Machine Learning Challenge
Updated: December 2, 2022

CS 5245: Data Programming
R. M. Parry

Using the built-in functions and classes in sklearn, build models to predict the class labels for the following problems. In particular, you should use one or more `Pipelines` and `GridSearchCV` to select hyperparameters.

There are many examples in the textbook, but I would draw your attention to the section "Hyperparameters and Model Validation" in chapter 5. For data sets 7-12, you will need to use some form of feature selection (`sklearn.feature_selection`) or dimensionality reduction (`sklearn.decomposition`).

For each competition, you can download the training set (with labels), test set (without labels), and an example submission file. When you have trained your model locally, predict every sample in the test set and produce an `npz` file. Upload it to web-cat. Your performance will be revealed on half of the test set. The other half will be used to compare the top models at the end of the competition.

In addition to submitting your test labels to Web-CAT, you must also submit your source code to asulearn along with a brief description of how you arrived at the "best" model for each problem. Your code will be graded by hand independent of your performance on Web-CAT. Specifically, your ability to apply concepts from class using sklearn will be evaluated.

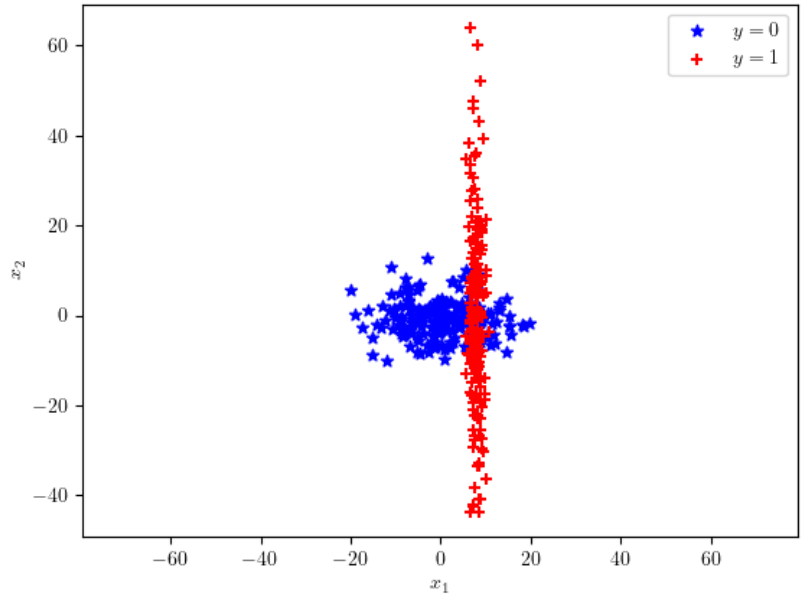
The following code loads the training data from the `npz` file and creates a scatter plot (shown for each 2D data set below).

```
import numpy as np
import matplotlib.pyplot as plt

plt.rcParams['text.usetex'] = True

problem = 1
problem_name = f'challenge_{problem:02d}'

train_file = f'{problem_name}.npz'
data = np.load(train_file)
x_train = data['x_train']
y_train = data['y_train']
x_test = data['x_test']

plt.clf()
plt.scatter(x_train[y_train==0, 0], x_train[y_train==0, 1], label=r'$y=0$',
            color='blue', marker='*')
plt.scatter(x_train[y_train==1, 0], x_train[y_train==1, 1], label=r'$y=1$',
            color='red', marker='+')
plt.xlabel(r'$x_1$')
plt.ylabel(r'$x_2$')
plt.title(f'Data set {problem} train')
plt.legend()
plt.axis('equal')
plt.show()

# model is your classifier (any object with a fit and predict method)
model.fit(x_train, y_train)
y_hat = model.predict(x_test)
```

