A2Q2

March 2, 2020

1 CMPT 423/820

1.1 Assignment 2 Question 2

- Seyedeh Mina Mousavifar
- 11279515
- sem311

```
In [3]: import pandas as pd

        # creating header for the dataset
        header = list()
        for i in range(14):
            if i == 0:
                header.append('label')
                continue
            string = 'feature' + str(i)
            header.append(string)


        # reading dataset and adding header
        data = pd.read_csv('data/a2q3.csv',
                           header=None,
                           names=header)
```

1.1.1 Part 1.

Plot the class densities for all 13 features, similar to A1Q7 Task 4.

```
In [4]: import matplotlib.pyplot as plt

        for i in range(1, 14):

            # select only feature_i and label columns
            string = 'feature' + str(i)
            split = data.filter(items=[string, 'label'])

            # creating plot
```
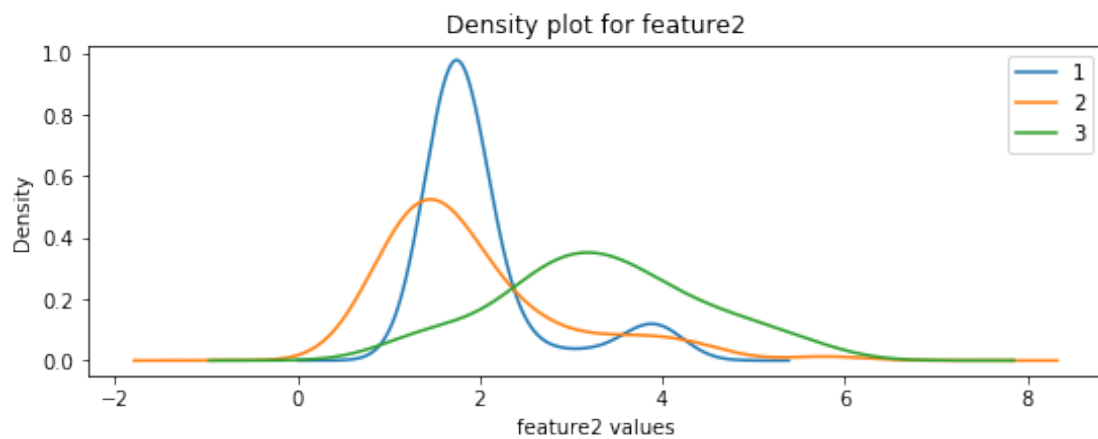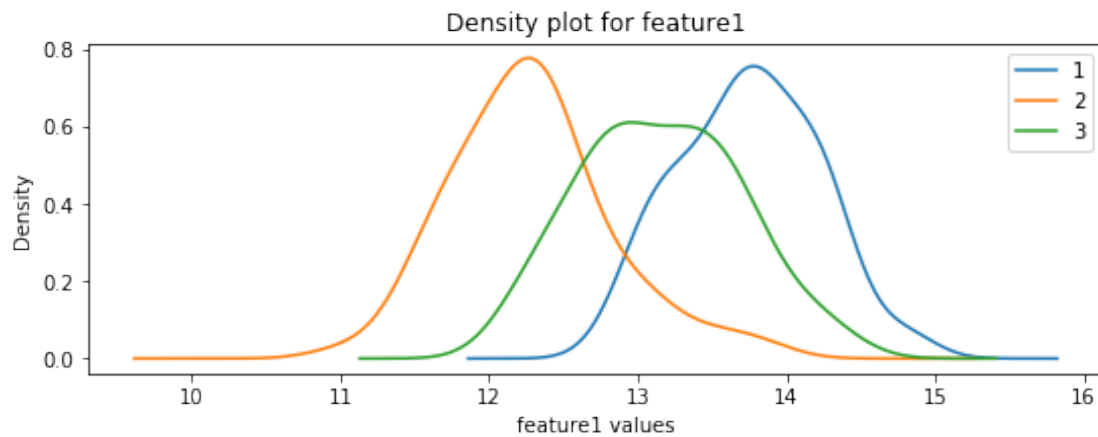
