CS/SE 4AL3 Homework 2: Classification with SVMs

Fall 2025
Due: October 7th

1 Overview

This assignment will get you started working with support vector machine classifiers and evaluation for classification problems. You will build a model to predict solar flare events. You will experiment with real solar activity data through a combination of four feature sets and evaluate with cross validation and confusion matrices.

2 Instructions

This assignment has one part and will be graded out of 25 points. You must work on this assignment individually.

You may use ONLY the following libraries in your code. You will receive a zero if any other libraries are used besides the ones listed below. Note that scikit learn is available for this assignment. You may add import statements within these packages for your convenience, e.g. `from sklearn.PACKAGE import CLASS`.

Listing 1: Available Packages

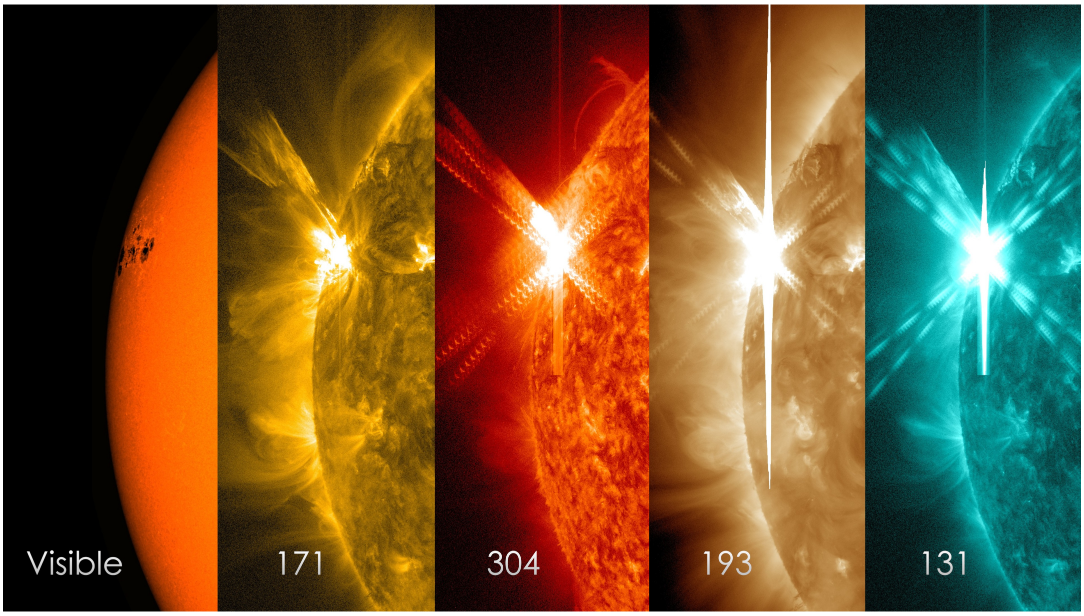
```
import os, sys
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn
import sklearn
import random
```

Your code should be written in Python and should include enough documentation/instructions for someone else to be able to run. You may submit your code as a Jupyter notebook or a Python file. Make sure that your file read/write operations use UTF-8 (if relevant). Your code should run with Python version 3.12. You should fill in the sections of the code that say “#CODE HERE”. The starter files for the notebook and python file are identical.

You are welcome to use code snippets from examples in class, things you find online, or from AI code generation tools. Just make sure to give proper attribution to code you did not write. Follow the syllabus instructions for how to report the use of AI tools. However, you may not copy code that does the entire assignment (e.g. someone who did this assignment in a previous semester).

3 Problem

Context: The sun’s solar flares are large bursts of electromagnetic energy that can significantly impact life on Earth. Powerful solar flares can knock out power grids, satellite systems, and all communication systems on the planet. To monitor solar activity and study such events, the United States National Oceanic and Atmospheric Administration (NOAA) operates the Geostationary Operational Environmental Satellite System; a series of geosynchronous satellites with specialized measurement instruments onboard. The GOES system provides us with solar imagery, magnetometer data, solar X-ray data, and data on high-energy solar protons that hit Earth. This data is sent to Earth and regularly updated on the GOES server which is open for public use.