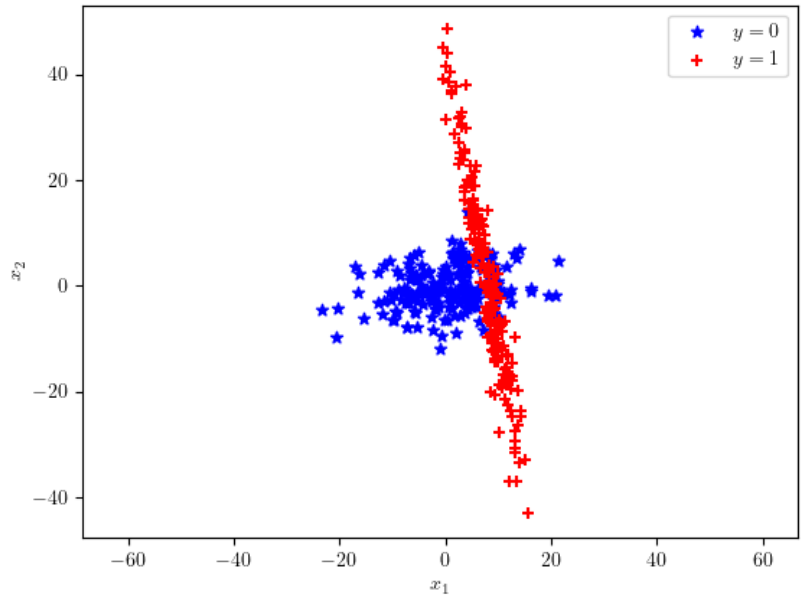
```
submission_file = f'{problem_name}_submission.npz'
np.savez(submission_file, y_test=y_hat)

# Then upload the submission npz to Web-CAT.
```

Challenges:

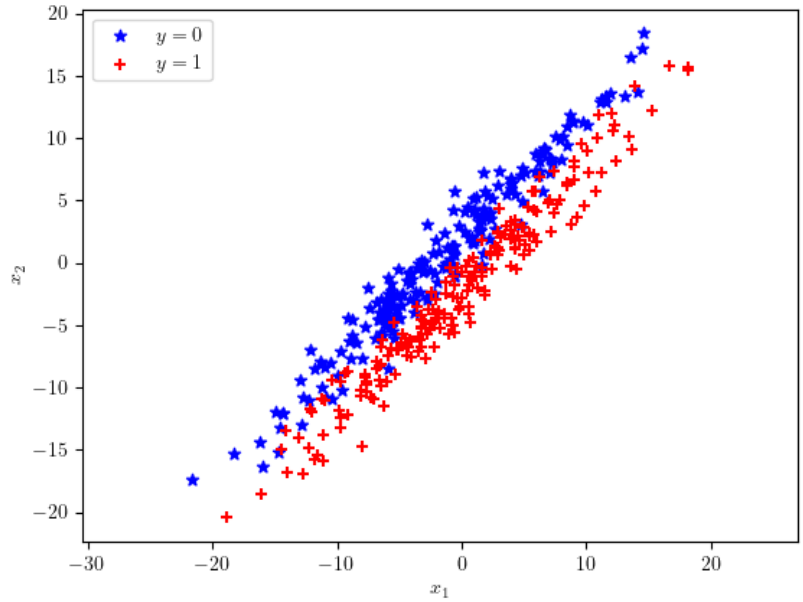
1. https://cs.appstate.edu/~rmp/cs5245/challenge_01.npz
   https://cs.appstate.edu/~rmp/cs5245/challenge_01_submission.npz



