PRIMROSE
Machine Learning Experts

Naive Bayse Exercise

Today, we are going to implement the Naive Bayes classifier in python and test it on the Pima Indians Diabetes dataset. The dataset classifies instances of people with/without diabetes based on different features, such as number of pregnancies, blood pressure, BMI, etc. While you are free to make your own implementation, it is recommended to follow the implementation steps below and test each one to make sure it works properly.

1. **Handle the Data**: Load the data from the CSV file in the following link and split it into training and test data: https://www.kaggle.com/uciml/pima-indians-diabetes-database
2. **Summarize the Data (train)**: summarize the properties in the training data by calculating every feature and class' (prediction value) mean and standard deviation (std).
3. **Write a prediction function**: use the summaries of the dataset (what you calculated in step 2) to generate a single prediction, which is based on the Gaussian distribution with the corresponding mean and std of each feature. You can find the equation for the probability of an event given a Gaussian distribution in: https://en.wikipedia.org/wiki/Naive_Bayes_classifier#Gaussian_naive_Bayes
4. **Make Predictions**: Generate predictions on the whole test data.
5. **Evaluate Accuracy**: Evaluate the accuracy of predictions made for the test data as percentage of correct predictions out of all predictions made.
6. **Tie it Together**: Use all of the code elements to present a complete and standalone implementation of the Naive Bayes algorithm.
   * (Optional) Try building it into a class with fit(train) method, which calculates the mean and std, and predict(test) method, which makes a Naive Bayes prediction for the test data.

We are going along the instructions from the following link:

http://machinelearningmastery.com/naive-bayes-classifier-scratch-python/