Unlike Earth, the sun's regions are not divided by countries, states, or cities. Instead, various patches on the sun are numbered by NOAA for scientific investigation purposes, and the most active patches are frequently monitored for high-intensity bursts of electromagnetic radiation. Each patch is called a HARP (HMI Active Region Patch); an enduring, coherent magnetic structure that produces an electromagnetic field. The regions provide measurable features that characterize that patch. There are two classes of solar flare events of particular interest to scientists: the M-class and X-class. Solar flares occur in HARP regions and the level of the energy burst is measured.

Monitoring these patches is helpful, but it is more useful to predict when a powerful burst will occur. Predicting an upcoming solar flare 24 hours in advance can give us time to mitigate damage. This is very challenging because solar flares are rare events. Most importantly, we do not know which features directly indicate an upcoming solar flare.

Challenge: In this assignment, you will build a classfier for binary classification using data from the Helioseismic and Magnetic Imager Instrument on NASA's Solar Dynamics Observatory (SDO) satellite that captures various solar events. It is the first instrument that continuously maps the vector magnetic field of the sun. The magnetic activity recorded using these instruments will serve as the feature set for the model you build.

A major solar event is defined as a burst of GOES X-ray flux of peak magnitude above the M1.0 level. A *positive solar event* is defined as an active HARP region that flares with a peak magnitude above the M1.0 level, as reported in the GOES database. A *negative solar event* is an active region that does not have such an event (no flare above M1.0 level) within 24 hours. The goal of the classification model should be to train on the given data and predict whether a major solar event will occur in the next 24 hours. This means that the classifier must predict whether a given solar event is positive (indicating a flaring active region) or negative (indicating a non-flaring region).

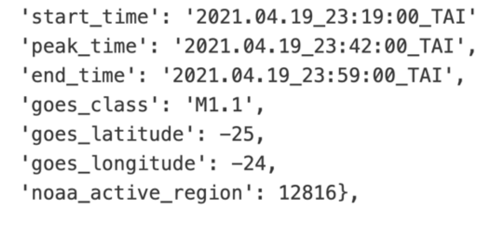
For this assignment, you will implement a Support Vector Machine (SVM) classifier that can distinguish between a positive and a negative solar flar event. Scientists M.G. Bobra and S. Couvidat used the SVM to study solar flares and published their findings in 2015. You may be interested to read the paper, however, it is optional and all information you need for the assignment has already been provided. You will evaluate your classifier using k-fold cross-validation and two different test sets. You will not mix data across these two datasets.

Feature Name	Description	Column Number and Filename
HARPNUM	HARP Number of the sun patch	Column 1-2: all_harps_with_noaa_ars.txt
NOAA_ARS	Corresponding NOAA assignments	
USFLUX	Total unsigned flux	Column 1-18: Main feature set (FS -I) pos_features_main_timechange.npy neg_features_main_timechange.npy
MEANGAM	Mean angle of field from radial	
MEANGBT	Mean gradient of total field	
MEANGBZ	Mean gradient of vertical field	
MEANGBH	Mean gradient of the horizontal field	
MEANJZD	Mean vertical current density	
TOTUSJZ	Total unsigned vertical current	
MEANALP	Mean characteristic twist parameter	
MEANJZH	Mean current helicity	
TOTUSJH	Total unsigned vertical current	
ABSNJZH	Absolute value of the net current helicity	
SAVNCPP	Sum of modulus of the net current per polarity	
MEANPOT	Mean photospheric magnetic free energy	
TOTPOT	Total photospheric magnetic free energy density	
MEANSHR	Mean shear angle	
SHRGT45	Fraction of area with shear >45 degrees	
R_VALUE	Sum of flux near polarity inversion line	
AREA_ACR	Area of strong field pixels in the active region	
Time-Change Features	Changes in the values of the above 18 properties every 6 hours for 24 hours prior to the time we make a prediction (4 times)	Column 19-90: feature set (FS -II) pos_features_main_timechange.npy neg_features_main_timechange.npy
Historical Activity Features	Feature that characterizes the activity history of each active region calculated by adding the scores of the M1.0 class events in the 24 hours prior to prediction time	Column 1: Historical Activity (FS -III) pos_features_historical.npy neg_features_historical.npy
MaxMin Features	The difference between maximum and minimum values in the 24 hours prior to prediction time for each feature	Column 1-18: Max Min Feature (FS -IV) pos_features_maxmin.npy neg_features_maxmin.npy