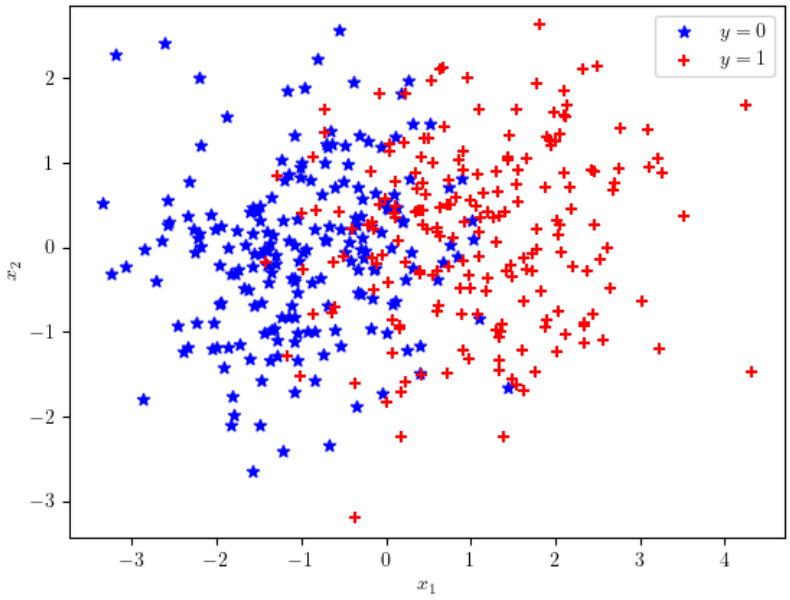
2. https://cs.appstate.edu/~rmp/cs5245/challenge_02.npz
   https://cs.appstate.edu/~rmp/cs5245/challenge_02_submission.npz



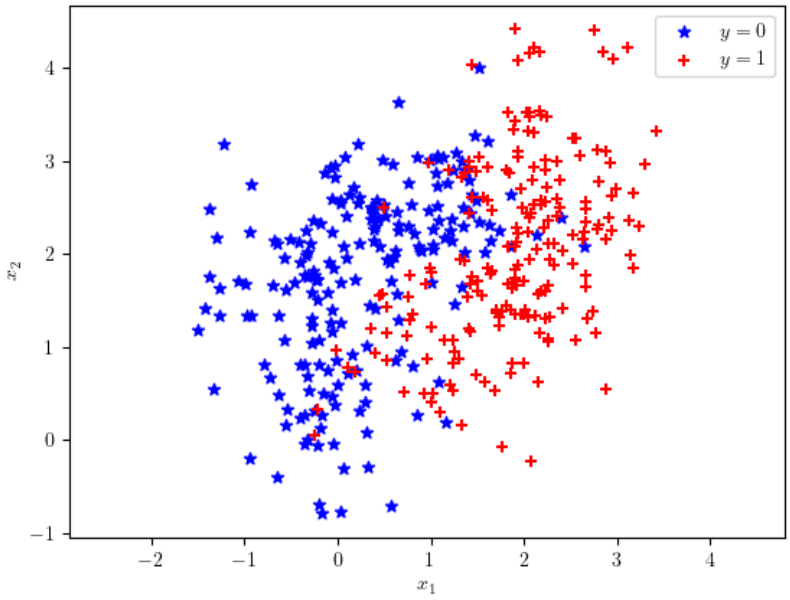
3. `https://cs.appstate.edu/~rmp/cs5245/challenge_03.npz`
   `https://cs.appstate.edu/~rmp/cs5245/challenge_03_submission.npz`



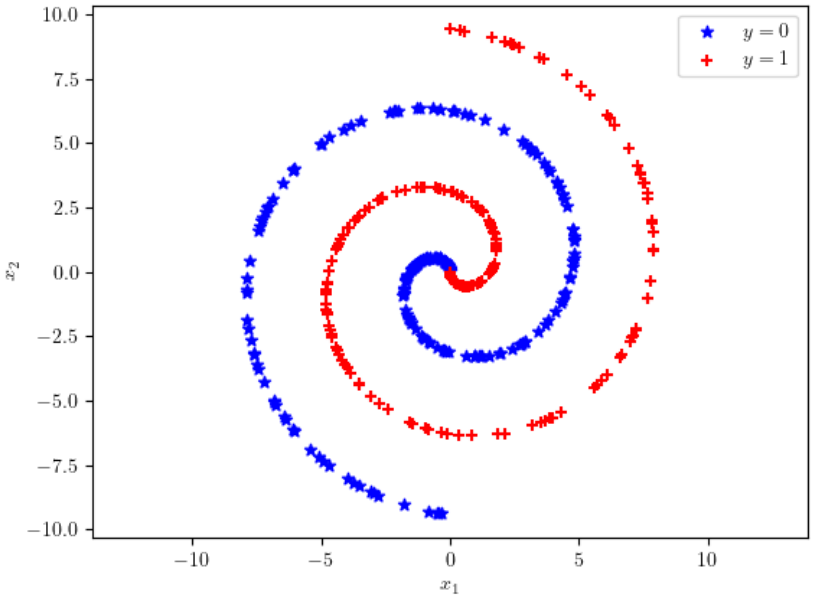
4. `https://cs.appstate.edu/~rmp/cs5245/challenge_04.npz`
   `https://cs.appstate.edu/~rmp/cs5245/challenge_04_submission.npz`



