:t-1}, g_{1:t-1}) \tag{3}$$

$$= \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} \sum_{x_{t-1}} P(x_{t-1}, x_t, s_t, g_t|s_{1:t-1}, g_{1:t-1}) \tag{4}$$

$$= \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} \sum_{x_{t-1}} P(s_t, g_t|x_{t-1}, x_t, s_{1:t-1}, g_{1:t-1}) P(x_{t-1}, x_t|s_{1:t-1}, g_{1:t-1}) \tag{5}$$

$$= \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} \sum_{x_{t-1}} P(s_t, g_t|x_{t-1}, x_t) P(x_t|x_{t-1}, s_{1:t-1}, g_{1:t-1}) P(x_{t-1}|s_{1:t-1}, g_{1:t-1}) \tag{6}$$

$$= \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} \sum_{x_{t-1}} P(s_t|x_{t-1}, x_t) P(g_t|x_{t-1}, x_t) P(x_t|x_{t-1}) P(x_{t-1}|s_{1:t-1}, g_{1:t-1}) \tag{7}$$

$$= \frac{1}{P(s_t, g_t|s_{1:t-1}, g_{1:t-1})} \sum_{x_{t-1}} P(s_t|x_{t-1}, x_t) P(g_t|x_t) P(x_t|x_{t-1}) P(x_{t-1}|s_{1:t-1}, g_{1:t-1}) \tag{8}$$

Eqn. (3) expresses the LHS in terms of the joint probability of x_t and x_{t-1} by summing over x_{t-1}.

In eqn. (4), the product rule is applied to express the joint probability of the variables of interest, i.e, (x_t, x_{t-1}, s_t, g_t) conditioned on the historical observations.

In eqn. (5) the chain rule is applied so the observation variables (s_t, g_t) are conditioned on the current and previous states (x_t, x_{t-1}).

In eqn. (6), the product rule is again applied, but this time to x_t and x_{t-1}. Here we also apply the sensor model assumption, which assumes conditional independence of (s_t, g_t) w.r.t. all previous observations $(s_{1:t-1}, g_{1:t-1})$, given the current and previous states (x_t, x_{t-1}).

In eqn. (7), we use the Markov model assumption, where the current state x_t is conditionally independent of all past states and observations, given x_{t-1}. We also apply the conditional independence of s_t and g_t, given (x_t, x_{t-1}) (which is the special condition in this problem; usually observations only depend on x_t).

Finally, in eqn. (8), we use the sensor model assumption for g_t, which is conditionally independent of x_{t-1}, given x_t (unlike s_t, which depends on both x_t and x_{t-1}.

If all the steps are correct, give 5 marks. If none of the steps are correct, give 0 marks. If the steps are partially correct, deduct 0.8 marks for each incorrect step.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Figure 3: Data Table for Q3

3. (4+3(Extra credit) points) (a) (2 points) What is the Shannon's entropy of a discrete random variable? Derive the Shannon's entropy for a discrete random variable that can take N values, all of which are equally likely. Show the steps and define all variables to get full credit.

   (b) (2 points) Define Information Gain and provide the expression for computing it. Introduce and properly define all the terms needed.

   (c) (3 (Extra Credit) points) Given the data for the Play Tennis example discussed in class (see Fig. 3), use the information gain criteria to identify the *best* variable at the first step of building a decision tree. Explain the steps and show the calculations of the steps leading to the computation

of the information gain. (**Note**: You may leave your calculations in fractions, after simplifying the expression as far as possible.)

Solution:

(a) The Shannon's entropy of a discrete random variable is given by:

$$H(X) = -\sum_{j=1}^{N} P(X = x_j) \log \left(P(X = x_j)\right) \tag{9}$$

where the random variable X can take N different values $x_1, x_2, \ldots, x_N$ and $P(X = x_j)$ denotes the probability with which X can take a value x_j. The log can be of any base which only changes the units of entropy.

