Machine Learning

Lab 05: Bayesian Classification

The goal of this lab exercise is to implement the naïve Bayesian classification algorithm.

Task 1.

In this task you will again use the Titanic passenger dataset, which you can download from Canvas. The goal is to train a naïve Bayesian classifier for predicting survival based on gender and passenger class.

A. Load the CSV file “titanic.csv” using Pandas. Extract the columns “Sex” and “Pclass” to use as feature vector, and the column “Survived” to use as target vector.
B. Split the dataset into training and test data.
C. Count the number of passenger, the number of survivors, and the number of casualties in the training data and calculate the priors $P[survived]$ and $P[casulty]$.
D. Count the number of male survivors, female survivors, male casualties, and female casualties in the training data and calculate the likelihoods $P[male|survived]$, $P[female|survived]$, $P[male|casulty]$, $P[female|casulty]$ applying Laplace smoothing with the parameter α$=10$.
E. Count the number of 1st, 2nd and 3rd class survivors, and 1st, 2nd, and 3rd class casualties in the training data and calculate the likelihoods $P[1st\ class|survived]$, $P[2nd\ class|survived]$, $P[3nd\ class|survived]$, $P[1st\ class|casulty]$, $P[2nd\ class|casulty], and\ P[3rd\ class|casulty]$ applying Laplace smoothing with the parameter α$=10$.

Task 2.

In this task you will implement and evaluate the naïve Bayesian classifier for predicting survival based on gender and passenger class on the test set.

A. Now go through all passengers in the test set and calculate the for each of these the posteriors $P[survived|sex, class]$ and $P[casulty|sex, class]$. Make sure to use logarithms to guarantee numerical stability. Compare the likelihood-ratio to the prior-ratio to predict survival for each passenger in the test set.
B. Calculate the confusion matrix to evaluate your classifier.