CSCI/DASC 6020: Written Assignment 02

Robert Johnson

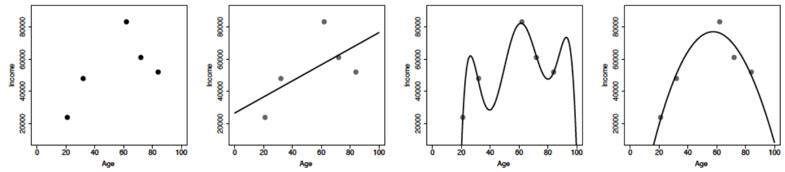
2024-09-24

Assignment Goal

The goal of this assignment is to demonstrate your understanding of fundamentals of machine learning – trade-off between prediction accuracy and model interpretability, supervised versus unsupervised learning, regression versus classification problems, measuring the quality of fit, and the bias-variance trade-off.

1 Question: Simple Regression Models

Consider the following figure:



Shown in the leftmost subfigure is the scatter plot of dataset. *Age* is the predictor variable and *Income* is the response/target variable. The next three subfigures are simple regression models which are referred to as M_1, M_2, and M_3. One of the models is an overfit, another is just right, and the remaining one is underfit. Which model is an overfit model? Underfit model? Just about right model? What is the basis for your answers?

Answer: M_1 Is underfit, you can tell since its just a straight line through the model. M_2 is overfit, the model is doing too much to get the line to every point in the model. M_3 is the right fit for the model above.

2 Question: Consistent Prediction Models

Consider the training data shown below, in which **ID**, **Occupation**, **Age**, and **Loan-Salary Ratio** are the predictor variables, and **Outcome** is the response/target variable.

ID	Occupation	Age	Loan Salary Ratio	Outcome
1	industrial	34	2.96	repaid
2	professional	41	4.64	default
3	professional	36	3.22	default
4	professional	41	3.11	default
5	industrial	48	3.80	default
6	industrial	61	2.52	repaid
7	professional	37	1.50	repaid
8	professional	40	1.93	repaid
9	industrial	33	5.25	default
10	industrial	32	4.15	default

Table 1: A machine learning application dataset.

Next consider the following prediction model (called M_1), which is developed using the data in the table above:

```
if Loan-Salary Ratio > 3 then
    Outcome='default'
else
    Outcome='repay'
end if
```

Why is this model a consistent prediction model? Explain. This model also uses two principles: feature design and feature selection. Explain these two principles.

Answer: This is a consistent prediction model because it accurately and consistently predicts the target variable using input variables. Feature engineering (aka design and selection) is, simply put, the act of converting raw data/observation into useful features. For this example we can see that when Loan Salary Ratio is > 3 Outcome is default, otherwise Outcome is repaid. So we select Loan Salary Ratio and design our model to get the desired feature Outcome.

3 Question: Consistent Prediction Model

Consider the training data shown in the following table. ID, Amount, Salary, Ratio, Age, Occupation, House, and Type are predictor variables, and Outcome is the response/target variable.

ID	Amount	Salary	Loan-Salary Ratio	Age	Occupation	House	Type	Outcome
2	90600	75300	1.20	41	industrial	farm	stb	repaid
3	195600	52100	3.75	37	industrial	farm	ftb	default
4	157800	67600	2.33	44	industrial	apartment	ftb	repaid
5	150800	35800	4.21	39	professional	apartment	stb	default
6	133000	45300	2.94	29	industrial	farm	ftb	default
7	193100	73200	2.64	38	professional	house	ftb	repaid
8	215000	77600	2.77	17	professional	farm	ftb	repaid
9	83000	62500	1.33	30	professional	house	ftb	repaid
10	186100	49200	3.78	30	industrial	house	ftb	default
11	161500	53300	3.03	28	professional	apartment	stb	repaid
12	157400	63900	2.46	30	professional	farm	stb	repaid
13	210000	54200	3.87	43	professional	apartment	ftb	repaid
14	209700	53000	3.96	39	industrial	farm	ftb	default
15	143200	65300	2.19	32	industrial	apartment	ftb	default
16	203000	64400	3.15	44	industrial	farm	ftb	repaid
17	247800	63800	3.88	46	industrial	house	stb	repaid
18	162700	77400	2.10	37	professional	house	ftb	repaid
19	213300	61100	3.49	21	industrial	apartment	ftb	default
20	284100	32300	8.80	51	industrial	farm	ftb	default

Table 2: Another machine learning application dataset.

ID	Amount	Salary	Loan-Salary Ratio	Age	Occupation	House	Type	Outcome
21	154000	48900	3.15	49	professional	house	stb	repaid
22	112800	79700	1.42	41	professional	house	ftb	repaid
23	252000	59700	4.22	27	professional	house	stb	default
24	175200	39900	4.39	37	professional	apartment	stb	default
25	149700	58600	2.55	35	industrial	farm	stb	default

Next consider the following prediction model (called M_2) which is developed using the data in the table above:

```
if Loan-Salary Ratio < 1.5 then
    Outcome='repay'
else if Loan-Salary Ratio > 4 then
    Outcome='default'
else if Age < 40 and Occupation ='industrial' then
    Outcome='default'
else
    Outcome='repay'
end if
```

Is this model a consistent prediction model? Explain. Which model is better? M_1 or M_2. Why?

Answer: Yes this model is a consistent prediction model since it will accurately and consistently determine the output/target variable. This model is likely better than M_1 as it uses more variables and can be used on more cases.

4 Question: Classification or Regression?

Explain whether each scenario is a classification or regression problem.

4.1 Scenario 1

We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

Answer: This is an example of regression. We are trying to evaluate the input variables (record profit, number of employees, industry) and attempting to draw a conclusion on a number value (CEO salary).

4.2 Scenario 2

We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

Answer: This is an example of classification. We are examining the input variables from the previous products and making a prediction on the output value (success or failure) of our own product. So, each previous product is classified already and we want to examine which input variables affect that classification and determine if our input variables will classify us as a success or failure.