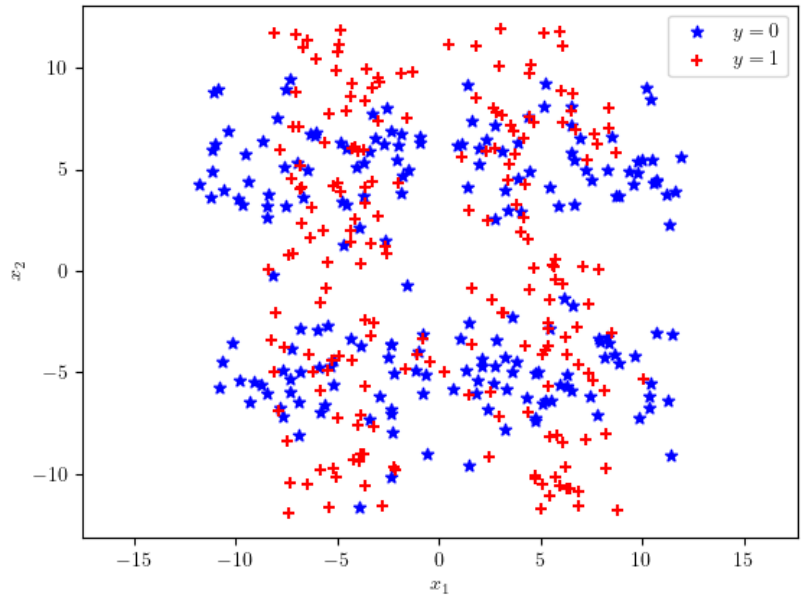
5. https://cs.appstate.edu/~rmp/cs5245/challenge_05.npz
   https://cs.appstate.edu/~rmp/cs5245/challenge_05_submission.npz



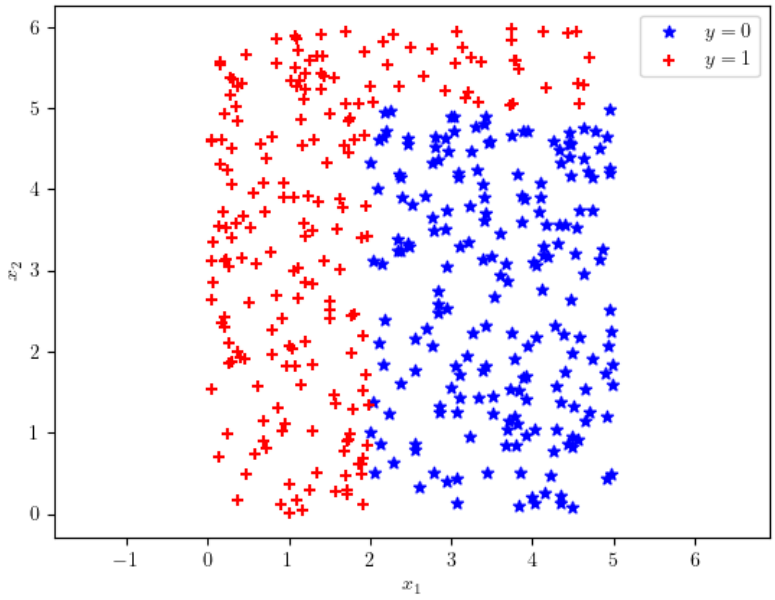
6. https://cs.appstate.edu/~rmp/cs5245/challenge_06.npz
   https://cs.appstate.edu/~rmp/cs5245/challenge_06_submission.npz