1 mark for correct definition of Shannon's entropy

For equally likely values, i.e., a discrete random variable distributed uniformly, we have $P(X = x_j) = 1/N, \forall j = 1, 2, \ldots, N$

$$H(X) = -\frac{1}{N} \sum_{j=1}^{N} \log_2(1/N) \tag{10}$$

$$= -\frac{1}{N} \sum_{j=1}^{N} (-\log_2 \ N) \tag{11}$$

$$= \frac{1}{N} \sum_{j=1}^{N} \log_2 \ N \tag{12}$$

$$= \frac{1}{N} \cdot N \log_2 \ N = \log_2 N \tag{13}$$

1 mark for correct derivation

(b) Information Gain $I(X, Y)$ is given as:

$$I(X, Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) \tag{14}$$

where $I(X, Y)$ is the information gain (or mutual information) between random variables X and Y, and $H(Y)$ is the Shannon's entropy as defined in eqn. (9) above. The term $H(Y|X)$ is defined as below:

$$H(Y|X) = \sum_{i=1}^{N} P(X = x_i) \sum_{j=1}^{M} P(Y = y_j|X = x_i) \log P(Y = y_j|X = x_i) \tag{15}$$

where X can take values $x_1, x_2, \ldots, x_N$ and Y can take values $y_1, y_2, \ldots, y_M$. The $P(\cdot)$ are the (conditional) probabilities as per the usual notation and the term $H(X|Y)$ can be defined similarly.

1 mark for correct definition and 1 mark for defining all the variables properly

(c) **Applying to the Play Tennis Example:**

Given the dataset, we have the following counts for the target variable "PlayTennis":

$$\text{Yes} = 9, \quad \text{No} = 5, \quad |S| = 14.$$

Step 1: Compute Entropy of S

$$p(\text{Yes}) = \frac{9}{14}, \quad p(\text{No}) = \frac{5}{14}$$

$$H(S) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) \approx 0.940.$$

Step 2: Compute Information Gain for Each Attribute

Outlook (Sunny, Overcast, Rain)

- S_{Sunny}: 5 instances (2 Yes, 3 No)

$$H(S_{\text{Sunny}}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) \approx 0.971.$$

- S_{Overcast}: 4 instances (4 Yes, 0 No)

$$H(S_{\text{Overcast}}) = 0.$$

- S_{Rain}: 5 instances (3 Yes, 2 No)

$$H(S_{\text{Rain}}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) \approx 0.971.$$

Weighted entropy for Outlook:

$$H(S, \text{Outlook}) = \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971) = \frac{10}{14}(0.971) \approx 0.694.$$

Information Gain:

$$I(S, \text{Outlook}) = H(S) - H(S, \text{Outlook}) \approx 0.940 - 0.694 = 0.246.$$

0.5 marks for entropy calculation of S, 0.5 marks for gain calculation of Outlook

Humidity (High, Normal)

- S_{High}: 7 instances (3 Yes, 4 No)

$$H(S_{\text{High}}) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right) \approx 0.986.$$

- S_{Normal}: 7 instances (6 Yes, 1 No)

$$H(S_{\text{Normal}}) = -\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right) \approx 0.592.$$

Weighted entropy for Humidity:

$$H(S, \text{Humidity}) = \frac{7}{14}(0.986) + \frac{7}{14}(0.592) = 0.5(1.578) = 0.789.$$

Information Gain:

$$I(S, \text{Humidity}) = H(S) - H(S, \text{Humidity}) \approx 0.940 - 0.789 = 0.151.$$

0.5 marks for gain calculation of Humidity

Wind (Weak, Strong)

- S_{Weak}: 8 instances (6 Yes, 2 No)

$$H(S_{\text{Weak}}) = -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right) \approx 0.811.$$

- S_{Strong}: 6 instances (3 Yes, 3 No)

$$H(S_{\text{Strong}}) = 1.0.$$

Weighted entropy for Wind:

$$H(S, \text{Wind}) = \frac{8}{14}(0.811) + \frac{6}{14}(1.0) \approx 0.462 + 0.429 = 0.891.$$

