Given a dataset, what would do based on it?

Labelled data (label: gender: 1 or 2)

1. Look at the size of the sample data

Note: the sample is too small

1. Look at the data to pre-process it
   1. Missing value

Methods: delete those rows/columns; impute (mean, median, majority)

1. Once finished the pre-processing (e.g., missing value), we are going to look at the sample data problem
   1. The entire dataset is small (our case here) ⬄ df.shape
   2. Or in the entire dataset, one class / label has very small sample
      1. 100 points: 99 points for gender 2; 1 point for gender 1 ⬄ imbalanced problem

df[‘gender’].value\_counts(…)

if the percentage is around 30% v.s. 70%, it’s good ⬄ no imbalanced problem

if the percentage is around 5% v.s. 95%, imbalanced ⬄

total 27 points, 1 point is for gender 1; 26 points are for gender 2

1. Cut off the large data points (cut off the 95% datapoints to the same 5%)
   1. We will only remain 1 point for gender 1; 1 point for gender 2.

(assume we have total 100,000: 5,000 for gender 1; 95,000 for gender 2)

1. Duplicate the small data points
   1. We will have 26 points for gender 1; 26 points for gender 2
   2. We duplicate 25 points for gender 1

if the accuracy for the train and validation data has huge gaps⬄ overfitting problem exits.

4. start training

1. consider: when we chose an algorithm, would there be over-fitting problems?

a. How can we examine if the over-fitting problem exist or not?

A picture containing icon

Description automatically generated b. If the over-fitting problem exist, how would we resolve it?

a. we will split the data into data\_train, data\_validation

if we have 100 points, 70 points are used to train; 30 points are used to validate

1. use the 70 points to train a model, decision tree

2. use the model from 1 to predict the labels of the train data

Accuracy based on the training data

(expect the high accuracy for this one)

3. use the model in 1 to predict the labels for the validation data

Accuracy based on the validation data

Model



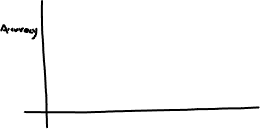
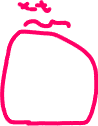
|  |  |  |  |
| --- | --- | --- | --- |
| age | salary | gender | Predicted label |
| 20 | 100 | 1 | 1 |
| 30 | 200 | 1 | 1 |
| 40 | 300 | 2 | 2 |
| 25 | 150 | 2 | 1 |
| 23 | 120 | 1 | 1 |
| 45 | 110 | 2 | 1 |
| 40 | 300 | 2 | 2 |
| 25 | 150 | 2 | 1 |
| 40 | 300 | 2 | 1 |
| 25 | 150 | 2 | 2 |











b. If the over-fitting problem exist, how would we resolve it?

k-folder method; cross c

100 points:

k = 5, 5 folder method

1.we will split the data and put into 5 equal-sized subgroups.

2.We will use the 4 subgroups data to train, use the left 1 to validate

3. we will repeat 1 and 2 for 5 times and each time, it randomly selects the 4 folders



Predicted label g1 predicted label g2

Real label g2

Real label g1

|  |  |
| --- | --- |
| 3 | 0 |
| 2 | 2 |

Confusion matrix. Calculate the accuracy = (3 + 2) / (3+0 + 2 + 2)= 5/7 = 0.71 (71%)

Precision and recall