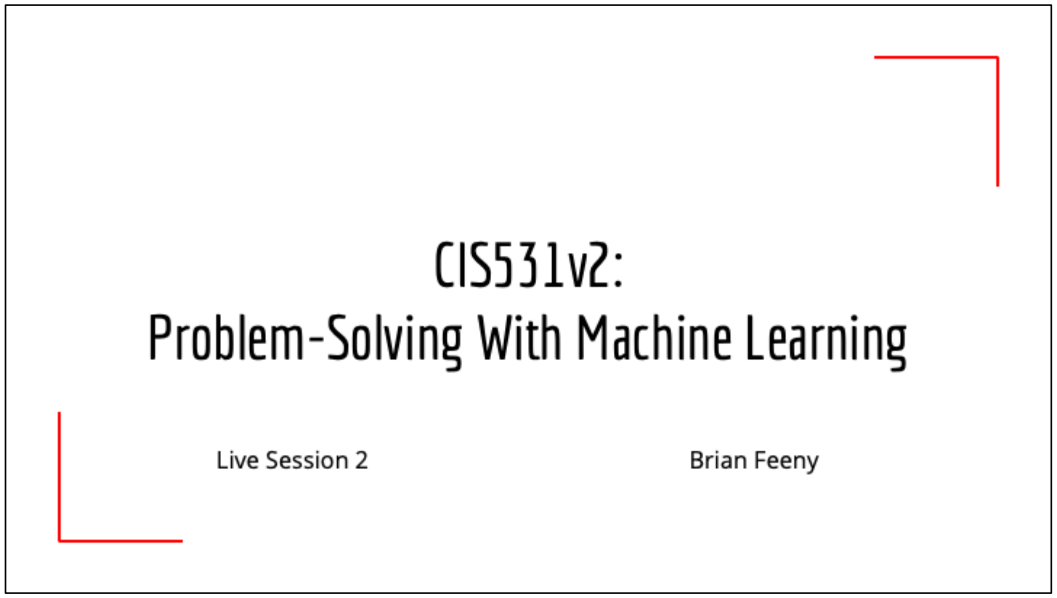


CIS531v2:
Problem-Solving With Machine Learning

Live Session 2

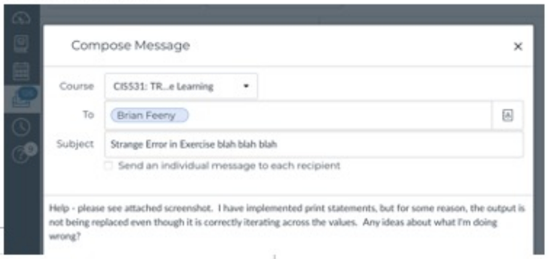
Brian Feeny

Note: This presentation is based substantially on work done by Melynda Eden, eCornell Machine Learning Facilitator, for version 1 of this course.

Brian Feeny, Course Facilitator

How to get assistance:

- Private message me through Canvas



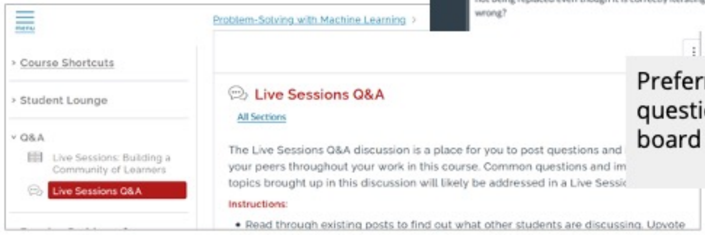
Course: CIS531: TR...e Learning

To: Brian Feeny

Subject: Strange Error in Exercise blah blah blah

Send an individual message to each recipient

Help - please see attached screenshot. I have implemented print statements, but for some reason, the output is not being replaced even though it is correctly iterating across the values. Any ideas about what I'm doing wrong?


Problem-Solving with Machine Learning ›
› Course Shortcuts
› Student Lounge
˅ Q&A
Live Sessions: Building a Community of Learners
Live Sessions Q&A

Live Sessions Q&A

All Sections

The Live Sessions Q&A discussion is a place for you to post questions and
your peers throughout your work in this course. Common questions and im
topics brought up in this discussion will likely be addressed in a Live Sessi

Instructions:

- Read through existing posts to find out what other students are discussing. Upvote

Preferred: You can also post questions to the Q&A discussion board

My name is Brian Feeny, and I am your course facilitator. If you have specific questions for me about the course material, assignments, or questions about project code, send me a message through Canvas. You can post questions or comments about the course material to the Live Sessions Q&A discussion board, but please do not post significant amounts of code, or full or partial solutions.

