# Problem 1 – Describe Statistical Learning

Statistical learning is divided in two main groups: Supervised and Unsupervised learning.

Supervised learning will try to find a prediction function from data. In those data we have one or more input variables, also called *inputs, features, predictors or independent variables* and usually one output variable, called *outcome, response or dependent variables*. The prediction can be whether regression, if the response variable is a quantitative variable or classification, if the response variable is a qualitative (or categorical) variable. Estimating those output variables means making a prediction.

In order to test the model and to know its performance, we usually separate the model in two parts: Training data and test data. Training data will be used to create and train the model. Test data will have the role of new data, not used in the training phase, to test the efficiency of the model. Repeating the train-test operation with different groups of train-test data is called cross-validation.

Once we have a prediction function, we might want to understand the relations between the variables in order to find unknown relations in the data. This is inference. Often, if a prediction model is very accurate, it will be harder to interpret.

To estimate the prediction function, we can use parametric or nonparametric methods. With the parametric method, we take a classical functional form (like linear, quadratic…) and fit the function my adjusting the function’s parameters. For the nonparametric method, we have more freedom in choosing the form but we have to choose a lot of parameters that requires a lot of observation and there is a risk of overfitting (modeling the noise).

Unsupervised learning is used when there is no output data. The goal is to find different groups (or clusters) in the data by finding relationships between the variables or the observations.

# Problem 2

<http://www.inf.ed.ac.uk/teaching/courses/mlsc/Notes/Lecture4/BiasVariance.pdf>

# Problem 3

# Problem 4 – Questions answering

## (b)

**Question:** Identify the column names corresponding to each of the data types mentioned in the introduction. How many observations do you have (in total, in the training set, in the test set)?

**Answer:**

|  |  |
| --- | --- |
| Ozone | Numerical (integer) |
| Radiation | Numerical (integer) |
| Temperature | Numerical (integer) |
| Wind | Numerical (integer) |

Test Set: 31 observations

Train Set: 80 observations

Total: 111 observations

## (c)

**Question:** What is the range of each input variable? What is the mean and standard deviation of each variable?

**Answer:**  This answer is given with the *apply* function. For the mean and the range, it is possible to use the *summary* function that also give the median, the first and the third quartiles.

> apply(ozone, 2, mean)

ozone radiation temperature wind

42.099099 184.801802 77.792793 9.938739

> apply(ozone, 2, range)

ozone radiation temperature wind

1 7 57 2.3

168 334 97 20.7

> apply(ozone, 2, sd)

ozone radiation temperature wind

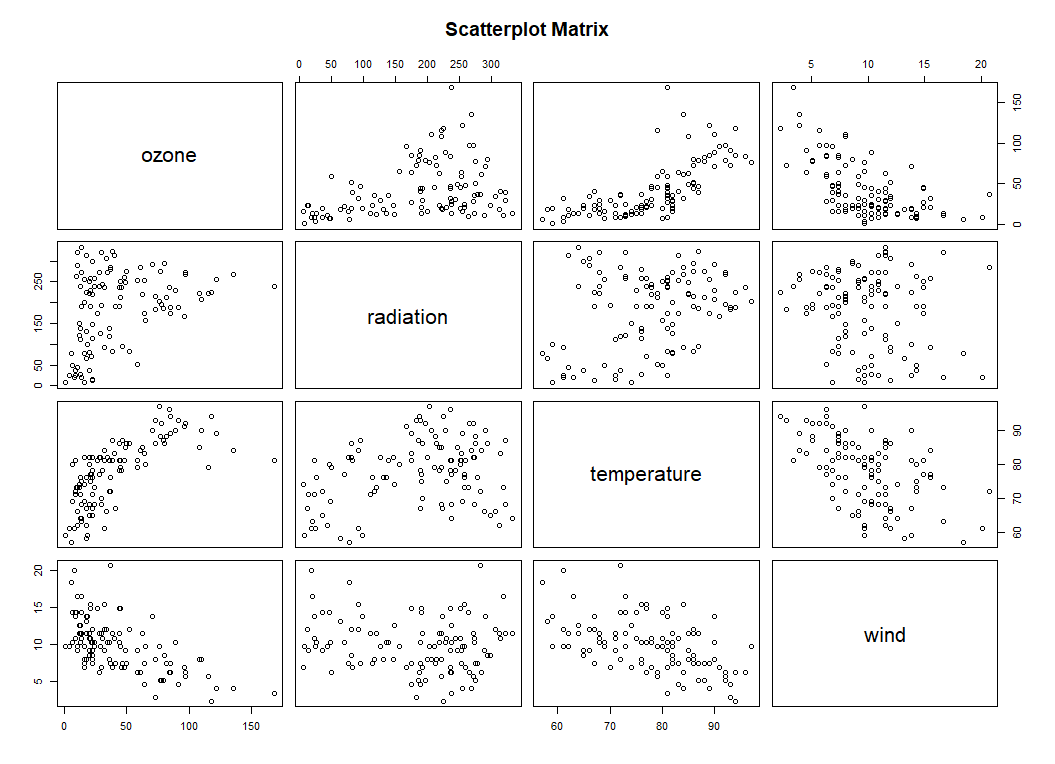
33.275969 91.152302 9.529969 3.559218

## (d)

**Question:**  Create scatterplots for each pair of feature.

**Code:** pairs(~ozone+radiation+temperature+wind,data = ozone, main = "Scatterplot Matrix")

**Result:**



**Question:** Calculate the Pearson correlation coefficients for each pair of datatypes

**Code:**

df = data.frame(ozone)

cor(df)

**Result:**

ozone radiation temperature wind

ozone 1.0000000 0.3483417 0.6985414 -0.6129508

radiation 0.3483417 1.0000000 0.2940876 -0.1273656

temperature 0.6985414 0.2940876 1.0000000 -0.4971459

wind -0.6129508 -0.1273656 -0.4971459 1.0000000

**Question:** In general, what is the range of the Pearson correlation coefficient? What does a correlation coefficient of 0 tell you about the relationship between two variables?

**Answer:** The Pearson correlation coefficient goes from -1 (totally uncorrelated) to 1 (totally correlated). The *0*-correlation coefficient means that there is no correlation between the two features. On the table above, one can see that each variable is totally corelated with itself (makes sense).

**Question:** What trends do you observe in the data according to the correlation coefficient? Can you see them directly from the plot (visually)?

**Answer:**

On the graph we can observe a correlation between *ozone* and *temperature* and also an inverted correlation between *ozone* and *wind*. On the other hand, we see no important correlation between *radiation* and *temperature* and between *ozone* and *radiation*.

The correlation matrix shows that the coefficient of correlation between *ozone* and *temperature* is 0.69 and between *ozone* and *wind* is -0.61. We can observe that the points on the graph seems to follow a curve. For the temperature, the more it is high, the more ozone is high and for wind, the more wind is high, the lower is ozone.