

1. Decision Trees (20 points)

Predicting whether a machine learning classifier will perform well on a given dataset is an important problem tackled by fields such as AutoML. You decide to use a decision tree to make this prediction based on certain attributes of the dataset and the classifier. You have collected the following data from 8 previous experiments:

Dataset Size	Feature Count	Classifier	Performed Well?
Small	Low	SVM	No
Large	High	Random Forest	Yes
Small	High	Neural Network	No
Large	Low	SVM	Yes
Small	Low	Random Forest	Yes
Large	High	Neural Network	Yes
Small	High	SVM	No
Large	Low	Random Forest	Yes

1. (10 points) Calculate the entropy of the target variable (Performed Well)

2. (10 points) Calculate the information gain for the attribute dataset size.

2. K-Nearest Neighbors (10 points)

There is a distance-based variant of the nearest neighbor algorithm. In this variant, you choose **all** points that are within a distance of d from the query point. Then you choose the class label using majority rule. For a fixed distance d , you want to make sure that your distance function does not retrieve too many points. Given the choice between Manhattan, Euclidean and Chebyshev distance functions, which would you choose and why? (Hint: think about the contour lines imposed by these functions and their respective areas).

3. Naïve Bayes Classifiers (15 points)

In the vast multiverse of "Comicopia," superheroes and supervillains abound. You are tasked with designing a classifier that can determine whether the newly discovered superpowered individuals across different realities as either "Hero" or "Villain". You have designed a Naïve Bayes classifier with three attributes:

1. Origin (Tragic/Privileged)
2. Power Source (Science/Magic)
3. Costume (Bright/Dark)

The classifier works well in many universes but fails catastrophically in others, leading to multiversal chaos. Your task is to unravel this paradox and propose solutions. Let us consider some cases.

Think carefully and give short answers with at most 3-4 sentences.

a. (5 points) Quantumania Universe: In this universe, it's discovered that a hero's Origin and Power Source are quantum-entangled. So, if you know the origin, you can accurately predict the value of power source. Explain how this phenomenon violates Naive Bayes assumptions and how it might lead to misclassifications.

b. (5 points) The Multiverse Dilemma: You noticed that the classifier's accuracy varies wildly across different universes. In some universes, it's nearly perfect; in others, it's no better than random guessing. Propose a hypothesis for this phenomenon, considering how the relationship between features and class labels might vary across realities.

c. (5 points) The Anti-Matter Universe: You discover a universe where all correlations between features and classifications are exactly inverted compared to the training data in the your universe. How would the Naive Bayes classifier perform in this universe? Explain your reasoning. If your classifier gets an accuracy of p in the prime universe, what will it get in the anti-matter universe?

4. Performance Evaluation of Classifiers (30 points)

Imagine you're a data scientist working for a major comic book publisher. You've been tasked with developing a machine learning model to predict which new comic book ideas will become bestsellers. The company wants to use this model to decide which comics to publish. You use accuracy, precision and recall as the performance metrics to evaluate the classifier.

Please be precise and concise. Each of these questions can be answered in 3-4 sentences.

a. (7.5 points) Most comic book ideas do not become bestsellers. How would this affect the use of accuracy and f-score in this situation?

b. (7.5 points) The company gets 100s of comic book ideas but can only publish 10 of these ideas. Given this constraint, would you recommend focusing on precision or recall? Justify your answer.

c. (7.5 points) Marvelous Comics' model achieves 90% accuracy in predicting bestsellers. However, when applied to comics from a newly acquired indie publisher, its accuracy drops to 60%. Explain how this could happen even if the model's precision and recall remain constant across both datasets.

d. (7.5 points) Describe how precision and recall would change if you switched the positive and negative labels in this binary classification problem (i.e., "non-bestseller" becomes the positive class and "bestseller" becomes the negative class). What does this tell you about these metrics?

5. Model Selection (25 points)

Consider three classifiers – KNN, Decision trees, and Naïve Bayes. For each of the circumstances described below, please provide a ranking of these three classifiers, **on average** (i.e., do not worry about extreme cases). So, a ranking of KNN, DT, NB means that KNN performs best while NB performs worst. If the information provided is insufficient to provide a full ranking, it is okay to provide ties. For example, KNN, DT-NB means that KNN performs best while NB and DT perform similarly.

Explain your justification in at most 3-4 sentences.

a. The dataset contains lot of attributes but only a small proportion of them are relevant.

b. At least 20% of the cells in the dataset have missing values.

c. The attributes have complex interactions between themselves that influences the target variable.

d. At least 10% of the labels in the training dataset are incorrect.

e. You do not have any control over feature engineering and the quantitative attributes are all in different scales.