College Privacy Prediction with K-Means Clustering and PCA

Benjamin Ward 2730444

Julio Lemus 2700504

Problem & Data Description

Goal: Use a clustering ML algorithm to classify colleges as either public or private based on the 17 features listed below. We also have the labels for the colleges to use for cross-validation and verify performance.

The data set is composed of 777 observations on the following 19 features:

• Name of the college • Private: labels (Not used in training) • Apps • Accept • Enroll • Top10perc	• Top25perc • F.Undergrad • P.Undergrad • Outstate • Room.Board • Books	• Personal, • PhD, • Terminal, • S.F.Ratio, • perc.alumni, • Expend, • Grad.Rate

Importance and Benefits

School's Viewpoint

Determine competitive acceptance rates, grants, etc for the current market and how your school relates to other public and private schools without manually classifying each one.

Student's Viewpoint

Students can use this to help them identify and connect with schools that align with their preferences and goals and optimize their chances of acceptance.

Code Approach

1. Center the data
2. PCA Dimension Reduction
3. Split the data up
    a. Training set: 85% (660 rows)
    b. Cross Validation: 15% (117 rows)
    c. Testing set: None because it would be the same as cross validation evaluation
4. K-Means Clustering
    a. Assign each point to their closest centroid cluster
    b. Calculate the new cluster by taking the mean of all its associated points
5. Performance Evaluation & Metrics

PCA Dimension Reduction

If we want to capture 99% of the variance, how many principal components do we need?

Principle Component	% Variance	% Cumulative Variance
1. Apps	46.36	46.36
2. Accept	40.72	87.08
3. Enroll	6.73	93.81
4. Top10perc	3.14	96.95
5. Top25perc	1.55	98.51
6. F.Undergrad	0.67	99.18

Clustering

Using the following algorithm below from the main notes and book, we performed K-Means clustering

The k-means clustering algorithm is as follows:

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$ randomly.

2. Repeat until convergence: {

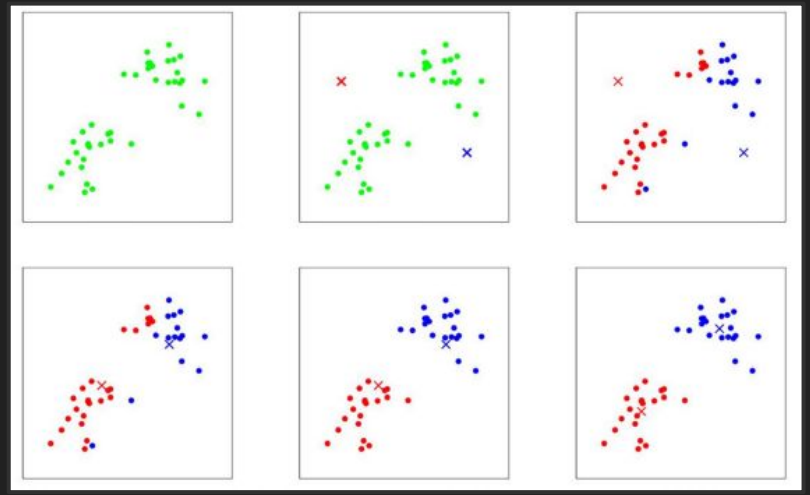
   For every i, set

   $$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

   For each j, set

   $$\mu_j := \frac{\sum_{i=1}^{n} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{n} 1\{c^{(i)} = j\}}.$$

   }



Evaluation Metrics

- True Privates
- False Privates
- True Publics
- False Publics

- Rand Index (true_privates + true_publics) / # of data points evaluated
- Precision true_pos / (true_pos + false_pos)
- Recall true_pos / (true_pos + false_neg)
- F1 2 * precision * recall / (precision + recall)

Results

Best Run With PCA

Cross-Validation (N = 116)	Predicted Private	Predicted Public
Actual Private (73)	True Private = 63	False Public = 12
Actual Public (44)	False Private = 32	True Public = 9

Rand Index: .6410	Precision: .6631	Recall: .8750	F1: .7545

Best Run Without PCA

Cross-Validation (N = 116)	Predicted Private	Predicted Public
Actual Private (73)	True Private = 59	False Public = 13
Actual Public (44)	False Private = 31	True Public = 13