Previous Recorded Session Topics:

Machine Learning Overview
- Supervised and Unsupervised
- Training, Testing, and Validation Data
- Loss Functions

k-NN overview

NumPy overview

Linear Algebra and Matrix Multiplication

Indexing and Slicing

Additional Resources

Lecture 1 "Supervised Learning Setup" -Cornell CS4780 Machine Learning for Decision Making SP17
https://www.youtube.com/watch?v=MrLPzBxG95I

Lecture 2 "Supervised Learning Setup Continued" -Cornell CS4780 SP17
https://www.youtube.com/watch?v=zj-5nkNKAow

Lecture 3 "k-nearest neighbors" -Cornell CS4780 SP17
https://www.youtube.com/watch?v=oymtGlGdT-k

Lecture 4 "Curse of Dimensionality / Perceptron" -Cornell CS4780 SP17
https://www.youtube.com/watch?v=BbYV8UfMJSA&t=2478s

Today’s Live Session

NumPy and Subsetting vs. Loops

Euclidean Distance Function Without Loops

Mode Function

Build a Facial Recognition System

Walk through of challenging lab survival tips

NumPy and Vectorization vs. Loops

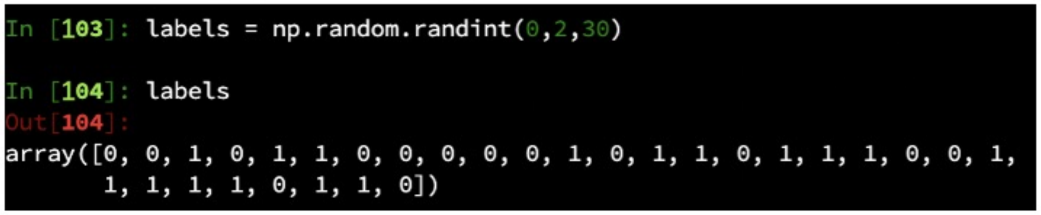
Python is an interpreted language, not compiled

Loop iterations are processed sequentially

NumPy has built-in array functions that first compile the instructions and run them in a lower-level language

A small performance hit for the first iteration due to compile time, but afterwards operations can be implemented in parallel

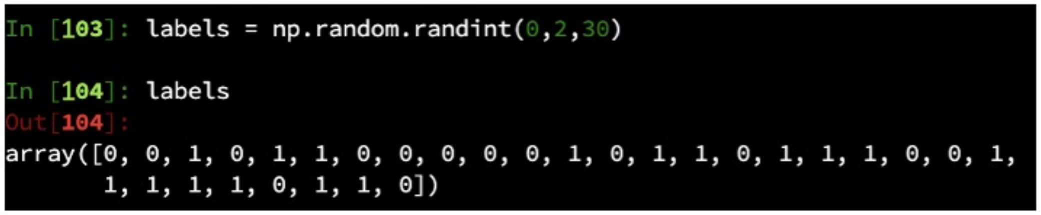
But what if you need to perform the same evaluation or operation across a large number of array values and no built-in NumPy function exists?



```
In [103]: labels = np.random.randint(0,2,30)

In [104]: labels
Out[104]:
array([0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1,
       1, 1, 1, 1, 0, 1, 1, 0])
```

Problem - we want to recode all 0's as -1's

We could write a loop to iterate over the array and check if the number == 0 and set the value to 1

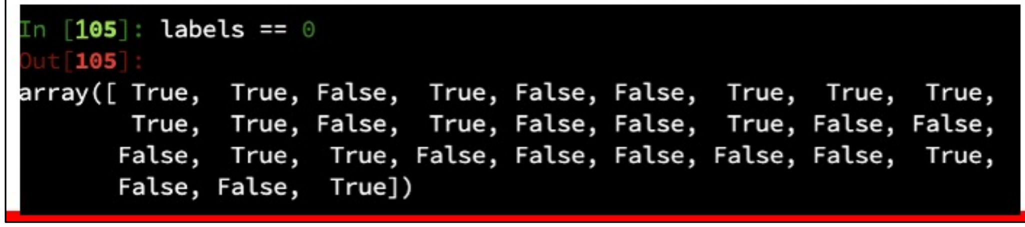


```
In [103]: labels = np.random.randint(0,2,30)

In [104]: labels
Out[104]:
array([0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1,
       1, 1, 1, 1, 0, 1, 1, 0])
```

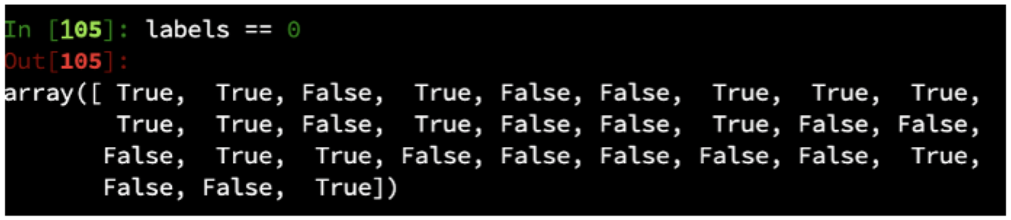
Problem - we want to recode all 0's as -1's

We could write a loop to iterate over the array and check if the number == 0 and set the value to 1

Or we can start by creating an index based on a simple evaluation (>, <, ==, etc.)