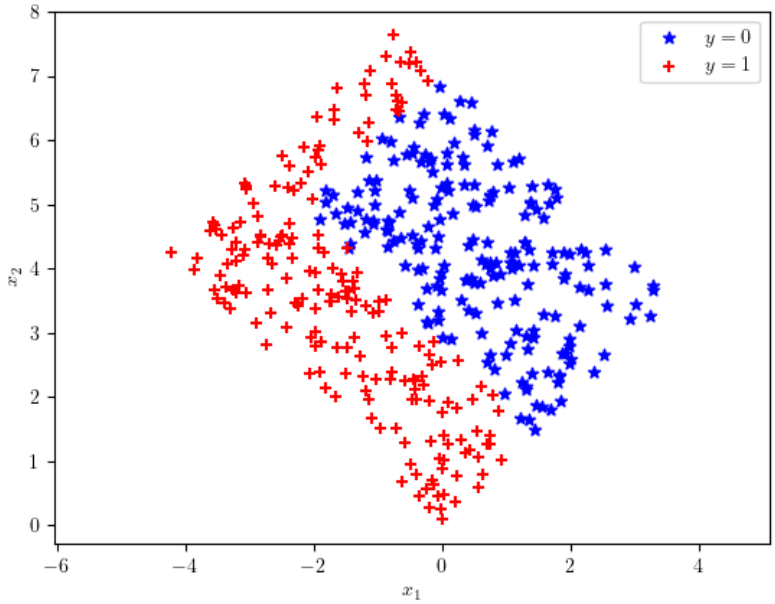
7. `https://cs.appstate.edu/~rmp/cs5245/challenge_07.npz`
   `https://cs.appstate.edu/~rmp/cs5245/challenge_07_submission.npz`



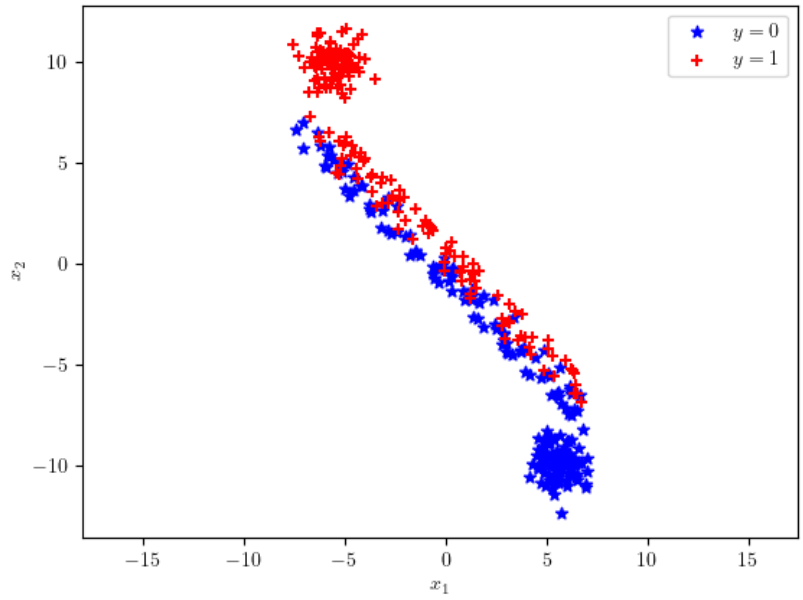
8. `https://cs.appstate.edu/~rmp/cs5245/challenge_08.npz`
   `https://cs.appstate.edu/~rmp/cs5245/challenge_08_submission.npz`



