Problem Set
Probability and Naive Bayes
Instructor: Ipshita Bonhi Upoma
BracU

	0	**1**	**2**
0	0.1	0.07	0.03
1	0.05	0.25	0.1
2	0.05	0.2	0.15

1. Consider two discrete random variables, X and Y, which represent the number of errors found in two different software modules during testing. The joint probability distribution of X and Y is given in the above table:
    - Calculate P(X=1,Y=2).
    - Calculate the marginal probability P(X=1).
    - Calculate the marginal probability P(Y=2).
    - Are X and Y independent? Justify your answer by comparing P(X=1,Y=2) with P(X=1)P(Y=2).

2. A market research company surveys customers regarding their preferences for two product features: Color (Red, Blue) and Size (Small, Large). The survey results are presented in the form of raw data from which the joint probability distribution needs to be derived.

    Raw Data Collected:

    - 100 respondents participated in the survey.
    - 20 respondents preferred Red and Small.
    - 30 respondents preferred Red and Large.
    - 10 respondents preferred Blue and Small.
    - 40 respondents preferred Blue and Large.

    Questions:

a. Construct the joint probability distribution table for the variables Color (C) and Size (S).
b. Calculate the joint probability $P(C=Red, S=Large)$.
c. Calculate the marginal probabilities $P(C=Red)$ and $P(S=Large)$.
d. Determine whether the preferences for Color and Size are independent.
e. What is the conditional probability that a respondent prefers a Small size given that they chose Red color $P(S=Small \mid C=Red)$?

3. A survey is conducted at a university to understand student preferences for study environment based on three factors: Location (Library, Cafe), Time of Day (Morning, Evening), and Noise Level (Quiet, Moderate). The survey involves 200 students, and the results are as follows:

   Raw Data Collected:

   - 40 students preferred the Library in the Morning with a Quiet environment.
   - 20 students preferred the Library in the Morning with Moderate noise.
   - 30 students preferred the Library in the Evening with a Quiet environment.
   - 10 students preferred the Library in the Evening with Moderate noise.
   - 25 students preferred the Cafe in the Morning with Quiet environment.
   - 15 students preferred the Cafe in the Morning with Moderate noise.
   - 45 students preferred the Cafe in the Evening with a Quiet environment.
   - 15 students preferred the Cafe in the Evening with Moderate noise.

   Questions:

   a. Create a joint probability distribution table for the variables Location (L), Time of Day (T), and Noise Level (N).
   b. Calculate the joint probability $P(L=Cafe, T=Evening, N=Quiet)$.
   c. Calculate the marginal probabilities $P(L=Library)$, and $P(N=Quiet)$.
   d. Determine if the factors Location, Time of Day, and Noise Level are independent. Hint: Check if $P(L=Library, T=Morning, N=Quiet)=P(L=Library)P(T=Morning)P(N=Quiet)$.
   e. What is the conditional probability that a student prefers the Library given that it is Morning and the environment is Quiet $P(L=Library \mid T=Morning, N=Quiet)$?

4. Suppose we have three events A, B, and C within a probability space. The events A and B are known to be conditionally independent given C. This means that $P(A \cap B \mid C)=P(A \mid C) \times P(B \mid C)$. Given the following probabilities:
   - $P(A \mid C)=0.4$
   - $P(B \mid C)=0.5$
   - $P(C)=0.2$

   Calculate the probability $P(A \cap B \cap C)$.

5. Let A, B, and C be events in a probability space, where A and B are conditionally independent given C. Assume the following probabilities:
   - $P(A|C)=0.3$
   - $P(B|{^C})=0.6$
   - $P(C)=0.5$

   Calculate: $P(A∩B|C)$, $P(A∩B|{^C})$ assuming A and B are conditionally independent given ^C with $P(A|{^C})=0.2$.

6. In a medical study, events A, B, and D represent having disease A, disease B, and taking drug D respectively. It is known that having disease A and B are conditionally independent given the use of drug D. The probabilities are given as:
   - $P(A|D)=0.4$
   - $P(B|D)=0.5$
   - $P(D)=0.3$
   - $P(A|{^D})=0.2$
   - $P(B|{^D})=0.3$

   Calculate: $P(A∩B|D)$ and, $P(A∩B|{^D})$ assuming A and B are also conditionally independent given ^D.

7. Consider three events E, F, and G in a probability space where E and F are conditionally independent given both G and ^G. You are given:
   - $P(E|G)=0.5$
   - $P(F|G)=0.6$
   - $P(E|{^G})=0.4$
   - $P(F|{^G})=0.3$
   - $P(G)=0.7$

   Calculate: $P(E∩F|G)$ and, $P(E∩F|{^G})$.