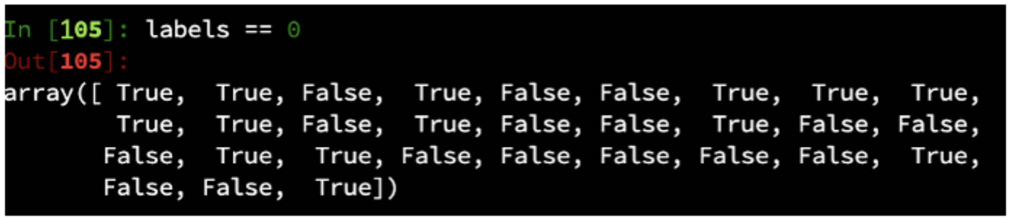


```
In [105]: labels == 0
Out[105]:
array([ True,  True, False,  True, False, False,  True,  True,  True,
        True,  True, False,  True, False, False,  True, False, False,
       False,  True,  True, False, False, False, False, False,  True,
       False, False,  True])
```



```
In [105]: labels == 0
Out[105]:
array([ True,  True, False,  True, False, False,  True,  True,  True,
        True,  True, False,  True, False, False,  True, False, False,
       False,  True,  True, False, False, False, False, False,  True,
       False, False,  True])
```

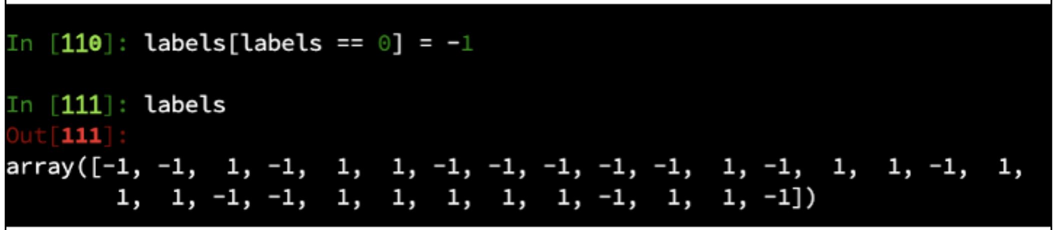
This is an index of all values that meet the criteria specified



```
In [105]: labels == 0
Out[105]:
array([ True,  True, False,  True, False, False,  True,  True,  True,
        True,  True, False,  True, False, False,  True, False, False,
       False,  True,  True, False, False, False, False, False,  True,
       False, False,  True])
```

This is an index of all values that meet the criteria specified

...which can be used to perform replacement operations in parallel



```
In [110]: labels[labels == 0] = -1

In [111]: labels
Out[111]:
array([-1, -1,  1, -1,  1,  1, -1, -1, -1, -1, -1,  1, -1,  1,  1, -1,  1,
        1,  1, -1, -1,  1,  1,  1,  1,  1, -1,  1,  1, -1])
```

How Much Faster Is It Really?

https://jasondeden.medium.com/eat-my-dust-loops-33e5279a01de

Summary:

- Operation on an array that takes > 20 seconds using loops and around 17 seconds using a list comprehension takes under 600 milliseconds using vectorization and around 300 milliseconds using NumPy

- Performance gain of up to 70x for using optimal approach

Use one array to index another one...

https://jasondeden.medium.com/array-indexing-slicing-vs-loops-753484854bc8

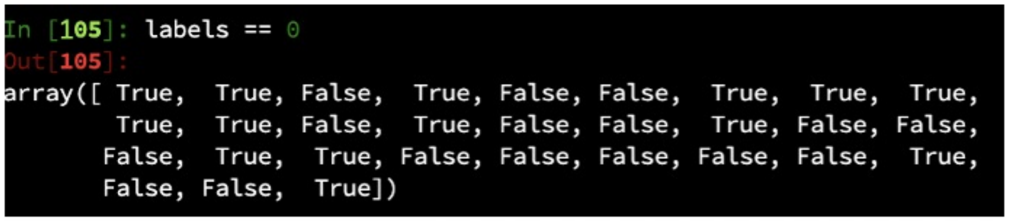
Array Indexing / Slicing vs. Loops

Jason Eden Apr 7 · 3 min read

A Simple, Straightforward Example of a Powerful Concept

Just a quick write-up (you'll be relieved to learn there are no jokes in this one. Except this one.) In the Cornell data science courses I have taken so far, they have heavily emphasized the need to use indexing and slicing

Math Operations on True/False Indexes



```
In [105]: labels == 0
Out[105]:
array([ True,  True, False,  True, False, False,  True,  True,  True,
        True,  True, False,  True, False, False,  True, False, False,
       False,  True,  True, False, False, False, False, False,  True,
       False, False,  True])
```

Question: Is the operation below valid, and if so, what would the response be?