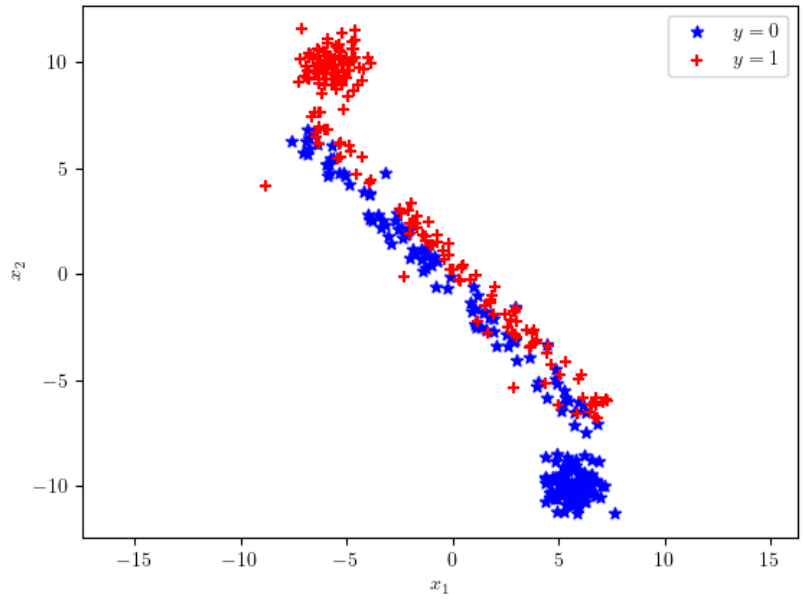
9. `https://cs.appstate.edu/~rmp/cs5245/challenge_09.npz`
   `https://cs.appstate.edu/~rmp/cs5245/challenge_09_submission.npz`



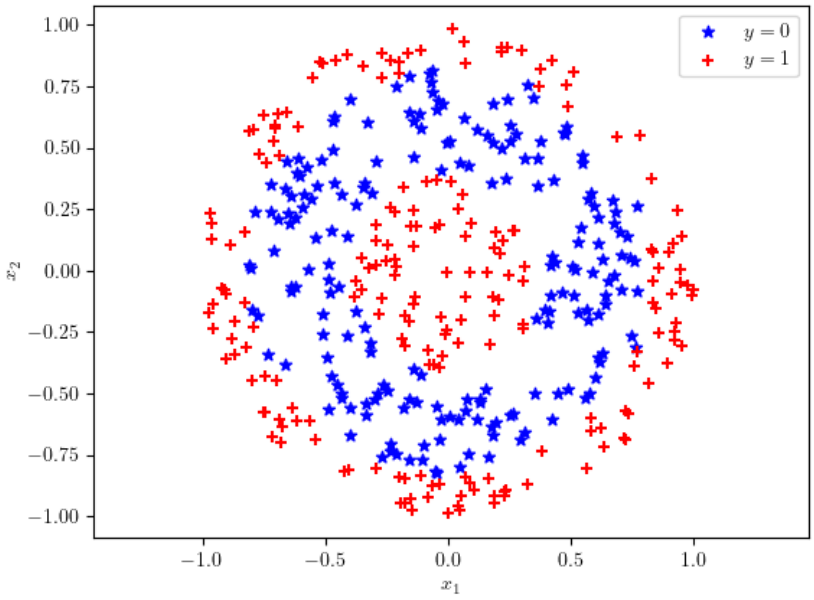
10. `https://cs.appstate.edu/~rmp/cs5245/challenge_10.npz`
    `https://cs.appstate.edu/~rmp/cs5245/challenge_10_submission.npz`



11. `https://cs.appstate.edu/~rmp/cs5245/challenge_11.npz`
    `https://cs.appstate.edu/~rmp/cs5245/challenge_11_submission.npz`



12. `https://cs.appstate.edu/~rmp/cs5245/challenge_12.npz`
    `https://cs.appstate.edu/~rmp/cs5245/challenge_12_submission.npz`


$y = 0$
$y = 1$
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
x_2
−1.0
−0.5
0.0
0.5
1.0
x_1