Information Gain:

$$I(S, \text{Wind}) = H(S) - H(S, \text{Wind}) \approx 0.940 - 0.891 = 0.049.$$

0.5 marks for gain calculation of Wind

Temperature (Hot, Mild, Cool)

- S_{Hot}: (2 Yes, 2 No), $H(S_{\text{Hot}}) = 1.0$.
- S_{Mild}: (4 Yes, 2 No), $H(S_{\text{Mild}}) \approx 0.918$.
- S_{Cool}: (3 Yes, 1 No), $H(S_{\text{Cool}}) \approx 0.811$.

Weighted entropy:

$$H(S, \text{Temperature}) = \frac{4}{14}(1.0) + \frac{6}{14}(0.918) + \frac{4}{14}(0.811) \approx 0.286 + 0.394 + 0.231 = 0.911.$$

Information Gain:

$$I(S, \text{Temperature}) = 0.940 - 0.911 = 0.029.$$

0.5 marks for gain calculation of Temperature

Step 3: Compare Information Gains

$$I(S, \text{Outlook}) = 0.246,$$
$$I(S, \text{Humidity}) = 0.151,$$
$$I(S, \text{Wind}) = 0.049,$$
$$I(S, \text{Temperature}) = 0.029.$$

Final Answer: The attribute with the highest information gain is *Outlook*.

0.5 marks for final answer with proper calculation, for a total of 3 marks

4. (6 points) (a) (3 points) Justify or refute the following statements. If the statement seems correct, provide a reasonable justification. If it is wrong, provide a recommendation to rectify the mistake. (**Note**: No credit for an incorrect justification or recommendation.)

i. (1 point) My student reported a high training accuracy and claimed that the ML model is correctly trained and ready to be deployed.

ii. (1 point) My student decided upon the max depth of the decision tree by consulting their dog - bark once implies max depth = 5; bark twice implies max depth = 10.

iii. (1 point) My student used the test set for optimizing the hyperparameters of the model and reported the test accuracy and claimed that the model is ready to be deployed.

(b) (3 points) Define the deterministic planning problem mathematically, with all the variables and functions considered clearly defined.

Solution:

(a) i. The statement is incorrect, because reporting only training accuracy is not sufficient for deployment. The student needs to perform cross-validation, or use three-way splits to tune hyperparameters, and report the performance on the test set.

ii. Barking of a dog can be assumed to be a random event. Randomly selecting hyperparameters is not a good strategy, hence the student will need to use a better strategy. Selecting the hyperparameter (max depth in this case) using cross-validation techniques like k-fold cross-validation by selecting the max depth that yields the smallest validation error would be a better choice.

iii. The test set should not be shown to the model for any purpose; strictly only to report the test error. The hyperparameters should be optimized using the training and validation subsets and once selected, the final model performance can be reported on the test set.

For each part, give 0.25 marks for correctly stating that the 'statement is incorrect' + 0.25 marks for stating why the statement is incorrect + 0.5 marks for correct recommendation

(b) The planning problem is to find a plan $P = (\Sigma, s_0, S_g)$, that finds a plan, which is a sequence of actions to transition from the initial state s_0 to one of the goal states $s \in S_g$ in the planning domain Σ as defined below.

1 marks for correct definition of $P = (\Sigma, s_0, S_g)$ and 0.5 marks for defining each term in the expression

The deterministic planning domain is defined as:

$$\Sigma = (S, A, \gamma, cost) \tag{16}$$

where:
S is the finite set of states that the agent can be in.
A is the finite set of actions that the agent can execute.
$\gamma : S \times A \rightarrow S$ is the state transition function, which captures the transition between states dependent on the action taken by the agent.
cost is the cost incurred upon each action taken and it can be a function of the state and action.

1 marks for correct definition of $\Sigma = (S, A, \gamma, cost)$ and 0.5 marks for defining each term in the expression. The *cost* in Σ is optional. Students should not be penalized for ignoring it in the definition.