uOttawa

Faculté de génie
Faculty of Engineering

Applied Machine Learning - Summer 2021

Assignment 2 - Parametric Methods

Submitted by:

Name	Email
Youssef Metwally	ymetw027@uOttawa.ca
Dina Ibrahim Morsy	dabde007@uOttawa.ca
Adel El Nabarawy	aelna025@uOttawa.ca

Submitted to:

Dr. Murat SIMSEK, SMIEEE

Part 1:

1. Given the training data in the table (Downsized Iris Dataset), predict the class of following example using Naïve Bayes Classification.

* For Class 0: "setosa"

Feature 1:

$$\mu_{11} = \frac{4.6 + 4.9 + 5.4 + 5.7 + 4.8}{5} = 5.08$$

$$\sigma_{11}^2 = \frac{(4.6-5.08)^2 + (4.9-5.08)^2 + (5.4-5.08)^2 + (5.7-5.08)^2 + (4.8-5.08)^2}{5}$$

$$= 0.1656$$

$$\sigma_{11} = 0.407$$

Feature 2:

$$\mu_{12} = \frac{3.4 + 3 + 3.4 + 4.4 + 3.4}{5} = 3.52$$

$$\sigma_{12}^2 = \frac{(3.4-3.52)^2 + (3-3.52)^2 + (3.4-3.52)^2 + (4.4-3.52)^2 + (3.4-3.52)^2}{5}$$

$$= 0.2176$$

$$\sigma_{12} = 0.466$$

Feature 3:

$$\mu_{13} = \frac{1.4 + 1.4 + 1.7 + 1.5 + 1.6}{5} = 1.52$$

$$\sigma_{13}^2 = \frac{(1.4-1.52)^2 + (1.4-1.52)^2 + (1.7-1.52)^2 + (1.5-1.52)^2 + (1.6-1.52)^2}{5}$$

$$= 0.0136$$

$$\sigma_{13} = 0.1166$$

Feature 4:

$$\mu_{14} = \frac{0.3+0.2+0.2+0.4+0.2}{5} = 0.26$$

$$\sigma_{14}^2 = \frac{(0.3-0.26)^2+(0.2-0.26)^2+(0.2-0.26)^2+(0.4-0.26)^2+(0.2-0.26)^2}{5}$$

$$= 0.0064$$

$$\sigma_{14} = 0.08$$

* For Class 1: 'Versicolor'

* Feature 1:

$$\mu_{21} = \frac{6.3 + 6.4 + 5.9 + 6.7 + 5.9}{5} = 6.24$$

$$\sigma_{21}^2 = \frac{(6.3 - 6.24)^2 + (6.4 - 6.24)^2 + (5.9 - 6.24)^2 + (6.7 - 6.24)^2 + (5.9 - 6.24)^2}{5}$$

$$= 0.094$$

$$\sigma_{21} = 0.307$$

* Feature 2:

$$\mu_{22} = \frac{3.3 + 3.2 + 3.2 + 3.1 + 3}{5} = 3.16$$

$$\sigma_{22}^2 = \frac{(3.3 - 3.16)^2 + (3.2 - 3.16)^2 + (3.2 - 3.16)^2 + (3.1 - 3.16)^2 + (3 - 3.16)^2}{5}$$

$$= 0.01$$

$$\sigma_{22} = 0.1$$

* Feature 3:

$$\mu_{23} = \frac{4.7 + 4.5 + 4.8 + 4.4 + 4.2}{5} = 4.52$$

$$\sigma_{23}^2 = \frac{(4.7 - 4.52)^2 + (4.5 - 4.52)^2 + (4.8 - 4.52)^2 + (4.4 - 4.52)^2 + (4.2 - 4.52)^2}{5}$$

$$= 0.045$$

$$\sigma_{23} = 0.213$$

* Feature 4

$$\mu_{24} = \frac{1.6 + 1.5 + 1.8 + 1.4 + 1.5}{5} = 1.56$$

$$\sigma_{24}^2 = \frac{(1.6-1.56)^2 + (1.5-1.56)^2 + (1.8-1.56)^2 + (1.4-1.56)^2 + (1.5-1.56)^2}{5}$$

$$= 0.018$$

$$\sigma_{24} = 0.136$$

* For class 2: 'virginica'

