1. **MNIST Database**

   One of the most famous applications of machine learning is classification of handwritten digits. A well known benchmark database that is extensively used in the literature is the MNIST, which contains a large number of handwritten digits, centered and normalized into fixed-size images (28×28 pixels).

   (a) Download the MNIST data from: `http://yann.lecun.com/exdb/mnist/` or `https://archive.ics.uci.edu/ml/databases/mnist/`. You will see the following

   - `database_train_images`- A 60000×784 matrix. Each row is an image, of size 28×28, unrolled into a vector.
   - `database_train_labels` - A 60000×1 vector. Each row is the digit that corresponds to `database_train_images`.
   - `database_test_images` - A 10000×784 matrix with test images.
   - `database_test_labels` - A 10000×1 vector with corresponding digits.

   (b) Visualization of data from MNIST database

   Many machine learning problems come from real-life applications. If possible, it is important to visualize the data, in order to gain intuition, and learn what methods we want to apply. In our case, the data observations are 28×28 gray-scale images of digits.

   i. Plot the first 25 images in the training set. Do all images of 9 look alike?

   ii. Plot 15 randomly selected images from the test set without looking at the corresponding labels and try to guess them. Were all of your guesses correct?

   iii. Let us explore the data even more. Find 2 different digits that look alike. Find 3 samples of the same digit that do not look alike at all.

   (c) Classification using KNN on MNIST database

   i. What is the nearest neighbor of a train sample, assuming it is included in the training set?

   ii. Write code for k-nearest neighbors with Euclidean metric (or use a software package). Find 5 nearest neighbors for the first 10 test samples and plot them together.

   iii. Test all 10000 digits in the test database with k nearest neighbors. Take decisions by majority polling. Plot train and test errors in terms of $1/k$ for $k \in \{1, 201, 401, \ldots, 10001\}$. You are welcome to use smaller increments of k. Which k^* is the most suitable k among those values?

   iv. Since the computation time depends on the size of the training set, one may only use a subset of the training set. Plot the *best error rate*, which is obtained by some value of k, against the size of training set, when the size of training set is $N \in \{5000, 10000, 15000, \ldots, 55000, 60000\}$. (You are welcome to choose smaller increments of N).

   Let us further enhance the classification error.

   (d) Plot the k nearest neighbors of some of misclassified samples.

(e) Replace the Euclidean metric with the following metrics and test them. Summarize the test errors (i.e., when $k = k^*$) in a table.

i. Minkowski Distance:

A. which becomes Manhattan Distance with $p = 1$

B. with $\log_{10}(p) \in \{0.1, .2, .3, \ldots, 1\}$

C. which becomes Chebyshev Distance with $p \to \infty$

ii. Mahalanobis Distance

iii. Hausdroff Distance

iv. (More Fun) with Hausdroff Distances 1, 2, 3 introduced in the paper *Hausdorff Distance with k-Nearest Neighbors by Jun Wang and Ying Tan*

(f) Replace the majority polling decision with another reasonable method devised by yourself. Use it with Euclidean, Manhattan, and Chebyshev distances and report the best test errors.

(g) What is the lowest error rate you achieved in this exercise?

2. **Forest Fire Data**

In this exercise, we investigate a difficult regression task, where the aim is to predict the burned area of forest fires in the northeast region of Portugal.

(a) Download the Forest Fire data from: `https://archive.ics.uci.edu/ml/datasets/Forest+Fires`.

(b) Exploring the data:

i. How many rows are in this data set? How many columns? What do the rows and columns represent?

ii. Explain why the transformation $Y_1 = \ln(1 + Y)$, where Y is the response variable is useful for this dataset. In the following, use Y_1 as the new response variable.

iii. Make pairwise scatterplots of the predictors (columns) in this data set with the dependent variable. Describe your findings.

iv. Make at least 16 pairwise scatterplots of predictors of your choice and describe your findings. You are welcome to make all possible scatter plots.

v. What are the mean, the median, range, first and third quartiles, and interquartile ranges of each of the variables in the dataset? Summarize them in a table.

(c) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions. Are there any outliers that you would like to remove from your data for each of these regression tasks?

(d) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

(e) How do your results from 2c compare to your results from 2d? Create a plot displaying the univariate regression coefficients from 2c on the x-axis, and the multiple regression coefficients from 2d on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

(f) Is there evidence of nonlinear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form

$$Y_1 = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

(g) Is there evidence of association of interactions of predictors with the response? To answer this question, run a full linear regression model with all pairwise interaction terms and state whether any interaction terms are statistically significant.

(h) Can you improve your model using possible interaction terms or nonlinear associations and between the predictors and response? Train the model on a randomly selected 70% subset of the data and test it on the remaining points and report your train and test results.

(i) KNN Regression: Note that for this problem, we have a mixture of categorical and quantitative predictors. There is not a unique way to define a distance metric in such a situation. Describe your findings and heuristics. Can your metric be specific to this problem? Use a reasonable distance metric to answer the following questions:

i. Use the first 4 predictors to perform k-nearest neighbor regression for this dataset. Find the value of k that gives you the best fit. Plot the train and test errors in terms of $1/k$.

ii. Use the last 4 predictors to perform k-nearest neighbor regression for this dataset. Find the value of k that gives you the best fit. Plot the train and test errors in terms of $1/k$.

iii. Use predictors 1,2, 9, 10, 11 to perform k-nearest neighbor regression for this dataset. Find the value of k that gives you the best fit. Plot the train and test errors in terms of $1/k$.

(j) Compare the results of KNN Regression with linear regression and provide your analysis.