```python
fig, ax = plt.subplots(figsize=(9, 3))

for label, df in split.groupby('label'):
    tmp_df = df[string]
    tmp_df.plot(kind="kde", ax=ax, label=label)

plt.title('Density plot for ' + string)
plt.xlabel(string + ' values')
plot = plt.legend()

plt.show()
```
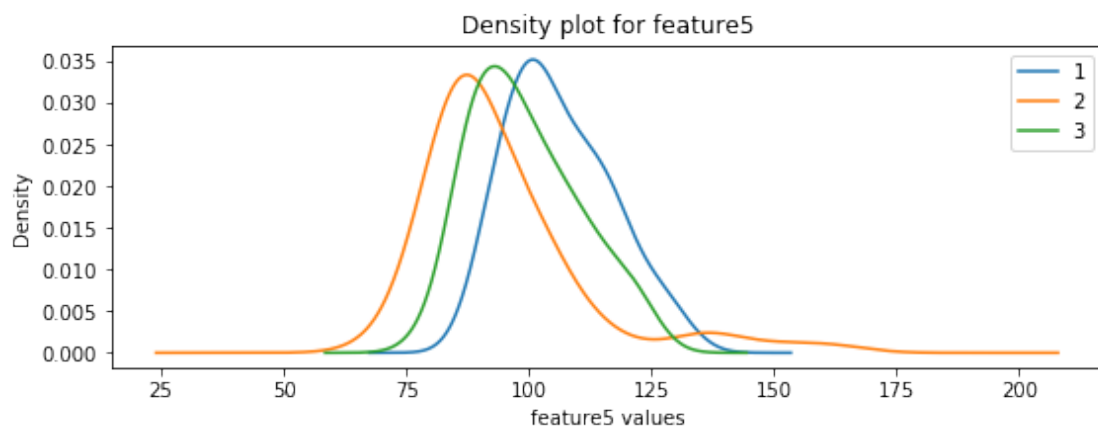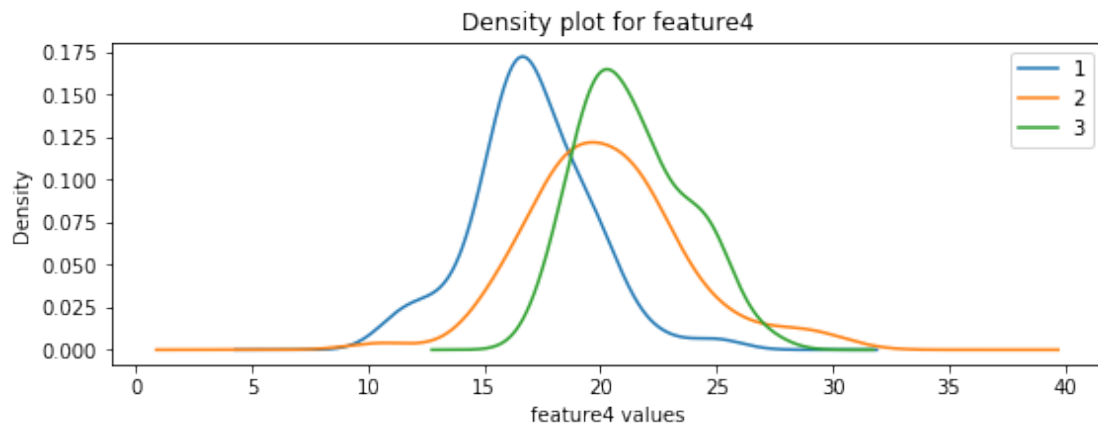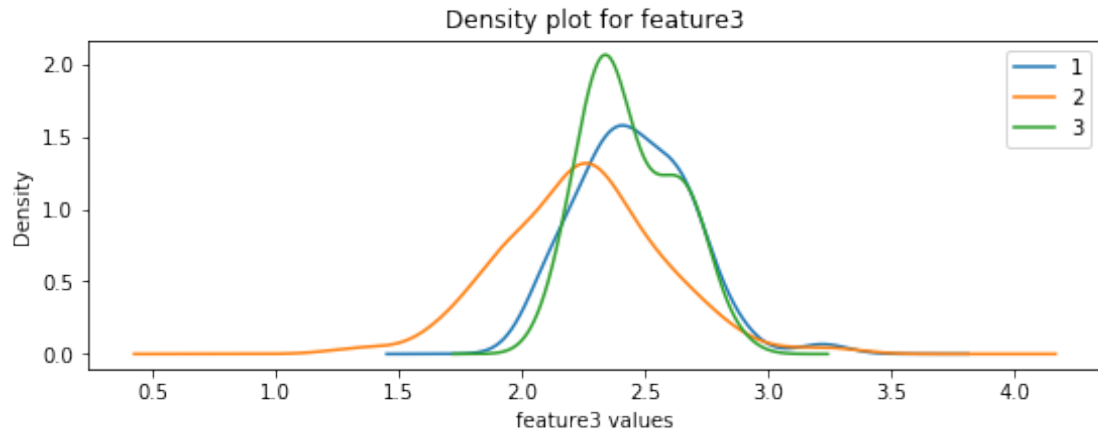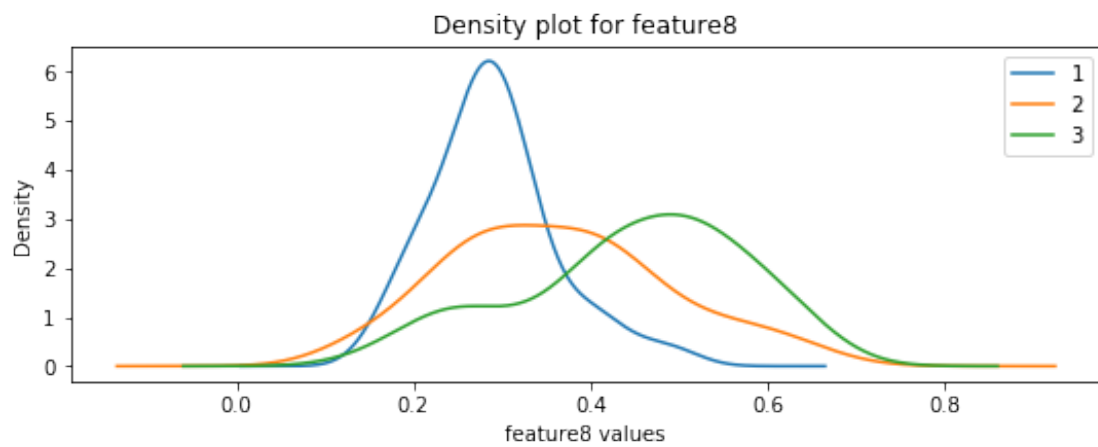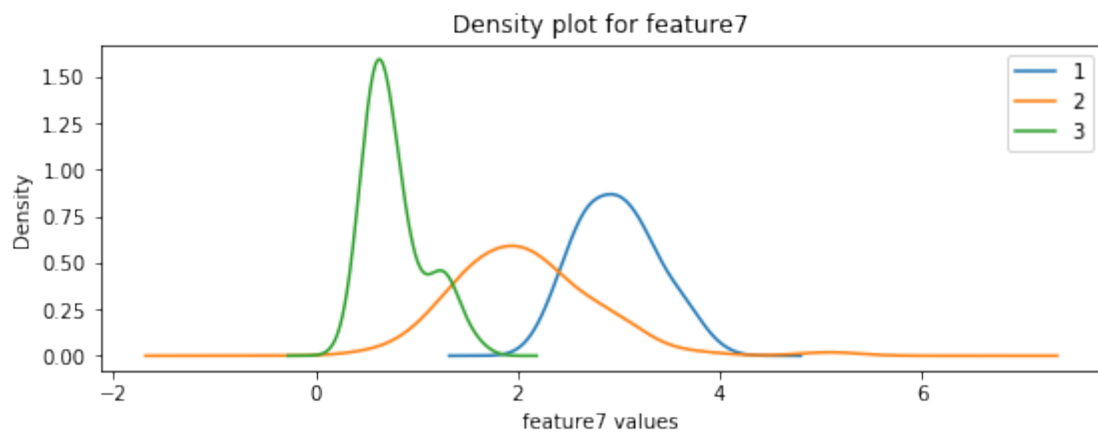
Density plot for feature1

