Drawing Conclusions

December 5, 2017

0.0.1 Calculating Errors

Here are two datasets that represent two of the examples you have seen in this lesson.

One dataset is based on the parachute example, and the second is based on the judicial example. Neither of these datasets are based on real people.

Use the questions below to assist in answering the quiz questions at the bottom of this page.

```
In [1]: import numpy as np
        import pandas as pd

        jud_data = pd.read_csv('judicial_dataset_predictions.csv')
        par_data = pd.read_csv('parachute_dataset.csv')
```

```
In [2]: jud_data.head()
```

```
Out[2]:    defendant_id    actual  predicted
        0         22574  innocent   innocent
        1         35637  innocent   innocent
        2         39919  innocent   innocent
        3         29610    guilty     guilty
        4         38273  innocent   innocent
```

```
In [3]: par_data.head()
```

```
Out[3]:    parachute_id actual predicted
        0          3956  opens     opens
        1          2147  opens     opens
        2          2024  opens     opens
        3          8325  opens     opens
        4          6598  opens     opens
```

```
In [35]: par_data.actual.unique()
```

```
Out[35]: array(['opens', 'fails'], dtype=object)
```

```
In [36]: par_data.predicted.unique()
```

```
Out[36]: array(['opens', 'fails'], dtype=object)
```

1. Above, you can see the actual and predicted columns for each of the datasets. Using the **jud_data**, find the proportion of errors for the dataset, and furthermore, the percentage of errors of each type. Use the results to answer the questions in quiz 1 below.

```
In [4]: jud_count = jud_data.shape[0]
        jud_count

Out[4]: 7283

In [5]: jud_error_count = jud_data[jud_data['actual'] != jud_data['predicted']].shape[0]
        jud_error_count

Out[5]: 307

In [6]: print('proportion of errors for the judidical dataset: {}'.format(jud_error_count/jud_co

proportion of errors for the judidical dataset: 0.042152958945489497


In [7]: jud_type1_count = jud_data.query("actual == 'innocent' and predicted == 'guilty'").count
        jud_type1_count

Out[7]: 11

In [8]: jud_type2_count = jud_data.query("actual == 'guilty' and predicted == 'innocent'").count
        jud_type2_count

Out[8]: 296

In [10]: print('proportion of type 1 errors for the judidical dataset: {}'.format(jud_type1_coun
         print('proportion of type 2 errors for the judidical dataset: {}'.format(jud_type2_coun

proportion of type 1 errors for the judidical dataset: 0.001510366607167376
proportion of type 2 errors for the judidical dataset: 0.04064259233832212
```

2. Using the **par_data**, find the proportion of errors for the dataset, and furthermore, the percentage of errors of each type. Use the results to answer the questions in quiz 2 below.

```
In [21]: par_count = par_data.shape[0]
         par_count

Out[21]: 5829

In [22]: par_error_count = par_data[par_data['actual'] != par_data['predicted']].shape[0]
         par_error_count

Out[22]: 233

In [23]: print('proportion of errors for the parachute dataset: {}'.format(par_error_count/par_c
proportion of errors for the parachute dataset: 0.039972551037913875
```

```
In [27]: par_type1_count = par_data.query("actual == 'fails' and predicted == 'opens'").count()[
         par_type1_count
```

```
Out[27]: 1
```

```
In [29]: par_type2_count = par_data.query("actual == 'opens' and predicted == 'fails'").count()[
         par_type2_count
```

```
Out[29]: 232
```

```
In [37]: print('proportion of type 1 errors for the parachute dataset: {}'.format(par_type1_coun
         print('proportion of type 2 errors for the parachute dataset: {}'.format(par_type2_coun
```

```
proportion of type 1 errors for the parachute dataset: 0.00017155601303825698
proportion of type 2 errors for the parachute dataset: 0.03980099502487562
```