Homework Assignment #1  
Due: 02/04/2005 in class  
(totals: 100 points)

Problem 1 (8 points). What are the major tasks of data mining? Explain each in 2-3 sentences.

Problem 2 (8 points). Define Euclidean distance between two $k$-dimensional vectors $a$ and $b$. Also, write a formula for Minkowski distance with $p \to \infty$ and provide intuition for the loss of the summation element as compared to the standard Minkowski distance with non-infinity parameter $p$.

Problem 3 (6 points). Explain the fundamental difference between supervised and unsupervised learning. In general, what is semi-supervised learning?

Problem 4 (8 points). Use Table 1.1 from the textbook (page 5) to answer the following questions: (i) how many features are there in the dataset; (ii) how many data points are in the dataset; (iii) what kind of a feature is Marital Status; (iv) how many numerical features are there in the dataset; (v) how many missing values are there in the whole dataset; (vi) is target numerical or categorical; (vii) how many missing values are there in the dataset; (viii) do you think feature ID can be useful variable (feature) for predicting Income. Explain your answer.

Problem 5 (6 points). Explain sample vs. population. What are estimates and what are parameters?

Problem 6 (7 points). Give an example of a biased sample which would not be a good representative of the population and which would potentially lead to incorrect inference. Please avoid examples from Chapters 1-3 of the textbook.

Problem 7 (7 points). Name five properties according to which we will compare data mining techniques.

-----

Programming assignments:

Problem 7 (25 points). From the UCI Machine Learning Repository download dataset Housing (ftp://ftp.ics.uci.edu/pub/machine-learning-databases/housing/). The dataset has 506 data points (instances, examples...), 13 features (attributes) and one target (MEDV) which is a median value of owner occupied homes in $1000. Please provide answers to the following questions:

(i) (4 points) What is your intuition about the relationship between feature 6 and the target? Explain your answer.

(ii) (4 points) What relationship would you expect to see between feature 1 and the target? Explain your answer.

(iii) (4 points) Calculate correlation coefficient between features 1 and 6. Also, calculate correlation coefficients between each feature and the target.
(iv) (8 points) Print a scatterplot matrix for the whole dataset. What features seem to be most correlated among themselves and most correlated with the target? Explain relationships for two pairs of most correlated features and relationships for two pairs or least correlated features. Was your intuition under (i) and (ii) in accordance with the data.

(v) (5 points) Print a box plot for the first three features. Provide detailed explanation of all elements of the box plot.

**Problem 8** (25 points). Build a function KNN.m whose first line should look like

```matlab
function [prediction, indices] = KNN (D, x, K)
```

where \( D \) is a \( n \)-by-\((k + 1) \) dataset/matrix (in which column \( k + 1 \) is the target), \( x \) is the 1-by-\( k \) vector and \( K \) is an integer. Variable \( \text{prediction} \) should the predicted target of \( x \) generated by the K-NN classifier (it is a single number), while \( \text{indices} \) is a \( K \)-by-1 vector of indices of the \( K \) closest examples (from \( D \)) to the input vector \( x \). Use Euclidean distance as a measure of closeness between \( k \)-dimensional vectors. Note that \( n \) and \( k \) are determined by the user of KNN.m via a matrix \( D \) and vector \( x \).

Please email the code to the instructor on or before the due date. The fastest correct Matlab code will be valued 5 extra points.

-----

For homework assignment policies please see the class syllabus located at

http://www.informatics.indiana.edu/predrag/2005springi400/syllabus.htm

-----

Good luck!