Density plot for feature2

Density plot for feature3

Density plot for feature4

Density plot for feature5

Density plot for feature6

Density plot for feature7

Density plot for feature8

Density plot for feature9

Density plot for feature10

Density plot for feature11

Density plot for feature12

Density plot for feature13
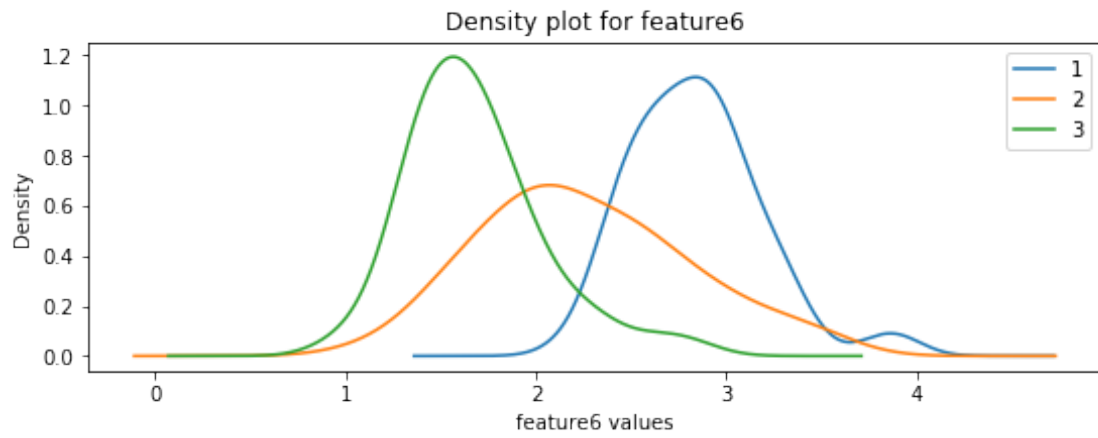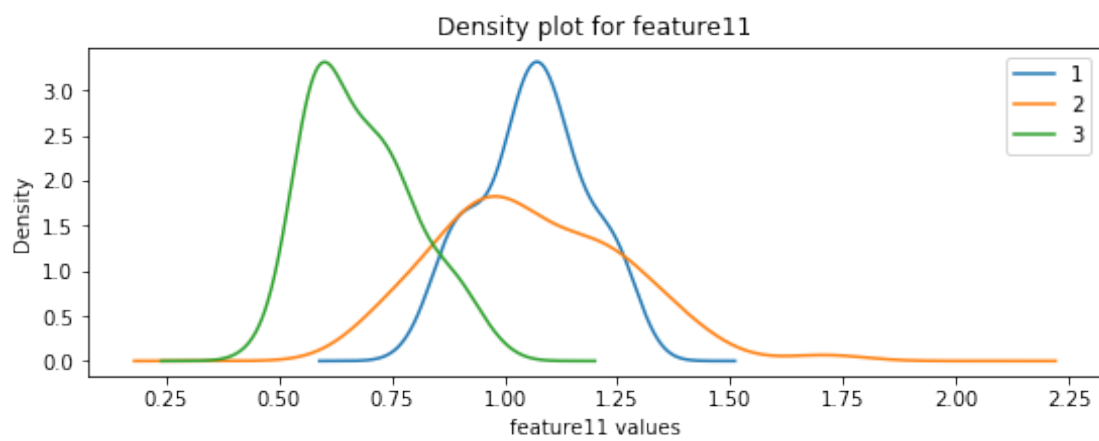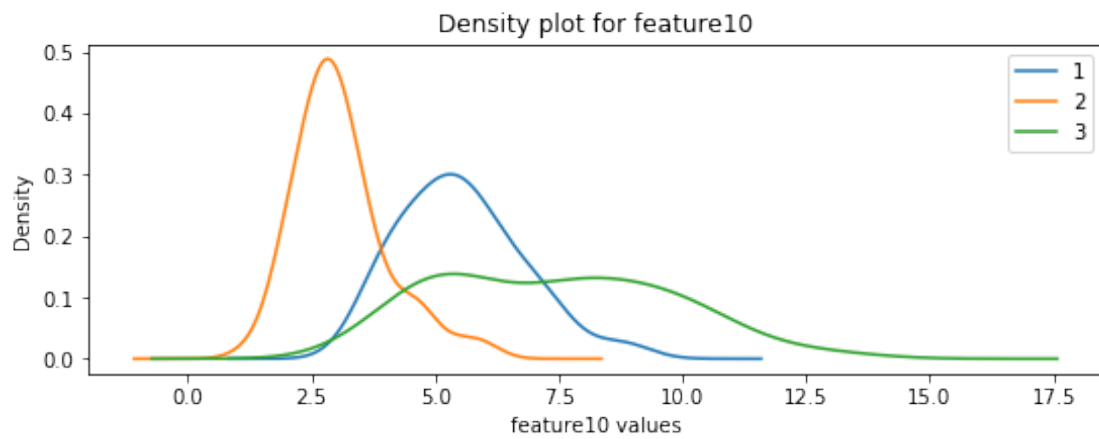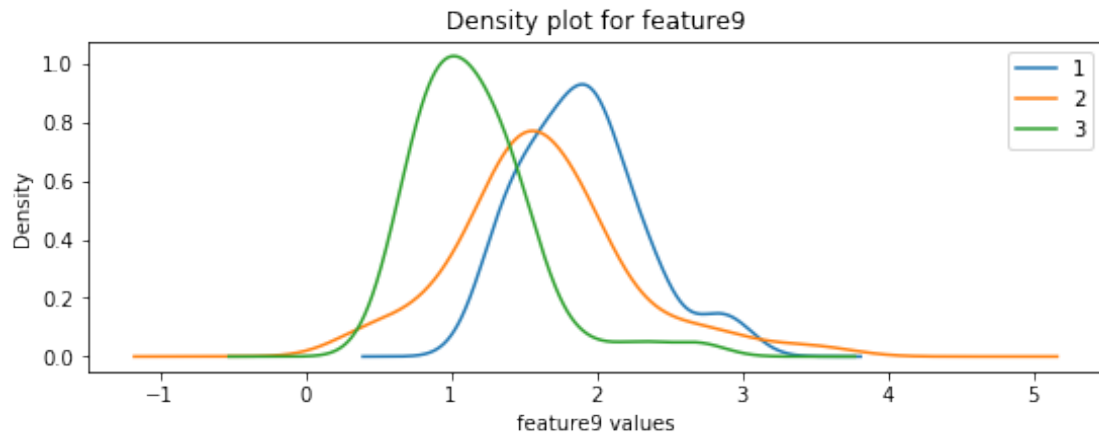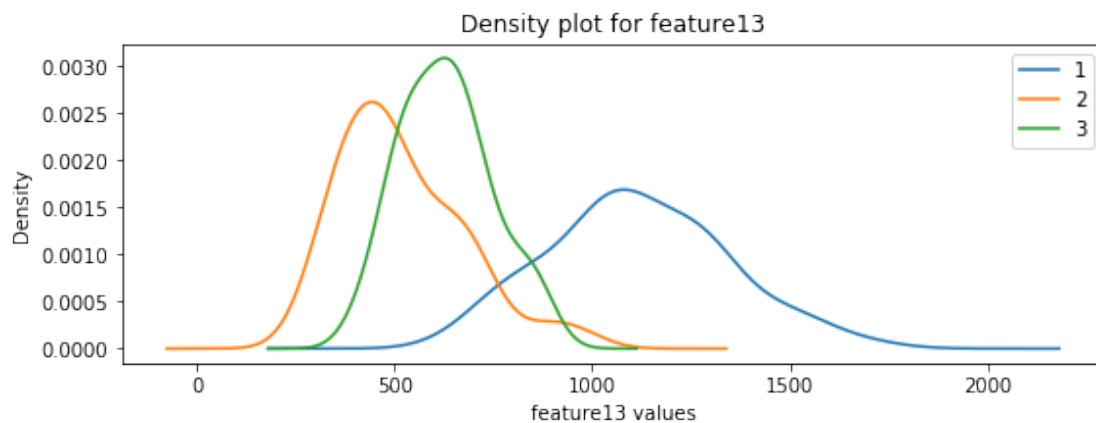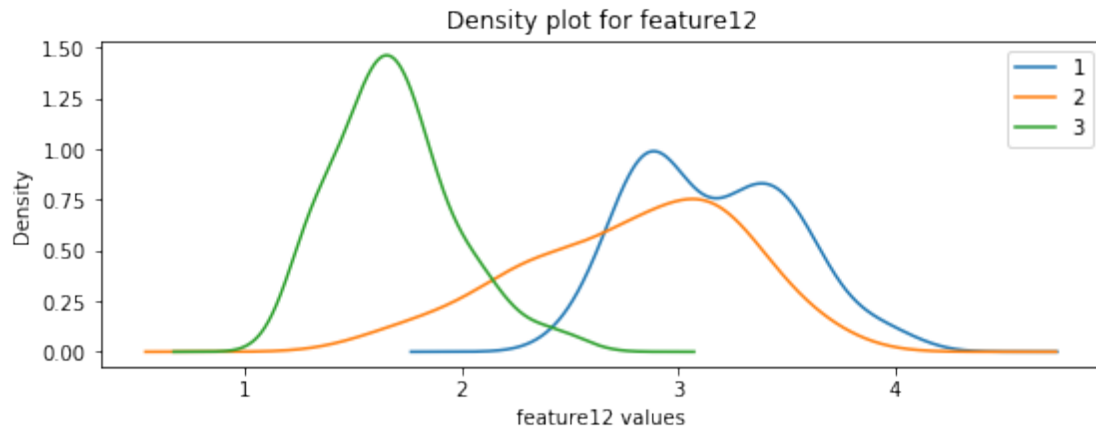
We compare the plots above, based on the curves overlap.

