Assignment 1
Classical Machine Learning methods

Part 1

Note:

1. Solve all the problems using MATLAB or any other computer language.
2. After calculating the best model given your data, review the given data and you may remove any possible outliers (a value with a possible error or high noise), then recalculate the model for cleaned data. Compare the two models. This exists only in problems 1 and 5.
3. Try by hand the linear and the logistic problems as well. You may use an excel sheet for this calculation.
4. In all your answers to each question, write down the equations of your solutions (after calculating their parameters).
5. Hint: it is better to take the model of a lower number of parameters if the gain in R^2 is not high enough.

1. The numbers of insured persons y with an insurance company for the years 1987 to 1996 are shown in the table.

Year	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
y	14000	13000	12000	11000	1050	10000	9500	9000	8700	8000

Make a scatterplot of the data, letting x represent the number of years since 1987.

a) Fit linear, quadratic, cubic, by comparing the values of R^2. Determine the function that best fits the data. (Hint: take care of note 4 above)
b) In all your answers in each model, write down the equations of your solutions (after calculating its parameters).
c) Graph the function of best fit with the scatterplot of the data.
d) With the best function found in part (b), predict the average number of insured persons in 1997.

2. The following data was obtained by throwing a rubber ball.

Time (sec)	Height (m)
0.0000	1.03754
0.1080	1.40205
0.2150	1.63806
0.3225	1.77412
0.4300	1.80392
0.5375	1.71522
0.6450	1.50942
0.7525	1.21410
0.8600	0.83173

a) Fit linear, quadratic, cubic, and power functions to the data. By comparing the values of R^2, determine the function that best fits the data.
b) In all your answers in each model, write down the equations of your solutions (after calculating its parameters).
c) Graph the function of best fit with the scatterplot of the data.
d) Determine the maximum height of the ball (in meters).
e) With the model you selected in part (b), predict when the height of the ball is *at least* 1.5 meters.

3. Develop a model for estimating heating oil used for a single-family home in the month of January based on average temperature and amount of insulation in inches.

Oil	Temp F	Insulation
275	40	4
360	27	4
160	40	10
40	73	6
90	65	7
230	35	40
370	10	6
300	9	10
230	24	10
120	65	4
30	66	10
200	41	6
440	22	4
323	40	4
50	60	10

a) Fit linear, quadratic functions to the data. By comparing the values of R^2, determine the function that best fits the data.
b) In all your answers in each model, write down the equations of your solutions (after calculating its parameters).
c) Then use the regression models for the functions in b to predict the needed oil if the temperature is 10 Fahrenheit and the insulation is 5 attic insulations inches.
d) What is your recommendation for the company?
e) You may review the data and remove what is outside the reasonable range (outlier), then recalculate the results and compare.

Part 2

Use the ReducedMNIST which is a reduced version of the MNIST data set.

- **ReducedMNIST training**: 1000 examples for each digit.
- **ReducedMNIST test**: 200 examples for each digit.

1. Use the ReducedMNIST data to generate these features for each of the images of the training and testing sets:
   a. DCT features (200 dimensions)
   b. PCA (use several dimensions so that the total variance is at least 95% of the total variance when using all the 784 dimensions).
   c. A feature of your creation
2. Then train these classifiers using the training set of the MNIST data for each of the above features:
   a. K-means for each class. Try 1, 4, 16, and 32 clusters for each class.
   b. SVM. You may try the linear and the nonlinear kernels (like RBF). In this case state clearly, what kernel have you used.

Then use the resulting models to classify the test set. Compare the different features and the different classifiers:
You must add a final table to summarize all your results (accuracy and processing time) in a comparative way, as the table is shown below.

1. **Only for the best result of each classifier put it in a confusion matrix among the 10 digits.**
2. **Add your final conclusions.**

		Features					
		DCT		PCA		Your features	
		Accuracy	Processing Time	Accuracy	Processing Time	Accuracy	Processing Time
Classifier							
K-means Clustering	1						
	4						
	16						
	32						
SVM	Linear						
	nonlinear*						

* Mention the kernel name and its specs