Cancer

August 4, 2020

Breast Cancer Data Prediction Using Daimensions

In this notebook, we'll be working with a dataset from the University of California Irvine's Machine Learning Repository. It has nine attribute columns to describe various aspects of cells and one classification column that classifies each cell as benign or malignant cancer. More information about the data can be found at (link).
We have two goals: one is to build a model predicting whether a cell is benign or malignant on future cell data and the other is to use attribute rank to learn about which attributes of the cell are most important for predicting cancer in cells. Daimensions' attribute rank option is useful for a lot of biomedical data like cancer cells because most of the time we are not only looking to predict which cells are cancerous but also what caused the cancer. Attribute rank helps us learn about this aspect of the data by telling us which attributes most closely correlate with a cell's classification. This greatly contributes to our understanding of the data and helps guide us toward probable cause.
Here is a look at our training data and the attributes we're using. For the target column, 2 is benign and 4 is malignant.

```
! head cancer_train.csv
```

```
Clump Thickness,Uniformity of Cell Size,Uniformity of Cell Shape,
   Marginal Adhesion,Single Epithelial Cell Size,Bare Nuclei,Bland
   Chromatin,Normal Nucleoli,Mitoses,Class
5,1,1,1,2,1,3,1,1,2
5,4,4,5,7,10,3,2,1,2
3,1,1,1,2,2,3,1,1,2
6,8,8,1,3,4,3,7,1,2
4,1,1,3,2,1,3,1,1,2
8,10,10,8,7,10,9,7,1,4
1,1,1,1,2,10,3,1,1,2
2,1,2,1,2,1,3,1,1,2
2,1,1,1,2,1,1,1,5,2
```

1. Get Measurements

We always want to measure our data before building our predictor in order to ensure we are building the right model. For more information about how to use Daimensions and why we want to measure our data beforehand, check out the Titanic notebook.

```
! ./btc_linux -measureonly cancer_train.csv

Brainome Daimensions(tm) 0.97 Copyright (c) 2019, 2020 by Brainome,
   Inc. All Rights Reserved.

Data:
Number of instances: 559
Number of attributes: 9
Number of classes: 2
Class balance: 63.15% 36.85%

Learnability:
Best guess accuracy: 63.15%
Capacity progression (# of decision points): [3, 4, 5, 5, 7, 7]
Decision Tree: 26 parameters
Estimated Memory Equivalent Capacity for Neural Networks: 56
   parameters

Risk that model needs to overfit for 100% accuracy...
using Decision Tree: 9.66%
using Neural Networks: 56.00%

Expected Generalization...
using Decision Tree: 20.70 bits/bit
using a Neural Network: 9.98 bits/bit

Recommendations:
Note: Maybe enough data to generalize. [yellow]
Note: Decision Tree clustering may outperform Neural Networks. Try with -f DT.

Time estimate for a Neural Network:
Estimated time to architect: 0d 0h 0m 0s
Estimated time to prime (subject to change after model architecting):
    0d 0h 3m 11s

Time estimate for Decision Tree:
Estimated time to prime a decision tree: a few seconds
```

2. Build the Predictor

Based on our measurements, Daimensions recommends we use a decision tree, which has lower risk of overfit and higher generalization for this dataset. We are also using -rank to prioritize certain attributes from our data, and we'll look at which attributes Daimensions decides are important later.

```
! ./btc_linux -v -v -f DT cancer_train.csv -o cancer_predict.py -e 10
    -rank
```

```
Brainome Daimensions(tm) 0.97 Copyright (c) 2019, 2020 by Brainome,
   Inc. All Rights Reserved.

Attribute Ranking:
Using only the important columns: Uniformity of Cell Size, Bare Nuclei,
Clump Thickness, Marginal Adhesion, Mitoses, Uniformity of Cell Shape

Time estimate for Decision Tree:
Estimated time to prime a decision tree: a few seconds
Note: Machine learner type DT given by user.

Estimated time to train a decision tree: less than a minute

Classifier Type:                      Decision Tree
System Type:                          Binary classifier
Best-guess accuracy:                  63.14%
Model accuracy:                       96.77% (541/559 correct)
Improvement over best guess:          33.63% (of possible 36.86%)
Model capacity (MEC):                 18 bits
Generalization ratio:                 30.05 bits/bit
Model efficiency:                     1.86%/parameter
System behavior
True Negatives:                       60.11% (336/559)
True Positives:                       36.67% (205/559)
False Negatives:                      0.18% (1/559)
False Positives:                      3.04% (17/559)
True Pos. Rate/Sensitivity/Recall:    1.00
True Neg. Rate/Specificity:           0.95
Precision:                            0.92
F-1 Measure:                          0.96
False Negative Rate/Miss Rate:        0.00
Critical Success Index:               0.92
Overfitting:                          No

Output: cancer\_predict.py
READY.
```

3. Validate the Model

Now we can validate our model on a separate set of data that wasn't used for training.

```
! python3 cancer_predict.py -validate cancer_valid.csv

Classifier Type:                      Decision Tree
System Type:                          Binary classifier
Best-guess accuracy:                  75.00%
Model accuracy:                       98.57% (138/140 correct)
Improvement over best guess:          23.57% (of possible 25.0%)
```

```
Model capacity (MEC):                18 bits
Generalization ratio:                7.66 bits/bit
Model efficiency:                    1.30%/parameter
System behavior
True Negatives:                      73.57% (103/140)
True Positives:                      25.00% (35/140)
False Negatives:                     0.00% (0/140)
False Positives:                     1.43% (2/140)
True Pos. Rate/Sensitivity/Recall:   1.00
True Neg. Rate/Specificity:          0.98
Precision:                           0.95
F-1 Measure:                         0.97
False Negative Rate/Miss Rate:       0.00
Critical Success Index:              0.95
```

4. Learn From Attribute Rank

From validating the data, we can see that the predictor has 98.57% accuracy. This is great for making predictions on future data. However, what might be of greater interest is looking at the output from building our predictor, specifically the attributes that Daimensions decided to use. Under the section of output called "Attribute Rank," Daimensions has listed the attributes used: Uniformity_of_Cell_Size, Bare_Nuclei, Clump_Thickness, Marginal_Adhesion, Mitoses, and Uniformity_of_Cell_Shape. This information about what attributes were the best predictors of malignant cancer cells is valuable to scientists looking for the causes of this cancer.

Citation

This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.
Sources:

- Dr. WIlliam H. Wolberg (physician), University of Wisconsin Hospitals, Madison, Wisconsin, USA
- Donor: Olvi Mangasarian (mangasarian@cs.wisc.edu), received by David W. Aha (aha@cs.jhu.edu)
- Date: 15 July 1992