- *feature 1*: This feature might be right for discriminating label 2 from other labels, but might have problems to identify label 1 and 3 due to their overlap.
- *feature 2*: This feature might be right for discriminating label 3 from other labels, but might have problems to identify label 1 and 2 because most of their data interfere with each other.
- *feature 3*: This might be one of the least separating features because all curves profoundly interfere with each other.
- *feature 4*: This feature might be right for discriminating label 1 from other labels, but might have problems to identify label 2 and 3 due to their overlap.
- *feature 5*: This feature has more classification power compared to feature 3, but compared to features seen above, it might struggle more in specifying groups because of the overlay.
- *feature 6*: This feature is the second-best distinguishing feature because it can determine group 1 from group 3, however group 2 overlap with each group notably.

- *feature 7*: In my opinion, this is the best feature for classification, because the curves have the least overlap with each other.
- *feature 8*: This might be one of the least separating features because all curves highly interfere with each other.
- *feature 9*: This feature can specify group 1 from group 3, however, group 2 overlap with each group notably.
- *feature 10*: This feature might be right for discriminating label 2 from other labels, but might have problems to identify label 1 and 3 because most of their data interfere with each other.
- *feature 11*: This feature might be right for discriminating label 3 from other labels, but might have problems to identify label 1 and 2 due to their overlap.
- *feature 12*: This feature might be right for discriminating label 3 from other labels, but might have problems to identify label 1 and 2 because most of their data interfere with each other.
- *feature 13*: This feature is the third-best discriminating feature, because it has high potential to classify label 1 correctly, and has one the least overlaps between label 2 and 3.

Which, if any, of the 13 features, would you pick as the single feature in a 1-feature classifier?

I would pick feature 7 as a single feature because its curves have the least interference with each other, so classification and recognizing groups differences would be easier.

Prior to building a classifier, do you think a classifier based on this data will have high accuracy?

If we use all of the features together, we might not get a good result, because the order of the curves aren't consistent and they highly overlap with each other. However, I think that by pruning misleading and conflicting features, we might get an acceptable result. Especially because feature 7 can highly discriminate the groups.