`np.sum(labels == 0)`

Euclidean Distance Without Loops

Compute Euclidean Distances in Matrix Form

- Be sure to read and understand the section in module 4 - **Compute Euclidean Distances in Matrix Form** - as this section explains the logic behind the innerproduct and l2distance functions that you will write
  - training data matrix **X**
  - test data matrix **Z**
  - Gram matrix

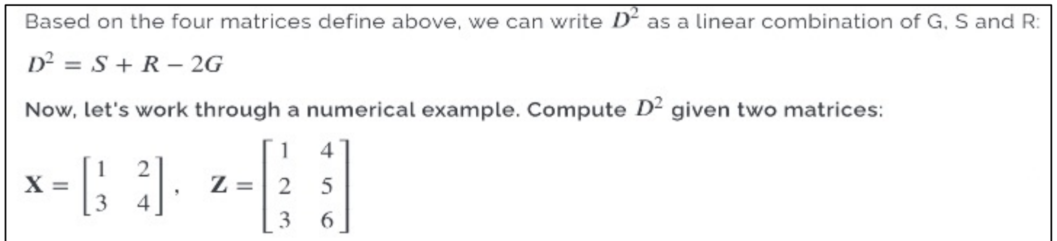
Euclidean Distance Example Walk-Through

Based on the four matrices define above, we can write D^2 as a linear combination of G, S and R:

$$D^2 = S + R - 2G$$

Now, let's work through a numerical example. Compute D^2 given two matrices:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$



Based on the four matrices define above, we can write D^2 as a linear combination of G, S and R:

$$D^2 = S + R - 2G$$

Now, let's work through a numerical example. Compute D^2 given two matrices:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

Squared Norm Review

$||x_1|| = 1^2 + 2^2 = 5$ $\qquad$ $||x_2|| = 3^2 + 4^2 = 25$

S = [5, 25] $\qquad$ Shape = (2,)

$||z_1|| = 1^2 + 4^2 = 17$ $\qquad$ ...

R = [17, 29, 45] $\qquad$ Shape = (3,)

Based on the four matrices define above, we can write D^2 as a linear combination of G, S and R:

$$D^2 = S + R - 2G$$

Now, let's work through a numerical example. Compute D^2 given two matrices:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

Squared Norm Review

$||x_1|| = 1^2 + 2^2 = 5$ $||x_2|| = 3^2 + 4^2 = 25$

S = [5, 25] Shape = (2,)

$||z_1|| = 1^2 + 4^2 = 17$ ...

R = [17, 29, 45] Shape = (3,)

Based on the four matrices define above, we can write D^2 as a linear combination of G, S and R:

$$D^2 = S + R - 2G$$

Now, let's work through a numerical example. Compute D^2 given two matrices:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

$$G = \mathbf{X}\mathbf{Z}^\top = \begin{bmatrix} 9 & 12 & 15 \\ 19 & 26 & 33 \end{bmatrix}$$

G.shape = (2,3)

S.shape = (2,)

R.Shape = (3,)

In order to add matrices, the shapes must match!

How to get there...

S = [5, 25]

Reshape S to get S = [5],
[25]

Repeat 3 times across appropriate axis

R = [17, 29, 45]

Reshape ?

Repeat 2 times across appropriate axis

$$S = \begin{bmatrix} 5 & 5 & 5 \\ 25 & 25 & 25 \end{bmatrix}$$

$$R = \begin{bmatrix} 17 & 29 & 45 \\ 17 & 29 & 45 \end{bmatrix}$$

$$S = \begin{bmatrix} 5 & 5 & 5 \\ 25 & 25 & 25 \end{bmatrix}$$

$$R = \begin{bmatrix} 17 & 29 & 45 \\ 17 & 29 & 45 \end{bmatrix}$$

$$G = \mathbf{X}\mathbf{Z}^\top = \begin{bmatrix} 9 & 12 & 15 \\ 19 & 26 & 33 \end{bmatrix}$$

$$S = \begin{bmatrix} 5 & 5 & 5 \\ 25 & 25 & 25 \end{bmatrix}$$

$$G = \mathbf{X}\mathbf{Z}^\top = \begin{bmatrix} 9 & 12 & 15 \\ 19 & 26 & 33 \end{bmatrix}$$

