EECS 445: Project 1 Quickstart

January 27, 2022

Agenda

1. Project Overview (20 min)
2. Project Setup (10 min)
3. Questions (rest of the time)

Serafina's Social SVMs

Serafina has been tasked with monitoring the EECS 445 piazza and wants to identify "gratitude" or "sadness" in a post

Serafina will test the feasibility of this idea using Reddit comments which have labels for gratitude or sadness!

She enlists a group of EECS 445 students who are versed in solving supervised learning problems to help her!

Project Logistics

Due on Wednesday, 2/9 at 10:00pm

Submit write-up to Gradescope

Submit challenge CSV to Canvas

Coding questions are highlighted in green, questions with written answers are highlighted in blue.

Sections

Section	*Points*	*Recommended Completion Date*
Ethics	9 pts	Friday, 1/28
Feature Extraction	12 pts	Friday, 1/28
Hyperparameter and Model Selection	35 pts	Wednesday, 2/2
Asymmetric Cost Functions and Class Imbalance	20 pts	Friday, 2/4
Challenge	14 pts	Tuesday, 2/8
Code Appendix	10 pts	Tuesday, 2/8

Sections

Section	*Points*	*Recommended Completion Date*
Ethics	9 pts	Friday, 1/28
Feature Extraction	12 pts	Friday, 1/28
Hyperparameter and Model Selection	35 pts	Wednesday, 2/2
Asymmetric Cost Functions and Class Imbalance	20 pts	Friday, 2/4
Challenge	14 pts	Tuesday, 2/8
Code Appendix	10 pts	Tuesday, 2/8

Dataset.csv and debug.csv

4 columns:

- text
- created_utc (timestamps)
- label (-1, 0, 1)
- emotion (matches label)

Heldout.csv does not contain labels

Dataset.csv and debug.csv

4 columns:

- text
- created_utc (timestamps)
- label (-1, 0, 1)
- emotion (matches label)

Heldout.csv does not contain labels

dataset

text	created_utc	label	emotion
Thank you. Yeah it's a nice thought, being sheltered and happy before any worries or drama. It would be nice...	1548126689.0	1	gratitude
Fake lous!	1548123196.0	0	neutral
Thank you kind internet stranger for listening to me. I feel a little less alone for the moment.	1547076010.0	1	gratitude
I meant it's a sign to others that he is a crypto-Nazi. My bad, I wasn't clear.	1548459137.0	0	neutral
) /sorry, the Excel formula part of my brain was driving me insane.	1548348522.0	-1	sadness
True sequels to what?	1547275184.0	0	neutral
So happy to see [NAME] regaining his form... we had to sacrifice [NAME] for it, but i'm okay with that...	1546568427.0	-1	sadness
I think the teacher was right though. Money would make a terrible aggregate.	1548355992.0	-1	sadness

Debug Dataset

DO:

- Use `data/debug.csv` for testing code
- Compare with `debug_output.txt`

Debug Dataset

DO:

- Use `data/debug.csv` for testing code
- Compare with `debug_output.txt`

DON'T:

- Use debug output as an exhaustive test suite
- Use debug output for analysis

Debug Dataset

DO:

- Use `data/debug.csv` for testing code
- Compare with `debug_output.txt`

DON'T:

- Use debug output as an exhaustive test suite
- Use debug output for analysis

To most closely match debug output:

1. Set `random_state=445`
2. Make sure libraries match versions in `requirements.txt`

Debug_output.txt

Run program on debug.csv and compare with debug.txt for sanity checks

Debug_output.txt:

```
Question 3(d): reporting dataset statistics:
The processed sentence is ['best', 'book', 'ever', 'it', 's', 'great']
d: 628
Average number of nonzero features: 11.530303030303031
Most common word: i
-----------------------------------------

Metric: accuracy
Best c: 0.100000
CV Score 0.8715

Metric: f1_score
Best c: 0.100000
CV Score 0.8591
```

