Homework 3 – Report

**Programming Exercises**

1. **Binary Classification on Text Data**

**Download**

A close-up of a computer screen

Description automatically generated

**Split training data**

A close-up of a text

Description automatically generated

**Preprocess Data**

A screenshot of a computer program

Description automatically generated

**Bag of Words of Model**

A computer code with black text

Description automatically generated with medium confidence

**Part a: Bernoulli Naïve Bayes**

A screenshot of a computer program

Description automatically generated

A computer code with text

Description automatically generated with medium confidence

The F1 score of the development set was 0.73906

**Part b: Model Comparison**

* The F1 score on the L1 regression was 0.70699 while the naïve bayes had an F1 score of 0.72968. Therefore, the naïve bayes model performed best in predicting whether a tweet is a real disaster or not. The pros of using generative vs discriminative models are that generative models can manage missing data better and can recognize speech patterns because of its ability to infer. However, the cons are that it’s more computationally expensive and that it can’t classify data very well.
* The assumptions of naïve bayes are different from logistic regression because it assumes that features are independent and logistic regression assumes that the features have a linear relationship.

1. **Gaussian Discriminant Analysis**

**A close-up of a computer screen

Description automatically generated**

Splitting the data

A computer code with black text

Description automatically generated with medium confidence

**Part a**

A screenshot of a computer code

Description automatically generated

A diagram of different colored dots

Description automatically generated

I notice that the Setosa class is pretty distinguished in terms of sepal length and sepal width ranging from width [2.8, 4.4] and sepal length [4.4, 5.6]. Meanwhile, Versicolor and Virginica tend to have similar ranges of sepal width around [2.3, 3.5] typically and while Versicolor has a smaller range of sepal length, there seems to be a substantial overlap between the sepal lengths of Versicolor and Virginica around [5.5, 7.0].

**Part b**

**A screenshot of a computer code

Description automatically generated**

**Part c**

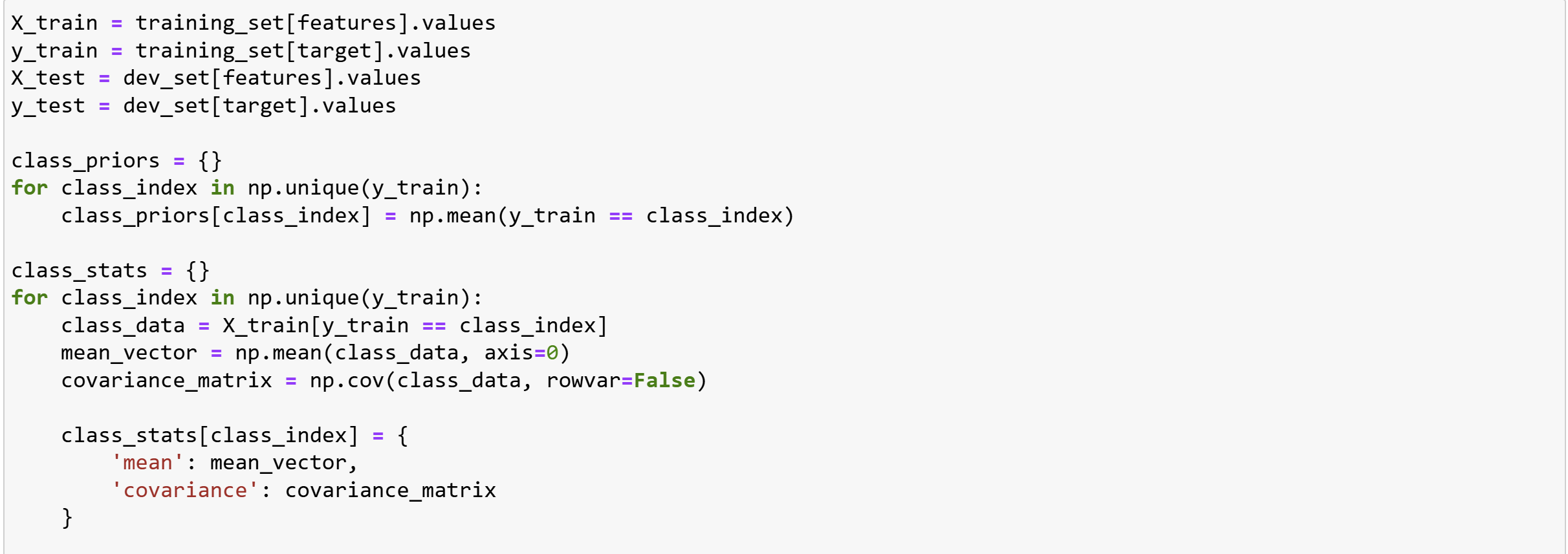
A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

**Part d i**

**** ****

**A screen shot of a computer screen

Description automatically generated**

**Part dii**

I will be using accuracy because I want to measure the performance of the classifier, as opposed to F1 that is about imbalance or precision which is concerned about false positives/negatives. The accuracy of the algorithm is 1.00 which is usually strange but makes sense in this context since the data set is extremely small and very well filtered.

**Part e**

A computer screen shot of a computer code

Description automatically generated

A diagram of a number of dots

Description automatically generated

I decided to try and run the clusters and it does seem to change, setosa in particular seems to take over the points where versicolor used to be. However, the virginica cluster is pretty similar to the one in part a, plus a few points from versicolor that were overlapping.

**Written Exercises**

**a. Naïve Bayes with binary features**

1. The Naïve Bayes assumption in this context is that bikes and skis are conditionally independent when given a student’s program (Master’s or phd)
2. The probability of a student in this group who neither bikes or skis being a master student is:

Prior probabilities:

P(y=0)=20/50​=0.4 and P(y=1)=30/50=0.6

Likelihoods:

P(x1 = 0, x2 = 0 | y = 0) =

P(x1 = 0 | y = 0) x P(x2 = 0 | y = 0) =

0.75 x0.75 = 0.5625 P(x1 = 0, x2 = 0 | y = 1) =

P(x1 = 0 | y = 1) x P(x2 = 0 | y = 1) =

0.333 x 0.5 = 0.16665

P(x1 = 0, x2 = 0) = P(x1 = 0, x2 = 0 | y = 0) x P(y = 0) + P(x1 = 0, x2 = 0 | y = 1) x P(y = 1)P(x1 = 0, x2 = 0) =

0.5625 x 0.4 + 0.16665 x 0.6 P(x1 = 0, x2 = 0) =

0.225 + 0.09999 = 0.32499

Posterior Probability:

P(y = 0 | x1 = 0, x2 = 0) =

(0.5625\*0.4)/(0.32499)P(y = 0 | x1 = 0, x2 = 0) =

0.225/0.32499 P(y = 0 | x1 = 0, x2 = 0) = 0.6924

1. It would not make sense to assume that the probability of biking and skiing are conditionally independent for a phd student. I would change my answer in part b by stating that if a student cannot bike, they cannot ski either and vice versa.

**b. Categorical Naïve Bayes**

**A white sheet of lined paper with blue writing

Description automatically generated**

**A white sheet of paper with writing on it

Description automatically generated**

**c. Weights for Clustering**

**A close-up of a math problem

Description automatically generated**

**A math equations on a piece of paper

Description automatically generated**