University of Ottawa

Introduction to Machine Learning CSI5155 2019

Practice Exercises for Midterm 2

Question 1

Consider the following results as obtained when evaluating three algorithms, using 10 fold cross validation against a single data set.

Fold	Decision Tree	Rule learner	Nearest neighbor
1	0.76	0.82	0.91
2	0.79	0.79	0.94
3	0.87	0.92	0.84
4	0.65	0.71	0.63
5	0.56	0.62	0.71
6	0.87	0.92	0.84
7	0.65	0.71	0.63
8	0.85	0.91	0.83
9	0.96	0.82	0.91
10	0.65	0.97	0.74

Show the steps to determine whether there is a statistical significance between the results of the three algorithms, using the paired t-test, with a significance level of $\alpha = 0.05$.

Answer:

The following table shows that distances between the three algorithms.

DT-RL	DT-NN	RL-NN
-0.06	-0.15	-0.09
0	-0.15	-0.15
-0.05	0.03	0.08
-0.06	0.02	0.08
-0.06	-0.15	-0.09
-0.05	0.03	0.08
-0.06	0.02	0.08
-0.06	0.02	0.08
0.14	0.05	-0.09
-0.32	-0.09	0.23

Here are the results:

	DT-RL	DT-NN	RL-NN
t-value	**-1.648863**	**-1.351951**	**0.5584**
p-value	**0.13358**	**0.20938**	**0.59019**

The p-values are higher than 0.05, so we could conclude that are no statistically significant differences. They are quite extreme, and we see that it is not a great idea to assume the normal distribution. So a t-test is not the best way to go and we cannot use the tables. (See below for the Mathematica code.)

```
In[1]:= a = Import["3L         .xlsx"]
Out[1]= {{{-0.06, 0., -0.05, -0.06, -0.06, -0.05, -0.06, -0.06, 0.14, -0.32},
    {-0.15, -0.15, 0.03, 0.02, -0.15, 0.03, 0.02, 0.02, 0.05, -0.09},
    {-0.09, -0.15, 0.08, 0.08, -0.09, 0.08, 0.08, 0.08, -0.09, 0.23}}}

In[2]:= a = a[[1]]
Out[2]= {{-0.06, 0., -0.05, -0.06, -0.06, -0.05, -0.06, -0.06, 0.14, -0.32},
   {-0.15, -0.15, 0.03, 0.02, -0.15, 0.03, 0.02, 0.02, 0.05, -0.09},
   {-0.09, -0.15, 0.08, 0.08, -0.09, 0.08, 0.08, 0.08, -0.09, 0.23}}

In[3]:= Print[p = Map[PairedTTest[#] &, a]];
        Print[mean = Map[Mean[#] &, a]];
        Print[std = Map[StandardDeviation[#] &, a]];
        Print[n = Map[Length[#] &, a]];
        Print[t = mean * Sqrt[n] / std];

PairedTTest::nortst : At least one of the p-values in {0.0267466}, resulting from a test
    for normality, is below 0.05`. The tests in {PairedT} require that the data is normally distributed. >>

PairedTTest::nortst : At least one of the p-values in {0.00888689}, resulting from a test
    for normality, is below 0.05`. The tests in {PairedT} require that the data is normally distributed. >>

PairedTTest::nortst : At least one of the p-values in {0.00165347}, resulting from a test
    for normality, is below 0.05`. The tests in {PairedT} require that the data is normally distributed. >>

General::stop : Further output of PairedTTest::nortst will be suppressed during this calculation. >>

{0.133576, 0.209385, 0.590193}

{-0.058, -0.037, 0.021}

{0.111235, 0.0865448, 0.118926}

{10, 10, 10}

{-1.64886, -1.35195, 0.558397}
```

Question 2

Suppose that we have 10 different data sets, and that we use a decision tree and a rule induction algorithm to construct models against it. Show the results when using the Wilcoxon's signed rank test.

Dataset	DT-RL	Rank
1	0.06	4
2	0.02	1
3	-0.05	3
4	0.09	5
5	0.12	8
6	0.03	2
7	-0.39	9
8	-0.40	10
9	-0.10	6.5
10	0.10	6.5

Answer:

Note that the two results for data sets 9 and 10 with the same values (0.10) are ranked as 6.5.

The critical value is 8.

The sum of positive ranks is 26.5, while the sum of negative ranks is 28.5, so *min(P, N)* $> \alpha$. This indicates that there is no statistical significant difference between the two algorithms.

Question 3

Consider the following results as obtained when evaluating three algorithms, against 10 different data sets. Show the steps to determine whether there is a statistical significance between the results of the three algorithms, using Friedman's test.

