Introduction to Machine Learning

Assignment 1

Submission date: 02/01/2020 , 23:59

General Instructions:

- The solution should be formatted as a report and running code should be included in a digital form.
- The solution can be done in pairs independently to other groups. Identical (or very similar solutions) are not allowed!
- You may choose whether to write your code in Python or Matlab.
- Please submit the assignment via Moodle.
- The code must be reasonably documented.

Problem 1: MNIST Classification

In this problem, we will perform handwritten digits classification using KNN algorithm.
Data: Download the training and test set from: http://yann.lecun.com/exdb/mnist/
Task:

A. Split the train set randomly into train and validation set. 80% of the training examples uses for training the model (training set), and 20% uses as a validation set to find the best parameter k for the KNN classifier.
B. Implement the KNN classifier and report the classification error.
    1. Report the performance of the classifier and argue which value of k you would choose according to the validation set. What is the classification rate on validation set of your chosen value k*?
    2. Compute the accuracy on test set of your chosen k*. Does the test performance correspond to the validation performance? Why or why not?

Theoretical Questions:

Question 1:

1. Suppose that we are given an independent and identically distributed sample of n points $\{y_i\}$ where each point $y_i \sim N(\mu, 1)$. Suppose that we use the estimator $\hat{\mu}$= 1 for the mean of the sample. Give the bias and variance of this estimator $\hat{\mu}$. Explain in a sentence whether this is a good estimator in general, and give an example of when this is a good estimator.
2. In this question you are going to analyze errors of Bayesian classifiers. Suppose that:
   - Y is boolean, X is real
   - $\mathrm{P(Y = 1) = 1/2}$
   - $\mathrm{p(X|Y = 1) = uniform[1, 4]}$
   - $\mathrm{p(X|Y = 0) = uniform[-4, -1]}$.

   a. Plot the two class conditional probability distributions $\mathrm{p(X|Y = 0)}$ and $\mathrm{p(X|Y = 1)}$.
   b. What is the error of the optimal classifier? Note that the optimal classifier knows $\mathrm{P(Y = 1)}$, $\mathrm{p(X|Y = 0)}$ and $\mathrm{p(X|Y = 1)}$ perfectly, and applies Bayes rule to classify new examples.
   c. Suppose instead that P(Y = 1) = 1/2 and that the class conditional distributions are uniform distribution with $\mathrm{p(X|Y = 1) = uniform[0, 4]}$ and $\mathrm{p(X|Y = 0) = uniform[-3, 1]}$. What is the error in this case? Justify your answer.
   d. Consider again the learning task from part (a) above. Suppose we train a Gaussian Naive Bayes classifier using n training examples for this task, where $\mathrm{n} \rightarrow \infty$.Of course our classifier will now (incorrectly) model $\mathrm{p(X|Y)}$ as a Gaussian distribution, so it will be biased: it cannot even represent the correct form of $\mathrm{p(X|Y)}$ or P(Y |X). Draw again the plot you created in part (a), and add to it a sketch of the learned/estimated class conditional probability distributions the classifier will derive from the infinite training data. Write down an expression for the error of the Gaussian Naive Bayes. (hint: your expression will involve integrals - please don't bother solving them).

Question 2:

Suppose a bank classifies customers as either good or bad credit risks. On the basis of extensive historical data, the bank has observed that 1% of good credit risks and 10% of bad credit risks overdraw their account in any given month. A new customer opens a checking account at this bank. On the basis of a check with a credit bureau, the bank believe that there is a 70% chance the customer will turn out to be a good credit risk.

a. Suppose that this customer's account is overdrawn in the first month. How does this alter the bank's opinion of this customer's creditworthiness?

b. Given (a), what would be the bank’s opinion of the customer’s creditworthiness at the end of the second month if there was not an overdraft in the second month?