Tutorial: Data Modeling

1. There are several proposals about what the intelligent data analysis process should look like. CRoss Industry Standard Process for Data mining (CRISP-DM) is one of them. Describe the components of the process and the relationships among them. Also discuss questions associated with each of the components.
2. There are mainly four method categories for data analysis problems: classification, regression, clustering and association. Explain and discuss each of them. What are differences among them? Examples are expected to support your discussion.
3. Most of data analysis methods share the same procedure.
    a. Select the Model Class
    b. Select the Score Function
    c. Apply the algorithm
    d. Validate the Results

    Explain and discuss the purpose and main issues of the four steps (what are they? What are the purpose and main issues).
4. Simplicity of the model is one of the important aspects. What are the reasons?
5. A perfect fit with zero error is suspicious in most of cases. Once we have fitted a model to given data, the fitting error can be composed into 4 components:
    a. the experimental error;
    b. the sample error;
    c. the model error;
    d. the algorithm error.

    Explain and discuss each of them.
6. Describe the cross-validation method. Discuss in which situation it should be employed.