Dataset	Decision Tree	Rule learner	Nearest neighbor
1	0.76	0.82	0.91
2	0.79	0.79	0.94
3	0.97	0.92	0.84
4	0.65	0.69	0.68
5	0.56	0.62	0.71
6	0.87	0.91	0.84
7	0.65	0.71	0.63
8	1.05	0.89	0.88
9	0.96	0.82	0.91
10	0.65	0.97	0.79

Answer. This results in the following Ranking.

Dataset	Decision Tree	Rule learner	Nearest neighbor
1	3	2	1
2	2.5	2.5	1
3	1	2	3
4	3	1	2
5	3	2	1
6	2	1	3
7	2	1	3
8	1	2	3
9	1	3	2
10	3	1	2

The resultant rankings are:

	Decision Tree	Rule learner	Nearest neighbor
Sum	21.5	17.5	21
Average	2.15	1.75	2.1

We have 3 algorithms, k = 3 and we have n = 10 datasets. From the χ^2 table, as in the textbook, the critical value is 7.81. If we use the Friedman test, the critical value will be 6.2.

The average rank is 2, the value of equation 2 is 0.95 and the value of equation 3 is equal to 0.975, which leads to a value (eq2/eq3) = 0.974. We cannot reject the NULL hypothesis that all the algorithms perform equally. So, there is again no statistically significant difference between these algorithms.

Question 4

Suppose that we are keeping track of customers' age and their level of fitness (range 1 to 10), as shown below.

Age	Fitness
10	5
14	4
18	6
22	3
26	7
30	10
34	6
14	7
18	6
22	10
26	9
30	4
34	6
38	8
42	10
46	3
50	4
54	5
58	6
62	4
66	7
70	9
74	4
78	1
82	10
86	7
22	9
46	7
50	4

Answers

The data sample is not very skewed, it approximates the normal distribution.

Below also the code for calculating the kurtosis (not asked).

Mean age	**42.14**
Median age	**38**
Mode age	**22**
Skew age	**0.451**

Mean fitness	**6.2414**
Median fitness	**6.0000**
Mode fitness	**4.0000**
Skew fitness	**-0.0015**

```
In[1]:= t = Import["table.xlsx"]
Out[1]= {{{10., 5.}, {14., 4.}, {18., 6.}, {22., 3.}, {26., 7.}, {30., 10.}, {34., 6.}, {14., 7.},
   {18., 6.}, {22., 10.}, {26., 9.}, {30., 4.}, {34., 6.}, {38., 8.}, {42., 10.},
   {46., 3.}, {50., 4.}, {54., 5.}, {58., 6.}, {62., 4.}, {66., 7.}, {70., 9.}, {74., 4.},
   {78., 1.}, {82., 10.}, {86., 7.}, {22., 9.}, {46., 7.}, {50., 4.}}, {{}}, {{}}}

In[2]:= t = t[[1]]
Out[2]= {{10., 5.}, {14., 4.}, {18., 6.}, {22., 3.}, {26., 7.}, {30., 10.}, {34., 6.}, {14., 7.},
  {18., 6.}, {22., 10.}, {26., 9.}, {30., 4.}, {34., 6.}, {38., 8.}, {42., 10.},
  {46., 3.}, {50., 4.}, {54., 5.}, {58., 6.}, {62., 4.}, {66., 7.}, {70., 9.},
  {74., 4.}, {78., 1.}, {82., 10.}, {86., 7.}, {22., 9.}, {46., 7.}, {50., 4.}}

In[3]:= Kurtosis[t]
Out[3]= {2.01839, 2.19106}

In[4]:= CentralMoment[t, 4] / StandardDeviation[t] - 3
Out[4]= {21176.9, 26.8347}

In[5]:= Kurtosis[t[[All, 1]]]
Out[5]= 2.01839
```

Question 5

Suppose that you have the data about the following three customers. Show how you would calculate the distance between these individuals.

Fitness	Income	Age	Gender	Profession
3	210000	35	M	Dentist
7	200000	40	M	Artist
5	130000	21	F	Teacher

Answer:

The first step would be to normalise the numeric data. We can do so by using min-max normalisation, and by setting the range according to the domain. In this case, we may have fitness [1, 10], Income [0, 500000], Age from [16, 80]. For the Gender attribute, we can convert it to [0, 1] or [-1, 1]. The categorical attribute poses a problem. One way is to convert the data into different binary categories, as shown below. Alternatively, we can create "categories" of professions and then assign a person to this category. In the second instance, we are losing some specific information. If this is not an option, assigning a numeric value to all professions and the converting it back to the source may be an option.

Fitness	Income	Age	Gender	Dentist	Artist	Teacher
0.22	0.42	0.30	-1	1	0	0
0.67	0.40	0.38	-1	0	1	0
0.44	0.26	0.08	1	0	0	1

Once we have converted the data, we can simply apply a distance function, such as the Euclidian distance to the data.

For instance, the Euclidian distance between the first and the second row is 1.486**.**