11320BAI500200 Introduction of Statistics and Machine Learning (II)

Homework (Bayes and Tree)

1. [8 points] Please select the most suitable type of machine learning approaches from "supervised learning", "unsupervised learning", "semi-supervised learning", "reinforcement learning", and "not a learning problem" which is best fit for the following using scenarios. Please first choose one of the five answers and explain your answer by specifying the input data, labels, and your reasons.
   (a) A model is trained using patient medical records, including genetic information and family history, to predict the likelihood of developing diseases like Alzheimer's or diabetes in the future. The dataset includes past cases labeled with whether the patient developed the disease or not.
   (b) A system is designed to optimize drug dosage recommendations for diabetes patients by continuously adjusting medication levels and monitoring blood sugar responses. It learns through trial and error, aiming to maintain stable glucose levels while minimizing side effects. The system receives feedback based on patient outcomes and refines its dosage recommendations over time.
   (c) A hospital's billing system calculates patient invoices based on standardized rates for treatments, medications, and services used. The system follows a predefined set of rules and does not improve over time based on data.
   (d) A wearable heart monitor records patients' heartbeats and detects unusual patterns that could indicate potential heart conditions. The system does not have predefined examples of anomalies but is learned by identifying patterns that deviate from normal heart activity.
2. [10 points] Assume a disease is so rare that it is seen in only one person out of every million. Assume also that we have a test that is useful in that if a person has the disease, there is a 90 percent chance that the test result will be positive; however, the test is not perfect, and there is a one in a thousand chance that the test result will be positive on a healthy person. Assume that a new patient arrives, and the test result is positive. What is the probability that the patient has the disease? Please write down the procedure to show how you derive your answer.

3. Given a dataset with features of weight loss, headache, fever, and cough representing the symptoms of patients. The last column "Prescription" indicates the use of a specific medicine. The task is to build a naïve bayes model to predict the "prescription" (a binary classifier).

Weight Loss	Headache	Fever	Cough	Prescription
Obvious	Yes	Yes	Yes	Yes
Obvious	Yes	Yes	No	Yes
No	Yes	Yes	Yes	No
Mild	No	Yes	Yes	No
Mild	No	No	Yes	No
Mild	No	No	No	Yes
No	No	No	No	No
Obvious	No	Yes	Yes	Yes
Obvious	No	No	Yes	No
No	No	Yes	No	Yes
Mild	No	No	Yes	Yes
Obvious	No	Yes	No	Yes
Obvious	Yes	No	No	Yes
Mild	No	No	No	No

(a) [10 points] Given that the features (Weight Loss, Headache, Fever, Cough) of a patient are (Obvious, Yes, No, No), please predict whether to use the medicine (prescription) using maximum a-posterior approach (naïve bayes). Derive the final answer by calculating prior, likelihood, and posterior. Please provide the procedure and details calculated by hand.

(b) [10 point] Please write a code for a naïve bayes classifier and validate your calculation in the previous question. Do not use any packages directly implementing the classifier. Please comment on each function with its inputs (parameter) and outputs (return) to obtain full credits in this problem. You can refer to classifer.py for some coding styles.

(a) [10 points] Please download the following dataset and run a naïve bayes model using scikit-learn. Please perform a 5-fold cross-validation. In each fold, we need 70% for training, 10% for validation, and 20% for testing. Please report the final performance using the classification report function in Scikit-learn.

Scikit-learn: https://scikit-learn.org/stable/

Download link: https://www.kaggle.com/datasets/manishkc06/patient-treatment-classification

(b) [2 points] What is the difference when we have features in discrete and continuous values (the previous coding problem versus this coding problem)?

Hint: If the features are continuous values, how do we compute likelihood?

4. Following the dataset in problem 4, please build a decision tree model to predict the "prescription" (a binary classifier). Note that we allow both 2-way and 3-way splits.

(a) [3 points] What is the entropy $\mathrm{H(Prescription)}$?

(b) [3 points] What is the entropy $\mathrm{H(Prescription|Weight\ Loss)}$?

(c) [3 points] What is the entropy $\mathrm{H(Prescription|Headache)}$?

(d) [5 points] Draw the full decision tree that would be learned for this dataset.

5. Please download the following dataset and run the tree-based models using scikit-learn and xgboost. Please split the data with 70% for training, 10% for validation, and 20% for testing, and report the performance in the three sets of data using classification report function in scikit-learn.

Scikit-learn: https://scikit-learn.org/stable/
Xgboost: https://xgboost.readthedocs.io/en/stable/get_started.html
Download link: https://www.kaggle.com/datasets/manishkc06/patient-treatment-classification

(a) [2 points] Please build a decision tree model.

(b) [2 points] Please build a random forest model.

(c) [2 points] Please build a gradient-boosting ensemble model.

(d) [10 points] Tune the parameters to find the best performance for the models built in the previous questions (a)(b)(c). Report the full results with your tuned parameters and illustrate the logic of your model selection procedure. What are the most influential parameters for each model (describe at least two parameters)?

(e) [5 points] Plot the feature importance values in a figure. Please perform feature selection by setting a threshold for feature importance and build a new model. Does the performance become better?

(f) [10 points] This experiment simulates the different sizes of the available dataset. Firstly leave 10% of data out for testing. Second, randomly downsample the left data with 90%, 80%, 70%, and so on, until 10% for model training and validation. You can choose to only implement the best-performed model (either decision tree, random forest, or gradient boosting) obtained in the previous problem. Please plot the performance curve with the different downsampling percentages.

(g) [5 points] Compare the above tree-based approaches with the naïve bayes classifier. Please state which classifier performs better and explain with reason. Your answer should include explanations about the differences between classifiers and the reason that the classifier is suitable for the data.

Language Models Usage Policy for This Homework

We allow students to use language models for coding problems, i.e., the delivery of Python code. Other problems are not allowed to use language models.

If students use any language model for their homework, they must:

1. **Describe how the language model was used** and provide the prompts (e.g., "I asked GPT to write the posterior calculation function of Bayes' Theorem with the prompt, …, then I refactor the code.").
2. **Show how they modified the response** (e.g., highlighting changes, adding their thoughts, or expanding explanations).
3. **Reflect on what they learned** (e.g., "GPT helped me understand reinforcement learning, but I also researched additional sources to confirm the explanation").

If students do not clearly explain their use of language models, the problem will receive a zero score.