$$R = \begin{bmatrix} 17 & 29 & 45 \\ 17 & 29 & 45 \end{bmatrix}$$

$$D^2 = S + R - 2G = \begin{bmatrix} 4 & 10 & 20 \\ 4 & 2 & 4 \end{bmatrix}.$$

$$S = \begin{bmatrix} 5 & 5 & 5 \\ 25 & 25 & 25 \end{bmatrix}$$

$$G = \mathbf{X}\mathbf{Z}^\top = \begin{bmatrix} 9 & 12 & 15 \\ 19 & 26 & 33 \end{bmatrix}$$

$$R = \begin{bmatrix} 17 & 29 & 45 \\ 17 & 29 & 45 \end{bmatrix}$$

$D^2{}_1 = S_1 + R_1 - 2G_1$

$$D^2 = S + R - 2G = \begin{bmatrix} 4 & 10 & 20 \\ 4 & 2 & 4 \end{bmatrix}.$$

$$S = \begin{bmatrix} 5 & 5 & 5 \\ 25 & 25 & 25 \end{bmatrix}$$

$$G = \mathbf{X}\mathbf{Z}^\top = \begin{bmatrix} 9 & 12 & 15 \\ 19 & 26 & 33 \end{bmatrix}$$

$$R = \begin{bmatrix} 17 & 29 & 45 \\ 17 & 29 & 45 \end{bmatrix}$$

$D^2{}_1 = S_1 + R_1 - 2G_1$

$S_1 = 5 \quad R_1 = 17 \quad G_1 = 9$

$$D^2 = S + R - 2G = \begin{bmatrix} 4 & 10 & 20 \\ 4 & 2 & 4 \end{bmatrix}.$$

$$S = \begin{bmatrix} 5 & 5 & 5 \\ 25 & 25 & 25 \end{bmatrix}$$

$$G = \mathbf{X}\mathbf{Z}^\top = \begin{bmatrix} 9 & 12 & 15 \\ 19 & 26 & 33 \end{bmatrix}$$

$$R = \begin{bmatrix} 17 & 29 & 45 \\ 17 & 29 & 45 \end{bmatrix}$$

$D^2{}_1 = S_1 + R_1 - 2G_1$

$S_1 = 5 \quad R_1 = 17 \quad G_1 = 9$

$D^2{}_1 = 5 + 17 - 2 * 9 = 22 - 18 = 4$

$$D^2 = S + R - 2G = \begin{bmatrix} 4 & 10 & 20 \\ 4 & 2 & 4 \end{bmatrix}.$$

$$S = \begin{bmatrix} 5 & 5 & 5 \\ 25 & 25 & 25 \end{bmatrix}$$

$$G = \mathbf{XZ}^{\top} = \begin{bmatrix} 9 & 12 & 15 \\ 19 & 26 & 33 \end{bmatrix}$$

$$R = \begin{bmatrix} 17 & 29 & 45 \\ 17 & 29 & 45 \end{bmatrix}$$

$D^2{}_1 = S_1 + R_1 - 2G_1$

$S_1 = 5 \quad R_1 = 17 \quad G_1 = 9$

$D^2{}_1 = 5 + 17 - 2 * 9 = 22 - 18 = 4$

$$D^2 = S + R - 2G = \begin{bmatrix} 4 & 10 & 20 \\ 4 & 2 & 4 \end{bmatrix}.$$

Further Reading

https://www.dabblingbadger.com/blog/2020/2/27/implementing-euclidean-distance-matrix-calculations-from-scratch-in-python

https://medium.com/swlh/euclidean-distance-matrix-4c3e1378d87f

https://numpy.org/doc/stable/user/basics.broadcasting.html

Matrix Diagonals

The course explains that S and R are are square norms and **not** dot products of their respective matrices

```
S = [5, 25]    X @ X.T = [[5,  11],
                          [11, 25]]
```

Matrix Diagonals

The course explains that S and R are are square norms and **not** dot products of their respective matrices.

```
S = [5, 25]    X @ X.T = [[5,  11],
                          [11, 25]]
```

Note, however, that the diagonal of X @ X.T equals [5, 25], which equals S!

Matrix Diagonals

The course explains that S and R are are square norms and **not** dot products of their respective matrices.

```
S = [5, 25]     X @ X.T = [[5,  11],
                           [11, 25]]
```

Note, however, that the diagonal of X @ X.T equals [5, 25], which equals S!

Matrix Diagonals

The course explains that S and R are are square norms and **not** dot products of their respective matrices.

