edX   **Microsoft:** DAT210x Programming with Python for Data Science

📖 **Bookmarks**

🔖 Bookmark

Welcome to Module 6's SVC Labs!

In order to complete the support vector classifier labs in this module, please make sure you download and unarchive this .zip file with all the datasets and files necessary.

Lab Assignment 1

In this lab, you'll get started with support vector classifier by revisiting UCI's wheat-seeds dataset. First, you will benchmark how long it takes to train and predict with SVC relative to how long K-Neighbors took to train and test, and then you'll compare the decision boundary plot produced by the two.

1. Start by reviewing the starter code located in Module6/**assignment1.py** and of course the wheat dataset in Module6/Datasets/**wheat.data**. Be sure to read through **everything**, so that you know exactly what is being asked of you...

2. Add in the KNN and SVC classifiers

3. Load up your dataset properly and prune it

4. Run the convenience benchmark and plotting functions

5. Then, answer the following questions.

Lab Question 1

(1/1 point)

In this lab, you run KNeighbors and SVC in high dimensional feature space, 5000 times! Ensure you don't have any spyware, movies, games, video editing, or other hardware intensive apps running while you execute the timed runs.

After the high dimensional classification and benchmarking finishes, the lab plots for you a matrix of 2D, feature vs feature plots against one another. This way, you can see how each pair of features stack against the other. Since a square matrix is produced, there are two copies of each plot, although the axes are swapped, and that is controllable by the FAST_DRAW parameter.

Which of the four tests produced the highest scoring result?

- ○ KNeighbors High Dimensionality Score

- ○ KNeighbors "Best" 2D Score

- ○ SVC High Dimensionality Score

- ⦿ SVC "Best" 2D Score ✔

EXPLANATION

SVC in high dimensions, even with a provided kernel, still attempts to find the best linearly separable plane to split your classes. If you have 'dirty' features thrown into the mix, it's entirely possible they will end up hurting your overall SVC performance, as opposed to just having a few really good features.

These are the results you should have received, if you did not alter the parameters before submitting your score:

KNeighbors Results

5000 Iterations Training Time: 1.88873505592

5000 Iterations Scoring Time: 3.78048992157

High-Dimensionality Score: 83.607

Max 2D Score: 90.164

SVC Results

5000 Iterations Training Time: 3.79915714264

5000 Iterations Scoring Time: 1.65462088585

High-Dimensionality Score: 86.885

Max 2D Score: 93.443

You have used 2 of 2 submissions

Lab Question 2

(1/1 point)
Change iterations from 5000 to 200,000 and then re-run your lab with the default parameters.

Ensure you don't have anything else processor intensive running on your machine for the duration of time it takes to do its number crunching (should be 1-10 minutes on modern laptops).

Which of these four computations was *fastest* to calculate?

○   KNeighbors Training Time

○   KNeighbors Testing Time

○   SVC Training Time

◉   SVC Testing Time   ✔

EXPLANATION

SVC Testing is extremely fast. It is essentially just seeing what side of the hyperplane a point is on.

You have used 1 of 2 submissions