EECS 445: Project 1 Quickstart

September 22, 2022

Agenda

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

Project Overview

Sachchit has spent a lot of time on online boards and forums lately, and he wants to be able to have better conversations. Specifically, he wants to know the sentiment that an online forum has about a certain topic, so that he can respond appropriately.

Goal: Help Sachchit by using support vector machines to predict the sentiment of an online post

Project Logistics

Due on Wednesday, 10/5 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*
Dataset Considerations	11 pts	Friday, 9/23
Feature Extraction	12 pts	Friday, 9/23
Hyperparameter and Model Selection	35 pts	Wednesday, 9/28
Asymmetric Cost Functions and Class Imbalance	20 pts	Friday, 9/30
Challenge	14 pts	Tuesday, 10/4
Code Appendix	8 pts	Tuesday, 10/4

Sections

Section	*Points*	*Recommended Completion Date*
Dataset Considerations	11 pts	Friday, 9/23
Feature Extraction	12 pts	Friday, 9/23
Hyperparameter and Model Selection	35 pts	Wednesday, 9/28
Asymmetric Cost Functions and Class Imbalance	20 pts	Friday, 9/30
Challenge	14 pts	Tuesday, 10/4
Code Appendix	8 pts	Tuesday, 10/4

Dataset Considerations

A discussion on the dataset which we will use to perform this classification task. You will answer questions on ethical concerns and the presence of noise in labels.

No code! Just answer the questions.

Feature Extraction

Write code to generate feature vectors for each example in dataset:

- Extract all unique words from the dataset.
- Build feature matrix based on whether words are contained in each sentence or not.

Hyperparameter + Model Selection

Determine best hyperparameter values for this classification problem:

- Learn to use `SVC` and `LinearSVC` classes from `scikit-learn`.
- Implement cross-validation for hyperparameter tuning.
- Implement hyperparameter search.

Analyze performance with different models:

- Experiment with non-linear classifiers with kernels.

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

Original dataset has three classes:

- Gratitude (positive emotion)
- Sadness (negative emotion)
- Neutral (no discernible emotion)

Goal: Train the most accurate classifier to distinguish between these 3 classes

Key: You have the opportunity to research and experiment with different methods!

Challenge

Training data:

- `multiclass_features`
- `multiclass_labels`

Run predictions on:

- `heldout_features`

Use `generate_challenge_labels()` to create a CSV.
Upload to Canvas!

Challenge

Grading: 14 points in total

- 8 points - Write-Up - Discuss your approach to the challenge, highlighting the choices and your reasoning
- 6 points - Accuracy - We will evaluate your classifier on the accuracy of your predictions on `heldout_features`
    - Your score is will be assigned relative to the rest of the class

Tips - Debug Dataset

DO:

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

Tips - 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

Tips - 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`

Tips - Performance Metrics

Values that characterize the performance of your classification model

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

		Actual Label	
		Positive	Negative
Predicted label	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Project Setup - Demo

Questions?