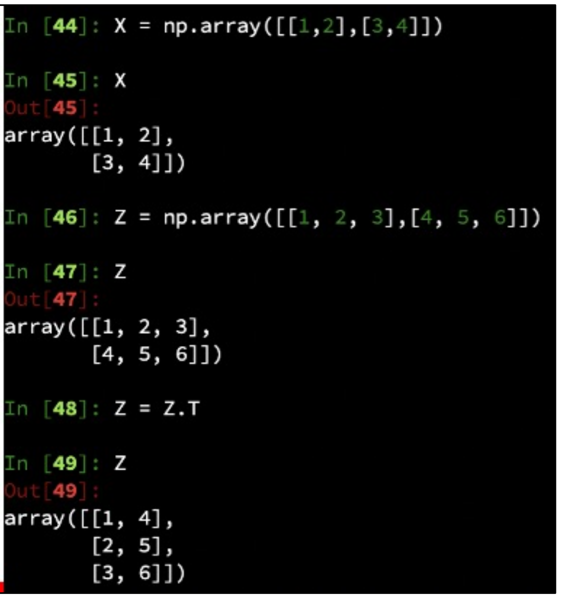
```
S = [5, 25]    X @ X.T = [[5,  11],
                          [11, 25]]
```

Note, however, that the diagonal of X @ X.T equals [5, 25], which equals S!

Does the same hold true for R?

Experiment and See!

Use a Codio terminal lab and try some simple experiments to find ways to simplify your code.



```
In [44]: X = np.array([[1,2],[3,4]])

In [45]: X
Out[45]:
array([[1, 2],
       [3, 4]])

In [46]: Z = np.array([[1, 2, 3],[4, 5, 6]])

In [47]: Z
Out[47]:
array([[1, 2, 3],
       [4, 5, 6]])

In [48]: Z = Z.T

In [49]: Z
Out[49]:
array([[1, 4],
       [2, 5],
       [3, 6]])
```

l2distance( ): Dealing with Negative Values in D^2

The l2distance() function is fairly straightforward. However, in your final D^2 calculation, you may end up with negative values as part of your matrix.

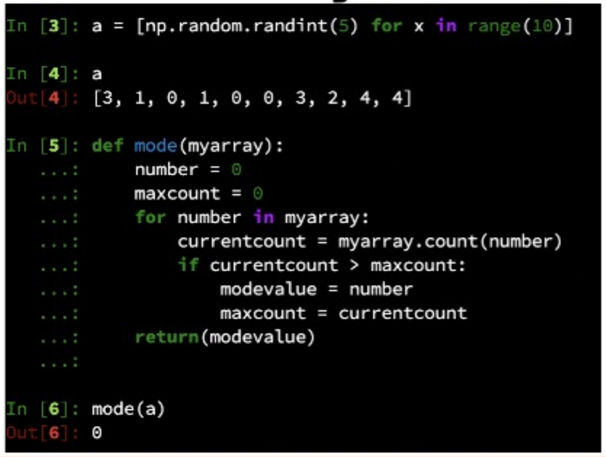
You can't compute the square root of negative values.

When converting D^2 to D, you need to think of a way to convert any values that are "< 0" to 0, without loops.

Hint: You've already seen how to do it in this presentation.

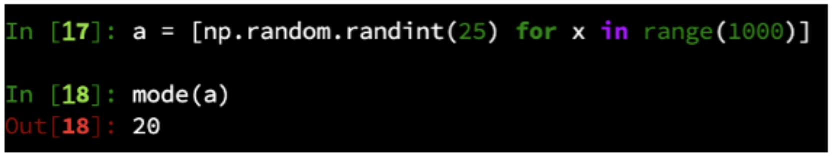
Mode Function Walk-Through

Mode Function Walk-Through



```
In [3]: a = [np.random.randint(5) for x in range(10)]

In [4]: a
Out[4]: [3, 1, 0, 1, 0, 0, 3, 2, 4, 4]

In [5]: def mode(myarray):
   ...:     number = 0
   ...:     maxcount = 0
   ...:     for number in myarray:
   ...:         currentcount = myarray.count(number)
   ...:         if currentcount > maxcount:
   ...:             modevalue = number
   ...:             maxcount = currentcount
   ...:     return(modevalue)
   ...:

In [6]: mode(a)
Out[6]: 0
```