Rand Index: .6153	Precision: .6555	Recall: .8194	F1: .7284

```
       Number elements in Centroid 1: 65
       Number elements in Centroid 2: 595
Iteration 4 Completed

The Centroids Have Converged


----------------------------------------------------------------------
Clustering Cross Validation Results
----------------------------------------------------------------------
       Predicted Private Schools: 96
          Actual Private Schools: 73

       Predicted Public Schools: 20
          Actual Public Schools: 44

  True Private School Predictions: 64
 False Private School Predictions: 32


  True Public School Predictions: 12
 False Public School Predictions: 8



     Rand Index: 0.6495726495726496
      Precision: 0.6666666666666666
         Recall: 0.8888888888888888
      F-1 Score: 0.761904761904762



Centroid 1 Original Location: [[0.10607578 6.50238159 2.25136363 8.83644907 0.9308919  5.76822097]]
Centroid 1 Original Location: [[4.47497353 1.16866717 1.28797684 0.70926511 7.98624502 8.35541873]]
----------------------------------------------------------------------
```

```
       Number elements in Centroid 1: 66
       Number elements in Centroid 2: 594
Iteration 8 Completed

The Centroids Have Converged


----------------------------------------------------------------------
Clustering Cross Validation Results
----------------------------------------------------------------------
       Predicted Private Schools: 95
          Actual Private Schools: 73

       Predicted Public Schools: 21
          Actual Public Schools: 44

  True Private School Predictions: 63
 False Private School Predictions: 32


  True Public School Predictions: 12
 False Public School Predictions: 9



     Rand Index: 0.6410256410256411
      Precision: 0.6631578947368421
         Recall: 0.875
      F-1 Score: 0.7544910179640719



Centroid 1 Original Location: [[1.59585063 0.55355528 9.08638042 9.58338012 4.55848761 4.21124616]]
Centroid 1 Original Location: [[0.61346222 0.06886816 3.42475684 0.48535856 0.79811773 1.28867567]]
----------------------------------------------------------------------
```

```
       Number elements in Centroid 1: 594
       Number elements in Centroid 2: 66
Iteration 9 Completed

The Centroids Have Converged


----------------------------------------------------------------------
Clustering Cross Validation Results
----------------------------------------------------------------------
       Predicted Private Schools: 21
          Actual Private Schools: 73

       Predicted Public Schools: 95
          Actual Public Schools: 44

  True Private School Predictions: 9
 False Private School Predictions: 12


  True Public School Predictions: 32
 False Public School Predictions: 63



    Rand Index: 0.3504273504273504
     Precision: 0.42857142857142855
        Recall: 0.125
     F-1 Score: 0.19354838709677416



Centroid 1 Original Location: [[2.90932749 4.00017956 8.90164515 1.60078967 2.13589884 1.57688945]]
Centroid 1 Original Location: [[5.39669474 7.22571595 6.46163388 1.44269913 4.62758029 4.20917214]]
----------------------------------------------------------------------
```

```
       Number elements in Centroid 1: 594
       Number elements in Centroid 2: 66
Iteration 10 Completed

The Centroids Have Converged


----------------------------------------------------------------------
Clustering Cross Validation Results
----------------------------------------------------------------------
       Predicted Private Schools: 21
          Actual Private Schools: 73

       Predicted Public Schools: 95
          Actual Public Schools: 44

  True Private School Predictions: 9
 False Private School Predictions: 12


  True Public School Predictions: 32
 False Public School Predictions: 63



     Rand Index: 0.3504273504273504
      Precision: 0.42857142857142855
         Recall: 0.125
      F-1 Score: 0.19354838709677416



Centroid 1 Original Location: [[3.97424248 2.48767495 2.02731822 0.62374001 4.55403696 2.32708594]]
Centroid 1 Original Location: [[9.39078973 6.19493426 5.51783808 5.31388662 9.33932271 9.84971286]]
----------------------------------------------------------------------
```

```
       Number elements in Centroid 1: 66
       Number elements in Centroid 2: 594
Iteration 12 Completed

The Centroids Have Converged


----------------------------------------------------------------------
Clustering Cross Validation Results
----------------------------------------------------------------------
       Predicted Private Schools: 95
          Actual Private Schools: 73

       Predicted Public Schools: 21
          Actual Public Schools: 44

  True Private School Predictions: 63
 False Private School Predictions: 32


  True Public School Predictions: 12
 False Public School Predictions: 9



    Rand Index: 0.6410256410256411
     Precision: 0.6631578947368421
        Recall: 0.875
     F-1 Score: 0.7544910179640719



Centroid 1 Original Location: [[7.77622726 9.48486604 5.71802694 3.55362056 7.80615203 3.00318105]]
Centroid 1 Original Location: [[9.06674541 5.38995549 6.48001482 0.1368199  7.84891048 2.44870364]]
----------------------------------------------------------------------
```

```
       Number elements in Centroid 1: 161
       Number elements in Centroid 2: 499
Iteration 13 Completed

The Centroids Have Converged


----------------------------------------------------------------------
Clustering Cross Validation Results
----------------------------------------------------------------------
       Predicted Private Schools: 77
          Actual Private Schools: 73

       Predicted Public Schools: 39
          Actual Public Schools: 44

  True Private School Predictions: 46
 False Private School Predictions: 31


  True Public School Predictions: 13
 False Public School Predictions: 26



    Rand Index: 0.5042735042735043
     Precision: 0.5974025974025974
        Recall: 0.6388888888888888
     F-1 Score: 0.6174496644295301



Centroid 1 Original Location: [[2.5468321  5.14304915 9.39150822 1.21836479 9.0201229  9.01261064]]
Centroid 1 Original Location: [[5.28413201 1.09357533 9.75480062 2.92043048 2.00546331 7.08161736]]
----------------------------------------------------------------------
```

```
       Number elements in Centroid 1: 594
       Number elements in Centroid 2: 66
Iteration 16 Completed

The Centroids Have Converged


======================================================================
Clustering Cross Validation Results
======================================================================
       Predicted Private Schools: 21
          Actual Private Schools: 73

       Predicted Public Schools: 95
          Actual Public Schools: 44

  True Private School Predictions: 9
 False Private School Predictions: 12


  True Public School Predictions: 32
 False Public School Predictions: 63



     Rand Index: 0.3504273504273504
      Precision: 0.42857142857142855
         Recall: 0.125
      F-1 Score: 0.19354838709677416



Centroid 1 Original Location: [[0.06118129 2.46097988 0.86481527 3.92381669 9.42546494 9.40663989]]
Centroid 1 Original Location: [[4.6834837  3.03979491 1.74075125 9.51870872 4.66423848 5.37237542]]
======================================================================
```

Contributions + Conclusion

By categorizing colleges as private or public, universities can better understand their position in the market and make informed decisions on grants, scholarships, and the types of students they wish to attract. This can help optimize their workflows and ensure that they are competing effectively with other institutions.

Contributions: Code and Presentation completed in equal parts by Ben and Julio

Questions?

Benjamin Ward 2730444

Julio Lemus 2700504