Ronald Adomako

HW 1

IST_5200

Data Science and Machine Learning with Python

Professor Langtao Chen

Book 1 Chapter 2:

Problems 2.1, 2.2, 2.5, 2.6, 2.7, and 2.10. Each problem deserves 5 points.

2.1

Assuming that data mining techniques are to be used in the following cases, identify whether the task required is supervised or unsupervised learning.

a. SUPERVISED

Deciding whether to issue a loan to an applicant based on demographic and financial data (with reference to a database of similar data on prior customers).

b. UNSUPERVISED

In an online bookstore, making recommendations to customers concerning additional items to buy based on the buying patterns in prior transactions.

c. SUPERVISED

Identifying a network data packet as dangerous (virus, hacker attack) based on comparison to other packets whose threat status is known.

d. UNSUPERVISED

Identifying segments of similar customers.

e. SUPERVISED

Predicting whether a company will go bankrupt based on comparing its financial data to those of similar bankrupt and nonbankrupt firms.

f. UNSUPERVISED

Estimating the repair time required for an aircraft based on a trouble ticket.

g. SUPERVISED

Automated sorting of mail by zip code scanning.

h. UNSUPERVISED

Printing of custom discount coupons at the conclusion of a grocery store checkout based on what you just bought and what others have bought previously.

2.2

Describe the difference in roles assumed by the validation partition and the test partition.

• The validation partition is used assess the accuracy of the primary model. The test partition is used for an alternate model otherwise known to be used for cross-validation.

2.5

Using the concept of overfitting, explain why when a model is fit to training data, zero error with those data is not necessarily good.

• Zero error on the training data assumes there is no noise in the data. Therefore, if there is noise in the the validation set, then the model won't be able to predict it leading to an inaccurate model. Chances are the data has noise. Machine learning models account for residuals (the difference from discrete result to true result) to better predict new data leading to an overall more accurate model.

2.6

In fitting a model to classify prospects as purchasers or nonpurchasers, a certain company drew the training data from internal data that include demographic and purchase information. Future data to be classified will be lists purchased from other sources, with demographic (but not purchase) data included. It was found that “refund issued” was a useful predictor in the training data. Why is this not an appropriate variable to include in the model?

• Because refunds would not accompany the new data, it is best to leave it out of the model to predict new data although it was relevant to the training partition.

2.7

A dataset has 1000 records and 50 variables with 5% of the values missing, spread randomly throughout the records and variables. An analyst decides to remove records with missing values. About how many records would you expect to be removed?

• Because we remove rows (records) that are missing and not variables (columns) we can expect 5% of the records to be removed. Note that the remaining data will have rows that have some missing data variables, but not entirely empty.

2.10

Two models are applied to a dataset that has been partitioned. Model A is considerably more accurate than model B on the training data, but slightly less accurate than model B on the validation data. Which model are you more likely to consider for final deployment?

- **Model A is overfitting on the training data. Choose Model B for deployment, since it is more accurate on the validation partition compared to the training partition.**