Mode Function In Real Life

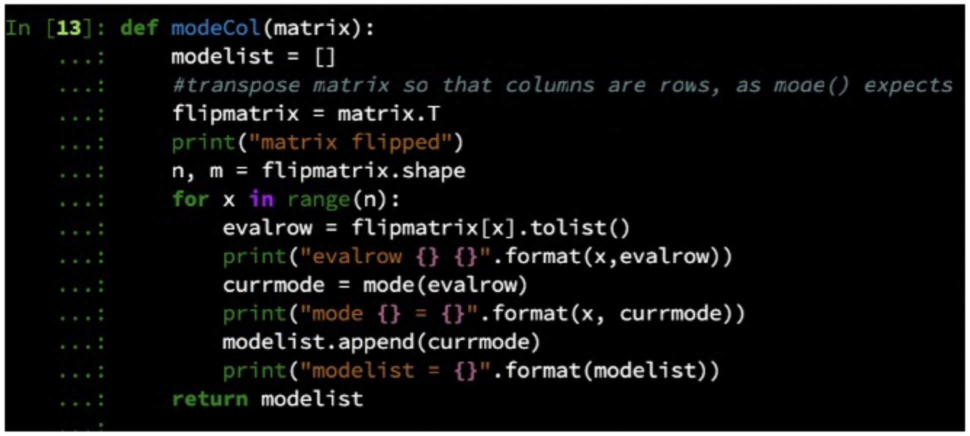


```
In [17]: a = [np.random.randint(25) for x in range(1000)]

In [18]: mode(a)
Out[18]: 20
```

Mode Using `np.unique()`

https://www.delftstack.com/howto/numpy/python-numpy-mode/

Mode over Columns



```
In [13]: def modeCol(matrix):
    ...:     modelist = []
    ...:     #transpose matrix so that columns are rows, as mode() expects
    ...:     flipmatrix = matrix.T
    ...:     print("matrix flipped")
    ...:     n, m = flipmatrix.shape
    ...:     for x in range(n):
    ...:         evalrow = flipmatrix[x].tolist()
    ...:         print("evalrow {} {}".format(x,evalrow))
    ...:         currmode = mode(evalrow)
    ...:         print("mode {} = {}".format(x, currmode))
    ...:         modelist.append(currmode)
    ...:         print("modelist = {}".format(modelist))
    ...:     return modelist
```

Facial Recognition Hints and Tips

Facial Recognition Project

findknn: use l2distance() to calculate the distance between all of your training data points (which are labeled) and test data points (which are the ones you're trying to predict), just like you were looking for the distance between all points in X and Z in the Euclidean Distance project. Then return a subset of the top k number of values and their index positions.

accuracy: evaluate the percentage of the time that values in one vector match another. Don't overthink this one. Remember True / False == 1 / 0.

Facial Recognition Project

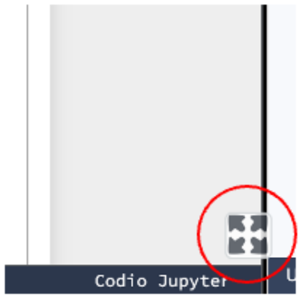
knnclassifier:

- Figure out how to pull only relevant training data labels from yTr
- Mode and Column Mode functions may be useful when k > 1
- The process may need to be different if k==1
    - Why?

Jupyter Notebooks Tips

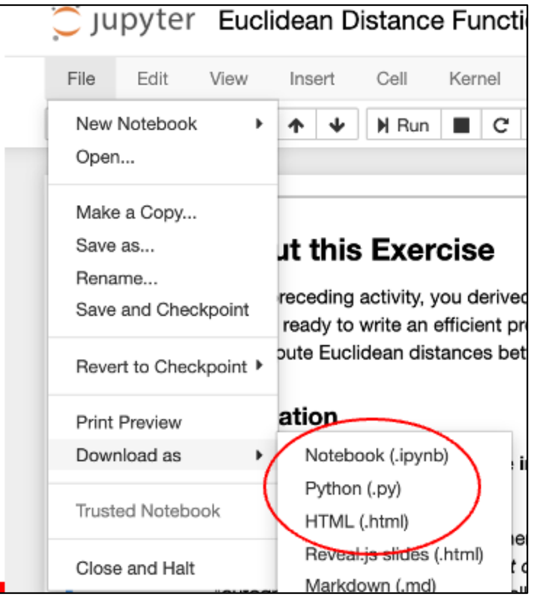
Full Screen Mode

Useful on a small screen. Just remember to save your work every so often and avoid losing to a timeout.



Save Your Work

In the file menu, you can download the notebook as an ipython notebook file, HTML (recommended), or other.



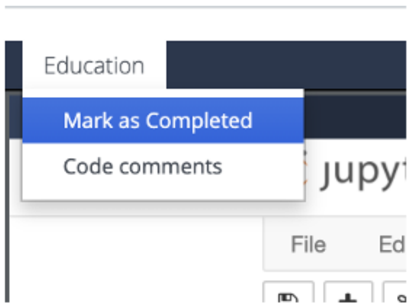
ut this Exercise

receding activity, you derived
ready to write an efficient pr
pute Euclidean distances bet

ation

Mark as Completed

The Education tab where you mark your work as completed and submit for grading is above the notebook menu (File, etc.)