* feature 1

$$\mu_{31} = \frac{4.9+5.8+6.9+6.4+6.4}{5} = 6.08$$

$$\sigma_{31}^2 = \frac{(4.9-6.08)^2+(5.8-6.08)^2+(6.9-6.08)^2+(6.4-6.08)^2+(6.4-6.08)^2}{5}$$

$$= 0.47$$

$$\sigma_{31} = 0.69$$

* feature 2

$$\mu_{32} = \frac{2.5+2.7+3.2+3.2+2.7}{5} = 2.86$$

$$\sigma_{32}^2 = \frac{(2.5-2.86)^2+(2.7-2.86)^2+(3.2-2.86)^2+(3.2-2.86)^2+(2.7-2.86)^2}{5}$$

$$= 0.082$$

$$\sigma_{32} = 0.287$$

* feature 3

$$\mu_{33} = \frac{4.5+5.1+5.7+5.3+5.3}{5} = 5.18$$

$$\sigma_{33}^2 = \frac{(4.5-5.18)^2+(5.1-5.18)^2+(5.3-5.18)^2+(5.3-5.18)+(5.7-5.18)^2}{5}$$

$$= 0.15$$

$$\sigma_{33} = 0.39$$

* Feature 4

$$\mu_{34} = \frac{1.7 + 1.9 + 2.3 + 1.9 + 2.3}{5} = 2.02$$

$$\sigma_{34}^2 = \frac{(1.7 - 2.02)^2 + (1.9 - 2.02)^2 + (2.3 - 2.02)^2 + (2.3 - 2.02)^2 + (1.9 - 2.02)^2}{5}$$

$$= 0.057$$

$$\sigma_{34} = 0.24$$

* Prior

$$P(C_1) = P(C_2) = P(C_3) = \frac{5}{15} = \frac{1}{3}$$

* likelihood For

(Sepal length = 6.9, Sepal width = 3.1, petal length = 5.4, petal width = 2.1)

* $P(X_i | C_1) =$

$$\frac{1}{\sqrt{2\pi(0.166)}} \exp\left(-\frac{(6.9-5.08)^2}{2\times(0.166)}\right) \times \frac{1}{\sqrt{2\pi(0.217)}} \exp\left(-\frac{(3.1-3.52)^2}{2\times(0.217)}\right) \times$$

$$\frac{1}{\sqrt{2\pi(0.014)}} \exp\left(-\frac{(5.4-1.52)^2}{2\times(0.014)}\right) \times \frac{1}{\sqrt{2\pi(0.006)}} \exp\left(-\frac{(2.1-0.26)^2}{2\times 0.0064}\right)$$

$$= 0$$

* $P(X_i | C_2) =$

$$\frac{1}{\sqrt{2\pi(0.094)}} \exp\left(-\frac{(6.9-6.24)^2}{2\times(0.94)}\right) \times \frac{1}{\sqrt{2\pi(0.01)}} \exp\left(-\frac{(3.1-3.159)^2}{2\times 0.01}\right)$$

$$\times \frac{1}{\sqrt{2\pi(0.046)}} \exp\left(-\frac{(5.4-4.52)^2}{2\times 0.046}\right) \times \frac{1}{\sqrt{2\pi(0.018)}} \exp\left(-\frac{(2.1-1.56)^2}{2\times 0.018}\right)$$

$$= 1.736\times 10^{-7}$$

$P(x_i | C_3) =$

$$\frac{1}{\sqrt{2\pi(0.47)}} \exp\left(-\frac{(6.9 - 6.08)^2}{2 \times 0.47}\right) \times \frac{1}{\sqrt{2\pi(0.08)}} \exp\left(-\frac{(3.1 - 2.86)^2}{2 \times 0.08}\right)$$

$$\times \frac{1}{\sqrt{2\pi(0.15)}} \exp\left(-\frac{(5.4 - 5.480)^2}{2 \times 0.15}\right) \times$$

$$\frac{1}{\sqrt{2\pi(0.058)}} \exp\left(-\frac{(2.1 - 2.02)^2}{2 \times (0.058)}\right)$$

$$\approx 0.381$$

* posterior

$$P(C_i|X_i) = \frac{P(X_i|C_i)\,P(C_i)}{P(X_i)}$$

$$P(C_1|X_i) = 0 / P(X_i) = 0$$

$$P(C_2|X_i) = 5.78 \times 10^{-8} / P(X_i)$$

$$P(C_3|X_i) = 0.127 / P(X_i)$$

$$\frac{P(C_2|X_i)}{P(C_3|X_i)} = \frac{5.78 \times 10^{-8} / P(X_i)}{0.127 / P(X_i)} = 4.55 \times 10^{-7} < 0$$

$\therefore P(C_3|X_i)$ $\therefore$ it's virginica.

