Exercises Guideline

1. Dissimilarity Between Asymmetric Binary Variables

You are given a dataset describing whether **three monitoring stations** (S1, S2, S3) detected certain environmental conditions during an inspection cycle. All attributes are **asymmetric binary variables.** 1 indicates the condition was detected, 0 indicates no detection.

Station	High-CO_2	Smoke	High-Noise	Vibration
S1	1	0	1	0
S2	1	1	0	0
S3	0	1	1	1

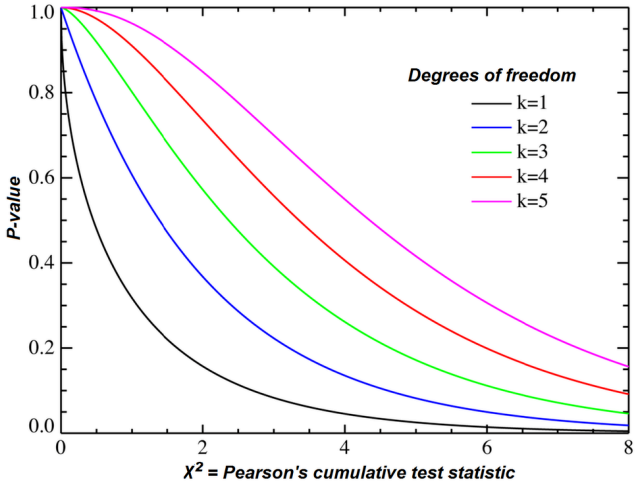
- For each pair of stations, construct the contingency table.
- For each pair of stations, compute the distances.
- Interpret the results:
  - Which two stations are most similar?
  - Which are most dissimilar?
  - How does ignoring double-zero matches (asymmetry) impact the distances?
  - What differences would occur if these were symmetric variables?

2. Chi-Square Calculation

You are given a contingency table showing whether people **own a dog** (row categories) and whether they **attend a fitness gym** (column categories). Each cell contains the **observed count** of people in that category. The last row and column show marginal totals.

	Attend Gym	Do not attend gym	Sum (row)
Owns a dog	45	55	100
Does not own a dog	30	90	120
Sum (col.)	75	145	220

- For each cell, compute the expected value.
- Compute the chi-square statistic.
- Determine the degrees of freedom, report the correlation confidence level, and decide whether to reject or accept the null hypothesis.



3. Attribute Selection with Information Gain

Given the following dataset, the goal is to predict the class label **“Subscribes Fitness Program"** based on several categorical attributes.

Age	Workout Frequency	Diet Quality	Membership Status	Subscribes Fitness Program
young	low	poor	none	no
young	medium	average	basic	no
middle	low	poor	none	yes
senior	medium	average	premium	yes
senior	high	good	premium	yes
senior	medium	average	none	no
middle	high	good	premium	yes
young	medium	average	none	no
young	high	good	basic	yes
senior	low	poor	none	no
young	high	good	premium	yes
middle	medium	good	basic	yes
middle	medium	average	premium	yes
senior	high	good	basic	no

Select the attributes with the highest information gain to be used in a **decision tree classifier**.
Use entropy as a measure of the expected needed information.
Which attribute would be chosen as the root attribute in a decision-tree split (according to the information-gain criterion)?

4. Naïve Bayes Classifiers

You are given the following dataset of students and whether they **pass a course** based on several categorical attributes.

study_hours	attendance	previous_grade	extra_practice	pass_course
low	low	C	no	no
low	low	B	no	no
medium	high	B	no	yes
high	medium	B	no	yes
high	medium	C	yes	yes
high	low	C	yes	no
medium	low	B	yes	yes
low	medium	C	no	no
low	high	C	yes	yes
high	medium	B	yes	yes
low	medium	B	yes	yes
medium	medium	B	no	yes
medium	high	B	yes	yes
high	medium	B	no	no

Use a **Naïve Bayes classifier (categorical)** to classify the following new student:

- study_hours = low
- attendance = medium
- previous_grade = B
- extra_practice = yes

Compare the two final probabilities and **give the final classification (pass or not pass)**.

5. Confusion Matrix Evaluation

A machine-learning model is used to classify emails as **Spam (Positive)** or **Not Spam (Negative)**. After evaluating the model on **200 emails**, the following confusion matrix is obtained:

	Actual Spam	**Actual Not Spam**
Predicted Spam	72	18
Predicted Not Spam	28	82

Using this confusion matrix, compute:

- **Sensitivity**
- **Specificity**
- **Accuracy**
- **Error Rate**
- **Precision**
- **Recall**
- **F1-measure**

Discuss whether this is a good, bad, or average model.