Dataset: The dataset is gathered from the GOES data server using SunPy, an open-source software for solar physics. The data is directly accessed from the server, providing a neat interface for data structures and methods to query. For the assignment, I have provided you with two datasets, one records the solar activity from May 1, 2010 until May 1, 2015. This is exactly the dataset on which Bobra's experiments were conducted. The second dataset records solar activity from May 1, 2020 until May 1, 2024. Both datasets have the same columns, just different data entries. The following table shows the features of the dataset and the files in which they are saved.

In addition to the above, there are 4 files provided:

1. The geos_data.npy contains all GOES event between the start and end date. Data sent every 12 minutes to the server. Data with no solar events for a given HARP and erroneous coordinate data has been filtered out.



```
{'event_date': '2021-04-19',
 'start_time': '2021.04.19_23:19:00_TAI',
 'peak_time': '2021.04.19_23:42:00_TAI',
 'end_time': '2021.04.19_23:59:00_TAI',
 'goes_class': 'M1.1',
 'goes_latitude': -25,
 'goes_longitude': -24,
 'noaa_active_region': 12816},
```

2. The data_order.npy, which corresponds to the order of the observations for which the model

should be trained.

3. pos_class.npy and neg_class.npy contain information about the positive and negative solar events based on HARP number, peak flare time, and class of energy burst characterizing the strength of electromagnetic field.

Goals: The two datasets are contained within this folder along with their feature files. Your goal is to create an input data array, provide appropriate labels, implement the SVM model, and evaluate it. Please implement functions that:

1. Preprocess the data to prepare it for the model by:
   (a) Normalizing the features
   (b) Removing missing values, if any
   (c) Assigning appropriate labels to positive and negative observations
2. Create an input data array by concatenating all features in a single 2D array. Write this function such that you can input any combination of feature sets and create the input data. For instance, (FS-I only), (FS-I, FS-II), (FS-I, FS-II, FS-IV) are all valid feature set combinations to try (Note that there are 15 total combinations of feature sets).
3. Perform classification using a Support Vector Machine. Use all relevant and possible feature set combinations to build different models and take note of the best performing model. You may want to experiment with the C, gamma, and kernel as hyperparameters (possibly others if you are interested).
4. Perform 5-fold cross validation. The output of your function should be the mean and standard deviation across all folds. For instance, if SVM1 is trained using FS-I, and SVM2 is trained using (FS-I and FS-II), the mean 5-fold CV is reported for SVM1 and SVM2 separately.
5. Calculate the true skill score (TSS) using the following equation: $TSS = \frac{TP}{TP+FN} - \frac{FP}{FP+TN}$. TSS is a good measure to predict rare events when we have a large class imbalance in the input data. Use this as the performance metric for all models.

4 Deliverables

You should submit the code you used as well a PDF report to Avenue. You should include any instructions to run your code in a README file and any disclaimers about the use of AI. Please submit all code as python files or a Jupyter Notebook. Submit these files on Avenue. Do not upload any datasets.

Your report should include the following with plots and results for **BOTH** of the two datasets (2010-15 and 2020-24):

1. Bar chart showing the performance of each model feature set combination across each fold.
2. Visualize performance of all feature sets using a confusion matrix. Make note of the best feature set and best hyperparameters.
3. Report all performance output for SVMs corresponding to all feature set combinations, including mean and standard deviation across folds.
4. Answer the following questions:
   (a) Which feature combination worked best and which feature set was worst? Report the performance.
   (b) Does adding additional feature sets improve the TSS score? What do you observe with each feature set?
   (c) Which dataset led to a better TSS score (2010 or 2020)? Why do you think this is the case?

5 Help

Please use the Teams channel and tutorials to ask questions about the assignment when you need guidance or pointers on this homework. You are free to discuss your approach and ideas with classmates, but should not share code or reuse data. You may use generative AI tools if you find them helpful, but please clearly document how they were used in the report and follow the guidelines in the syllabus for what you **must** include when using generative AI. If you use generative AI and do not report it, you may receive a 0 for the assignment. You take full responsibility for the deliverables you submit.