Written Assignment 02

AUTHOR
Owen Senowitz

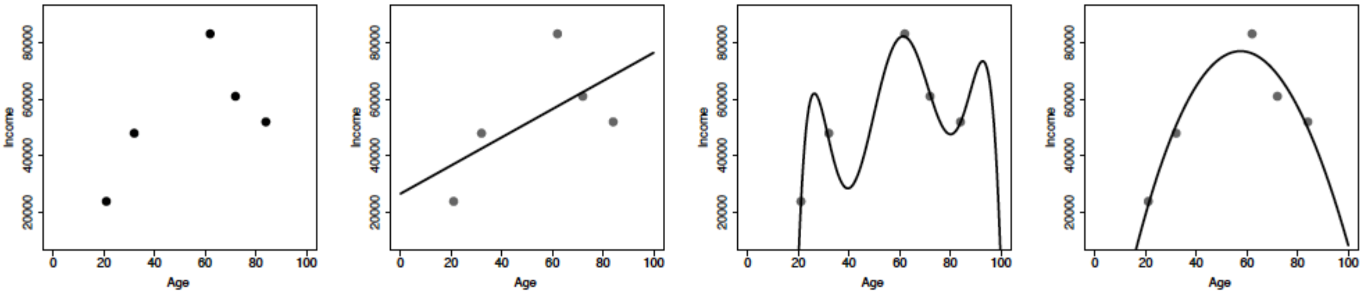
PUBLISHED
September 2, 2024

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, and regression versus classification problems.

1 Question: Simple Regression Models

Consider the following figure:



Shown in the leftmost subfigure is the scatter plot of dataset. is the predictor variable and 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 Under Fit Model: Model M_1 looks like the underfit model. It’s too simple and doesn’t capture the pattern in the data very well. It probably won’t do well on either this data or new data.

M_2 Overfit Model: I think model M_3 is the overfit one because it seems like it’s trying too hard to match every single point in the data. It might do really well on this dataset, but it’ll probably mess up on new data.

M_3 Just Right Model: Model M_2 seems just right. It captures the pattern in the data without being too complicated. It should work well on both this data and new data.

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

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:

Q1) Why is this model a consistent prediction model?

The model is consistent because it correctly predicts the “Outcome” for all the examples in the training data. Here’s how it works. The model says that if the Loan-Salary Ratio is greater than 3, the outcome should be “default.” If the Loan-Salary Ratio is 3 or less, the outcome should be “repaid.”

Now, if you look at the data. For all the entries where the Loan-Salary Ratio is greater than 3, the actual outcome is “default.” For all the entries where the Loan-Salary Ratio is 3 or less, the actual outcome is “repaid.”

Q2) This model also uses two principles: feature design and feature selection.

This is about creating or choosing the right features (variables) that will help the model make accurate predictions. In this case, the model is designed around the “Loan-Salary Ratio” as the main feature. It assumes that this ratio is a strong indicator of whether someone will repay or default on a loan. This is about picking the most important features out of all the available ones. Here, the model is only using the "Loan-Salary Ratio" and ignoring other features like Occupation and Age. The idea is that the Loan-Salary Ratio is the most useful feature for making the prediction.

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.

```
# A tibble: 25 × 9
      ID Amount Salary `Loan-Salary Ratio`   Age Occupation   House Type  Outcome
   <dbl>  <dbl>  <dbl>               <dbl> <dbl> <chr>        <chr> <chr> <chr>
 1     1 245100  66400                3.69    44 industrial   farm  stb   repaid
 2     2  90600  75300                1.2     41 industrial   farm  stb   repaid
 3     3 195600  52100                3.75    37 industrial   farm  ftb   default
 4     4 157800  67600                2.33    44 industrial   apar… ftb   repaid
 5     5 150800  35800                4.21    39 profession… apar… stb   default
 6     6 133000  45300                2.94    29 industrial   farm  ftb   default
 7     7 193100  73200                2.64    38 profession… house ftb   repaid
 8     8 215000  77600                2.77    17 profession… farm  ftb   repaid
 9     9  83000  62500                1.33    30 profession… house ftb   repaid
10    10 186100  49200                3.78    30 industrial   house ftb   default
# ℹ 15 more rows
```

ID	Amount	Salary	Loan-Salary Ratio	Age	Occupation	House	Type	Outcome
1	245100	66400	3.69	44	industrial	farm	stb	repaid
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

ID	Amount	Salary	Loan-Salary Ratio	Age	Occupation	House	Type	Outcome
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
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

Another machine learning application dataset.

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:

Model M_1 is very simple it only uses the Loan-Salary Ratio to decide if someone will “default” or repay. It worked perfectly with the first dataset because it only had one clear rule: if the ratio is above 3, “default” otherwise, repay.

Model M_2 is more complex. It uses not just the Loan-Salary Ratio, but also Age and Occupation to make predictions. Because of this, it can handle more varied data, like the second dataset, where there are more factors to consider.

So, Model M_2 is better for this dataset because it can correctly predict the outcomes even when other factors (like Age and Occupation) are involved. It’s more flexible and can handle more complex situations than Model M_1.

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:

In this scenario, the goal is to understand which factors affect CEO salary. Since salary is a numerical value, and you’re trying to predict it based on other factors (profit, number of employees, industry), this is a regression problem. Regression is used when the target variable is continuous and numerical, like predicting a 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:

In this scenario, the goal is to predict whether a new product will be a success or a failure based on data from similar products. Since the outcome (success or failure) is a categorical variable, this is a classification problem. Classification is used when the target variable is categorical, like determining whether something will be a success or failure.