Student output:

```
Question 3(d): reporting dataset statistics:
The processed sentence is ['best', 'book', 'ever', 'it', 's', 'great']
d: 829
Average number of nonzero features: 11.530303030303031
Most common word: i
-----------------------------------------

Metric: accuracy
Best c: 0.100000
CV Score 0.8715

Metric: f1_score
Best c: 0.100000
CV Score 0.8591
```

Debug_output.txt

Run program on debug.csv and compare with debug.txt for sanity checks

Debug_output.txt:

```
Question 3(d): reporting dataset statistics:
The processed sentence is ['best', 'book', 'ever', 'it', 's', 'great']
d: 628
Average number of nonzero features: 11.530303030303031
Most common word: i
-----------------------------------------

Metric: accuracy
Best c: 0.100000
CV Score 0.8715

Metric: f1_score
Best c: 0.100000
CV Score 0.8591
```

Student output:

```
Question 3(d): reporting dataset statistics:
The processed sentence is ['best', 'book', 'ever', 'it', 's', 'great']
d: 829
Average number of nonzero features: 11.530303030303031
Most common word: i
-----------------------------------------

Metric: accuracy
Best c: 0.100000
CV Score 0.8715

Metric: f1_score
Best c: 0.100000
CV Score 0.8591
```

Ethics

No code! Just answer the questions.

Feature Extraction

Extract all unique words from the dataset.

Build feature matrix based on whether words are contained in each sentence or not.

Hyperparameter + Model Selection

Learn to use `SVC` and `LinearSVC` classes from `scikit-learn`.

Implement cross-validation for hyperparameter tuning.

Implement hyperparameter search.

Experiment with non-linear classifiers with kernels.

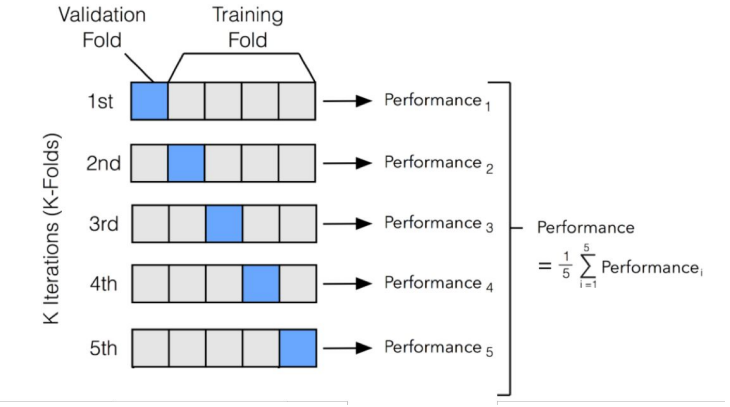
K-fold Cross Validation

Split training data into K folds

Train model K times, using a different fold each time for validation

Average the performance

Repeat for different hyperparameter values and compare to find ideal value



K-fold Cross Validation

Scikit-Learn StratifiedKFold Object

split() function which returns iterable object which can be looped through to give arrays of train and test indices for each fold

sklearn.model_selection.StratifiedKFold

class sklearn.model_selection.StratifiedKFold(*n_splits=5*, *, *shuffle=False*, *random_state=None*) [source]

Stratified K-Folds cross-validator.

Provides train/test indices to split data in train/test sets.

This cross-validation object is a variation of KFold that returns stratified folds. The folds are made by preserving the percentage of samples for each class.

Imbalanced Data

What happens if the dataset does not have 50/50 split between positive and negative labels?

Can we weight data points to adjust for this?

How does class imbalance affect performance metrics?

Challenge

Train a three-class SVM (we now include the neutral class)

Training data:

- `multiclass_features`
- `multiclass_labels`

Run predictions on:

- `heldout_features`

Use `generate_challenge_labels()` to create a CSV. Upload to Canvas as <uniqname>.csv with 1 column of 3000 predictions

Appendices

Review appendices at end of the spec for helpful info

Expected runtimes, topics covered, etc.

Particularly useful for challenge

Project Setup

Questions?