**Homework Set 10**

**Problem 1 (Naïve Bayes)**

Naïve Bayes is a simple and eﬀective machine learning approach. In this assignment you will determine a person’s willing in purchasing a computer using Naïve Bayes. Suppose we have collected historic data of a person’s purchase willing with the attributes: 1) **Age**: how old the person is, 2) **Income**: the income of the person, 3) **Student**: if the is a student, and 4) **Credit rating**: credit rating of the person. The historic data is shown in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age** | **Income** | **Student** | **Credit rating** | **Purchase** |
| ≤30 | High | No | Fair | No |
| ≤30 | High | No | Excellent | No |
| 31-40 | High | No | Fair | Yes |
| >40 | Medium | No | Fair | Yes |
| >40 | Low | Yes | Fair | Yes |
| >40 | Low | Yes | Excellent | No |
| 31-40 | Low | Yes | Excellent | Yes |
| ≤30 | Medium | No | Fair | No |
| ≤30 | Low | Yes | Fair | Yes |
| >40 | Medium | Yes | Fair | Yes |
| ≤30 | Medium | Yes | Excellent | Yes |
| 31-40 | Medium | No | Excellent | Yes |
| 31-40 | High | Yes | Fair | Yes |
| >40 | Medium | No | Excellent | No |

Suppose a person with attribute value {**Age**≤ 30, **Income** = medium, **Student** = yes, **Credit rating**= Fair} walks into a store. Determine if the person will buy a computer or not using the naïve Bayes method.

**Problem 2 (k-nearest neighbor)**

This problem considers performing kNN classification using different distance measures. Given two vectors and . Define a modified distance measure as:

Consider the following labelled training data points:

1. Develop a 1-NN classifier on these points using the usual Euclidean distance. Draw the decision boundary for this classifier. Write down the equations for the different sections of the decision boundary. Clearly mark each region in your drawing with the label assigned by the classifier.
2. Repeat the procedure in part (a) but using the defined distance measure . In a separate figure, draw the decision boundary for this classifier. Write down the equations for the decision boundary.

**Problem 3 (k-nearest neighbor)**

In this problem, you are asked to classify images of digits using kNN. Download the files “10HW3\_train.txt”, “10HW3\_test.txt“, and “10HW3\_validate.txt” from the class website. These files contain your training, test, and validation datasets. The digit images are already converted into vectors of pixel colors. The data files are in ASCII text format, and each line of the files contains a feature vector of size 784, followed by its label. The coordinates of the feature vector are separated by spaces.

1. For *k* = 1, 3, 5, 11, 16, and 21, build kNN classifiers from the training data. For each of these values of *k*, write down a table of training errors (error on the training data) and the validation errors (error on the validation data). Which of these classifiers performs the best on validation data? What is the test error of this classifier?
2. Construct a 3-NN classifier from the training data. Compute the confusion matrix of the classifier based on the data in “10HW3\_test.txt“. The confusion matrix is a 10×10 matrix, where each row is labelled 0,…,9 and each column is labelled 0,…,9. The entry of the matrix at row *i* and column *j* is *Cij*/*Nj*, where *Cij* is the number of test examples that have label *j* but are classified as label *i* by the classifier, and *Nj* is the number of test examples that have label *j*. Based on your answers, which digits do you think are the easiest and the hardest to classify?
3. Identify one falsely classified vector. Convert the vector back to a 28×28 image. Report your observations.