EE 4540 Introduction to Machine Learning – Group assignment – version1

In groups of 1 or 2 people, using data sets from the following websites:

http://people.sc.fsu.edu/~jburkardt/datasets/regression/regression.html

http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/frame.html

- estimate the relevant feature y (typically contained in the last data column) as a function of the other features X, using linear regression based on a MMSE estimator. Compare the de-normalized estimate

  $$\hat{y}_d = \hat{y} \cdot \sigma + m$$

  (where σ is the standard deviation and m is the mean value of the original data y), with the correct value y, and evaluate the accuracy of the estimate (i.e. the normalized mean square error).

  $$\alpha = \frac{E\{(\hat{y}_d - y)^2\}}{E\{(y)^2\}}$$

- Consider different number of testing data (50%, 60%, 70%, 80%, 90%), and compare the accuracy obtained in the various cases. Do you observe an optimal ratio between number of data used for training and for testing?

You must have at least 4 features, and you should only use the meaningful features (not the columns containing only constant values). List the meaning of the relevant feature y and of the features used for the regression.

Optional questions:

1. Is the relationship among y and the other features linear? How can you check it? Should you use a different model?
2. Does your accuracy improve or worsen if you use a numerical algorithm for optimization? Do your performance vary if you initialize gradient and steepest descent algorithm with random values (you can use the command rand())
3. Interpret your results. Which feature is more relevant in your estimate? How can you tell? Does it make sense? Are you discovering something unexpected?
4. Are your features X redundant? How can you tell?

Try to visualize your results in an effective way. Save your work with the "publish" option and write your answers in the Matlab file as comments

EE 4540 Introduction to Machine Learning – Group assignment – version2

In groups of 1 or 2 people, using data sets from the following websites:

http://people.sc.fsu.edu/~jburkardt/datasets/regression/regression.html

http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/frame.html

- estimate the relevant feature y (typically contained in the last data column) as a function of the other features X, using linear regression based on a MMSE estimator. Compare the de-normalized estimate

  $$\hat{y}_d = \hat{y} \cdot \sigma + m$$

  (where σ is the standard deviation and m is the mean value of the original data y), with the correct value y, and evaluate the accuracy of the estimate (i.e. the normalized mean square error).

  $$\alpha = \frac{E\{(\hat{y}_d - y)^2\}}{E\{(y)^2\}}$$

- Use 85% of the data for training, and pick an example with a large number of features.

Questions:

1. Are your features X redundant? How can you tell?
2. Reduce the number of features eliminating those that are less correlated with y and identify how many feature you must eliminate before you observe a rapid increase in α
3. Reduce the number of features by eliminating those that are highly correlated with other features in X and identify how many feature you must eliminate before you observe a rapid increase in α
4. Using the Principal Component Analysis theory, change the basis for your feature space into a basis of uncorrelated random variables, and eliminate those with a lower variance. Identify how many feature you must eliminate before you observe a rapid increase in α
5. Derive your conclusions.

Try to visualize your results in an effective way. Save your work with the "publish" option and write your answers in the Matlab file as comments