Challenging Lab Approaches

Psuedocode

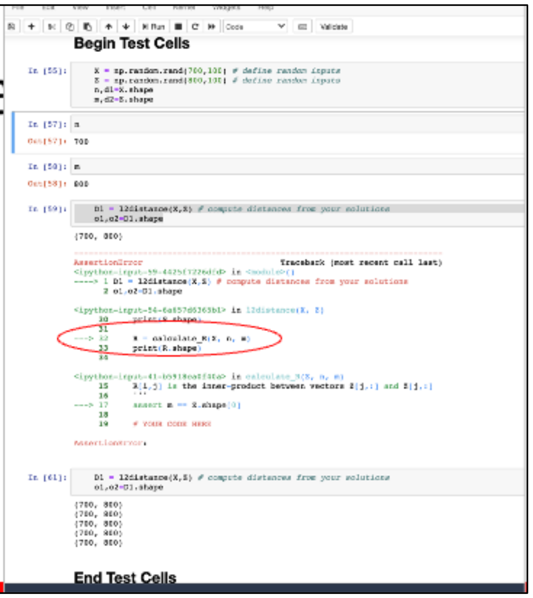
Use comments to write out the code logic you need to figure out

```
#First, run knnclassifier to get k nearest observations for each test point


#Then…
```

Challenging Lab Experimentation

Challenging Lab Experime

- b to add cells (suggest between function and test cells) and label beginning and ending
- Copy test code and run in an experimentation cell, look at outputs.
- Break up their function into data generation and function testing pieces. Evaluate code errors for tests.



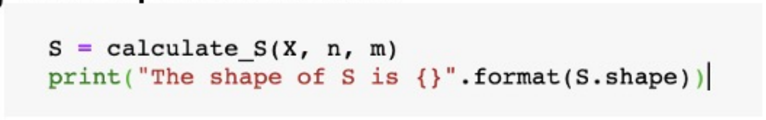
Begin Test Cells

```
In [55]:  X = np.random.rand(700,100) # define random inputs
          Z = np.random.rand(800,100) # define random inputs
          n,d1=X.shape
          m,d2=Z.shape

In [57]: n
Out[57]: 700

In [58]: m
Out[58]: 800

In [59]:  D1 = l2distance(X,Z) # compute distances from your solutions
          o1,o2=D1.shape

(700, 800)

---------------------------------------------------------------------------
AssertionError                            Traceback (most recent call last)
<ipython-input-59-4425f7226dfd> in <module>()
----> 1 D1 = l2distance(X,Z) # compute distances from your solutions
      2 o1,o2=D1.shape

<ipython-input-54-6a857d6363b1> in l2distance(X, Z)
     30     print(X.shape)
     31
---> 32     R = calculate_R(Z, n, m)
     33     print(R.shape)
     34

<ipython-input-41-b5918ea8f4fa> in calculate_R(Z, n, m)
     15     R[i,j] is the inner-product between vectors Z[j,:] and X[j,:]
     16     '''
---> 17     assert m == Z.shape[0]
     18
     19     # YOUR CODE HERE

AssertionError:

In [61]:  D1 = l2distance(X,Z) # compute distances from your solutions
          o1,o2=D1.shape

(700, 800)
(700, 800)
(700, 800)
(700, 800)
(700, 800)
```

End Test Cells

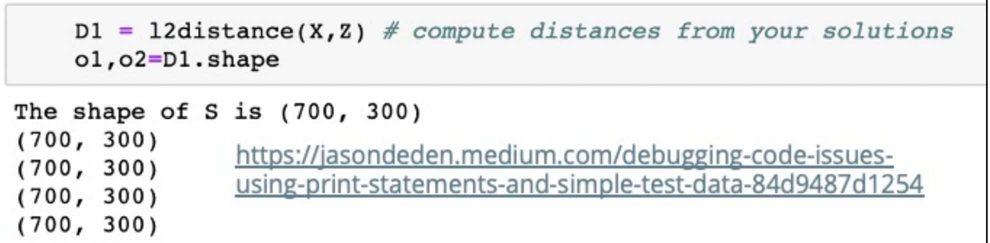
Challenging Lab Experimentation

Use a lot of print statements.



```
S = calculate_S(X, n, m)
print("The shape of S is {}".format(S.shape))
```

Practice them. Get good at them. Comment them out when no longer needed.



```
    D1 = l2distance(X,Z) # compute distances from your solutions
    o1,o2=D1.shape
```

```
The shape of S is (700, 300)
(700, 300)
(700, 300)
(700, 300)
(700, 300)
```

https://jasondeden.medium.com/debugging-code-issues-using-print-statements-and-simple-test-data-84d9487d1254

Grader Functions

Check the output of grader functions, if available, and compare to the output of your own functions.

Reminder: It’s fine to use them in experiments. It is **not** OK to use them in your actual functions. (Yes, we will check for this, and points awarded by the autograder will be removed.)

Questions?

Thank You

End of Live Session 2