8. In a study on social media influence, researchers are trying to understand if sharing political content (Event A) and engaging in political discussions (Event B) are conditionally independent given a user's political affiliation (Event C).

   Given Probabilities: $P(A|C)=0.4$, $P(B|C)=0.5$, $P(C)=0.3$

   Calculate the probability that a randomly selected user from the sample shares political content and engages in discussions, given their political affiliation.

9. A company runs two different types of marketing campaigns simultaneously: email marketing (Event D) and social media ads (Event E). They want to see if these campaigns independently attract new customers (Event F) when considering the customer segment targeted (youth, adults, etc.).

   Given Probabilities: $P(D|F)=0.6$, $P(E|F)=0.7$, $P(F)=0.2$, $P(D|{^F})=0.2$, $P(E|{^F})=0.1$

Calculate the probability that a new customer is attracted by both the email marketing and social media ad campaigns, given that they are a part of the targeted segment.

10. In a school district, students can participate in a sports program (Event G) and a music program (Event H). The district wants to know if participation in these two programs is conditionally independent given the students' grade level (Event I).

    Given Probabilities: $P(G|I)=0.3$, $P(H|I)=0.4$, $P(I)=0.5$.

    Determine the probability that a student participates in both sports and music programs given their grade level.

11. A streaming service uses a Naive Bayes classifier to predict the genre of movies based on two features: presence of action scenes (A) and presence of scary scenes (R). The genres considered are action (X) and horror (Y).

    Given Probabilities: $P(X)=0.6$, $P(Y)=0.4$, $P(A|X)=0.7$, $P(A|Y)=0.3$, $P(R|X)=0.2$ $P(R|Y)=0.8$.

    Calculate the probability that a movie is a horror given that it contains action scenes and scary scenes.

12. A career counseling tool uses Naive Bayes to advise students on potential career paths based on their interest in mathematics (M) and their interest in biology (B). The career paths suggested are engineering (E) and medicine (D).

    Given Probabilities: $P(E)=0.7$, $P(D)=0.3$, $P(M|E)=0.8$, $P(M|D)=0.3$, $P(B|E)=0.2$, $P(B|D)=0.7$.

    What is the probability that a student is advised to pursue medicine given they have an interest in both mathematics and biology?

13. A political analyst uses a Naive Bayes classifier to predict voter behavior based on two issues: support for environmental policies (EP) and support for economic policies (EC). The classifications are progressive voter (P) and conservative voter (C).

    Given Probabilities: $P(P)=0.5$, $P(C)=0.5$, $P(EP|P)=0.8$, $P(EP|C)=0.3$, $P(EC|P)=0.4$, $P(EC|C)=0.7$.

    Estimate the probability that a voter is progressive given their support for both environmental and economic policies.

14. A digital marketing team uses Naive Bayes classification to predict if a user will click on an ad based on their age group (young Y or old O) and browsing history (frequent F or

infrequent I browser).

Given Probabilities:

- $P(Click)=0.3$, $P(No\ Click)=0.7$
- $P(Y \mid Click)=0.4$, $P(Y \mid No\ Click)=0.2$
- $P(F \mid Click)=0.7$, $P(F \mid No\ Click)=0.3$

What is the probability that a user will click on the ad if they are young and a frequent browser?

15. A health insurance company uses Naive Bayes classification to assess the risk of chronic illness based on smoking status (smoker S or non-smoker N) and exercise frequency (regular R or irregular I).

Given Probabilities:

- $P(High\ Risk)=0.25$, $P(Low\ Risk)=0.75$
- $P(S \mid High\ Risk)=0.6$, $P(S \mid Low\ Risk)=0.3$
- $P(R \mid High\ Risk)=0.3$, $P(R \mid Low\ Risk)=0.7$

Calculate the probability that an individual is at high risk for chronic illness if they are a smoker and do not exercise regularly.

16. A political analyst uses Naive Bayes to predict whether a citizen will vote for Party A or Party B based on their age group (young Y, middle-aged M, old O), income level (low L, medium D, high H), and education level (high school HS, college C, postgraduate PG).

Given Probabilities:

- $P(Party\ A)=0.45$, $P(Party\ B)=0.55$
- $P(Y \mid Party\ A)=0.3$, $P(M \mid Party\ A)=0.4$, $P(O \mid Party\ A)=0.3$
- $P(L \mid Party\ A)=0.2$, $P(D \mid Party\ A)=0.5$, $P(H \mid Party\ A)=0.3$
- $P(HS \mid Party\ A)=0.25$, $P(C \mid Party\ A)=0.50$, $P(PG \mid Party\ A)=0.25$

What is the probability that a middle-aged, high-income, college-educated voter will choose Party A?

