CS 4/565: Introduction to AI

Project 3: Naive Bayes for Binary Classification

Due: April 17th, 11:59 PM

1 Introduction

Implement a Naive Bayes classifier from scratch (no machine learning libraries) to classify whether a mushroom is edible or poisonous using the UCI Mushroom dataset. The UCI Mushroom dataset is a widely used benchmark for classification tasks, particularly with categorical data. It consists of 8,124 instances, each representing a mushroom. It aims to predict whether the mushroom is `edible (e)` or `poisonous (p)` based on 22 categorical attributes such as Cap shape, surface, Odor, Spore print color, and so on. The dataset is well-suited for evaluating classification models such as Naive Bayes, decision trees, and association rule learners. For more information, visit the UCI Mushroom Dataset page.

2 Tasks

We need to implement the Naive Bayes Classification manually (i.e., you can not use the library directly). Please fill the following task 1 and task 2, with a report to introduce them.
Task 1: I have uploaded the code with missing steps; please finish the missing code (step 5 and step 6) for the naive Bayes classification.

Step 5) Training Phase: Compute the prior probability $P(c)$ for each class c, i.e., the frequency of each class in the training data. For each class c and each feature x_j, compute the conditional probability $P(x_j \mid c)$, representing the likelihood of feature value x_j given class c.

Step 6) Prediction Phase: For each test sample, compute the log posterior probability for each class:

$$\log P(c) + \sum_j \log P(x_j \mid c)$$

Predict the class with the **highest log posterior** as the output label.
Task 2: Please only use the 10 categorical attributes in the original dataset and then redo the training and test process.

3 Deliverables

There are two deliverables: report and code.

1. **Report (30 points)** The report should be delivered as a separate pdf file, and it is recommended for you to use the NIPS template to structure your report. You may include comments in the Jupyter Notebook, however, you will need to duplicate the results in the report. The report should describe your results, experimental setup, details, and comparison between the results obtained from different settings of the algorithm and dataset.

2. **Code (70 points)**

   The code for your implementation should be in Python only. The name of the Main file should be project3.ipynb. Please provide necessary comments in the code and show some essential steps for your group work.