17. A streaming service uses Naive Bayes to decide whether to show a new sci-fi series or a romantic comedy to a user, based on their previous genre preferences (sci-fi SF, romance RM), viewing time (peak PK, off-peak OP), and subscription type (basic B, premium P).

Given Probabilities:

- $P(Sci\text{-}Fi)=0.6$, $P(Rom\text{-}Com)=0.4$

- $P(SF|\text{Sci-Fi})=0.7$, $P(RM|\text{Rom-Com})=0.8$
- $P(PK|\text{Sci-Fi})=0.8$, $P(OP|\text{Rom-Com})=0.6$
- $P(B|\text{Sci-Fi})=0.5$, $P(P|\text{Rom-Com})=0.5$

Estimate the probability that a premium user, who prefers sci-fi and watches during peak times, will be shown the new sci-fi series.

18. A health app predicts whether a user is at low or high risk for diabetes based on their physical activity level (active A, sedentary S), diet type (balanced B, high-sugar HS), and family history (yes Y, no N).

    Given Probabilities:

    - $P(\text{Low Risk})=0.7$, $P(\text{High Risk})=0.3$
    - $P(A|\text{Low Risk})=0.8$,$P(S|\text{High Risk})=0.7$
    - $P(B|\text{Low Risk})=0.9$, $P(HS|\text{High Risk})=0.6$
    - $P(Y|\text{High Risk})=0.4$, $P(N|\text{Low Risk})=0.85$

    What is the probability that a sedentary user with a high-sugar diet and a family history of diabetes is at high risk?

19. A political analyst uses Naive Bayes to estimate support (Support S or Oppose O) for a candidate based on voter registration status (Registered R, Not Registered N) and past voting frequency (Frequent F, Infrequent I, Never V).

    Given Probabilities:

    - $P(S)=0.7$, $P(O)=0.3$
    - $P(R|S)=0.9$, $P(N|S)=0.1$
    - $P(F|S)=0.6$, $P(I|S)=0.3$, $P(V|S)=0.1$
    - $P(R|O)=0.6$, $P(N|O)=0.4$
    - $P(F|O)=0.2$, $P(I|O)=0.5$, $P(V|O)=0.3$

    What is the probability that a voter supports the candidate if they are registered and have never voted before?

20. A tech company uses a Naive Bayes Classifier for screening resumes to predict the suitability of candidates for software engineering roles. The classifier uses features such as "coding experience" (C), "university ranking" (U), and "candidate race" (R). The candidate race feature can take two values: "minority" (M) and "non-minority" (N). The classifier has been trained on historical data, and there are concerns about racial bias. The company wants to assess if there is a bias toward "non-minority" candidates.

    Given Probabilities:

    - $P(\text{Hire})=0.7$, $P(\text{Not Hire})=0.3$

- $P(C \mid Hire)=0.9$, $P(U \mid Hire)=0.85$
- $P(C \mid Not\ Hire)=0.4$, $P(U \mid Not\ Hire)=0.3$
- $P(M \mid Hire)=0.2$, $P(M \mid Not\ Hire)=0.8$
- $P(N \mid Hire)=0.8$, $P(N \mid Not\ Hire)=0.2$

Calculate the probability that a "minority" candidate with excellent coding experience and from a top-ranking university is hired, and compare it with the probability that a "non-minority" candidate with the same qualifications is hired. Use these calculations to discuss potential racial bias in the hiring algorithm.

21. A factory produces 90% of the widgets and a competitor produces 10%. 5% of the factory's widgets are defective, and 10% of the competitor's widgets are defective. If a randomly selected widget is defective, what is the probability that it was produced by the factory?

22. In a town, 60% of people own a car, and 40% do not. The probability that a person with a car gets into an accident in a year is 0.03, while the probability of someone without a car getting into an accident is 0.01. If a randomly selected person got into an accident this year, what is the probability that they own a car?

23. A hospital test for a disease has the following characteristics:
    - The test correctly identifies 95% of people with the disease (True Positive Rate).
    - The test correctly identifies 90% of people without the disease (True Negative Rate).
    - 2% of the population actually has the disease.

    If a person tests positive, what is the probability they actually have the disease?

24. A student is taking an exam with two types of questions: multiple-choice and true/false. The student is equally likely to encounter either type of question. The student answers 70% of multiple-choice questions correctly and 80% of true/false questions correctly. If the student answered a question correctly, what is the probability that it was a multiple-choice question?

25. In a factory, there are two machines (Machine A and Machine B) producing parts. Machine A produces 60% of the parts, and Machine B produces 40%. The defect rates for the machines are:
    - Machine A: 3% of the parts are defective.
    - Machine B: 5% of the parts are defective.

    A defective part is selected at random. What is the probability that it was produced by Machine A?