Part 2:

1. Load the Iris dataset:

```
iris_ds = load_iris()
data = iris_ds.data
labels = iris_ds.target
```

2. Drop the petal length and petal width features to form a 2D Iris dataset:

```
X = data[:,:2]
y = labels[:]
```

3. Plot the likelihoods of first feature (Sepal length) for each class as given below and apply Naïve Bayes Classifier to 2D Iris dataset to predict classes. Plot posterior probabilities and calculate the accuracy:

```
# Plot the likelihoods of first feature (Sepal length)
x1 = X[:, 0][y == 0]
sigma1 = np.std(x1)
mean1 = np.mean(x1)
x2 = X[:, 0][y == 1]
sigma2 = np.std(x2)
mean2 = np.mean(x2)
x3 = X[:, 0][y == 2]
sigma3 = np.std(x3)
mean3 = np.mean(x3)
x=np.linspace(1, 10, 100)
plt.plot(x, stats.norm.pdf(x, mean1, sigma1), label=iris_ds.tar-
get_names[0])
plt.plot(x, stats.norm.pdf(x, mean2, sigma2), label=iris_ds.tar-
get_names[1])
plt.plot(x, stats.norm.pdf(x, mean3, sigma3), label=iris_ds.tar-
get_names[2])
plt.xlabel('Sepal length')
plt.ylabel('Likelihood')
plt.legend()
plt.show()
```

```
# apply Naïve Bayes Classifier
cls = GaussianNB()
cls.fit(X, y)
y_pred = cls.predict(X)

# calculate the accuracy
print("accurecy: ", accuracy_score(y, y_pred))

# Plot posterior probabilities
plt.plot(cls.predict_proba(X)[:, 0], label=iris_ds.target_names[0])
plt.plot(cls.predict_proba(X)[:, 1], label=iris_ds.target_names[1])
plt.plot(cls.predict_proba(X)[:, 2], label=iris_ds.target_names[2])
plt.legend()
```

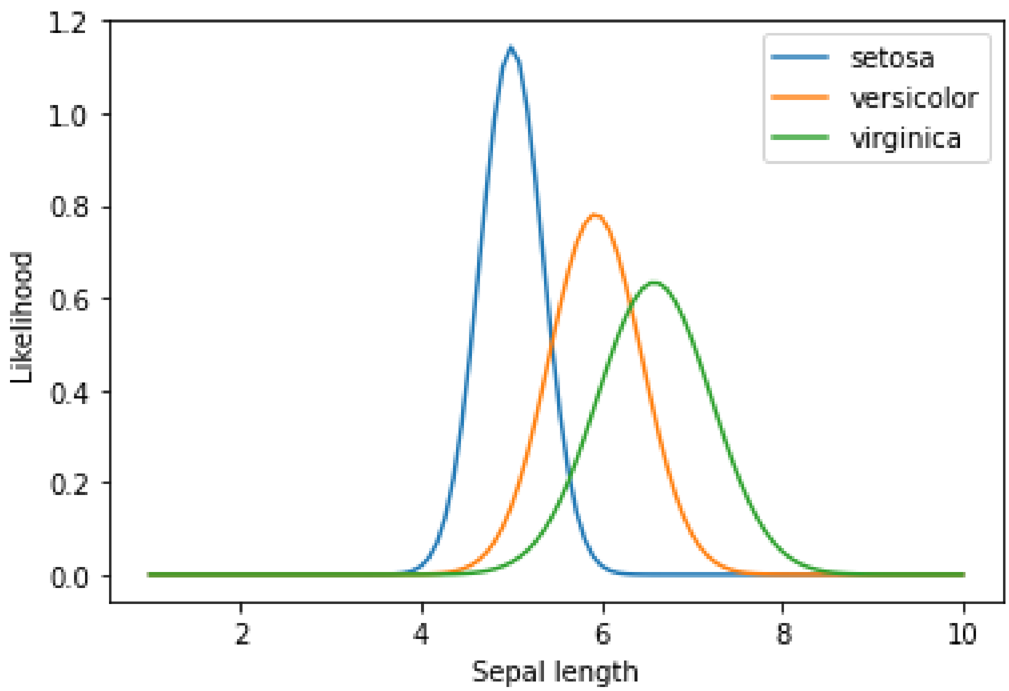


Figure 1: Likelihoods of first feature

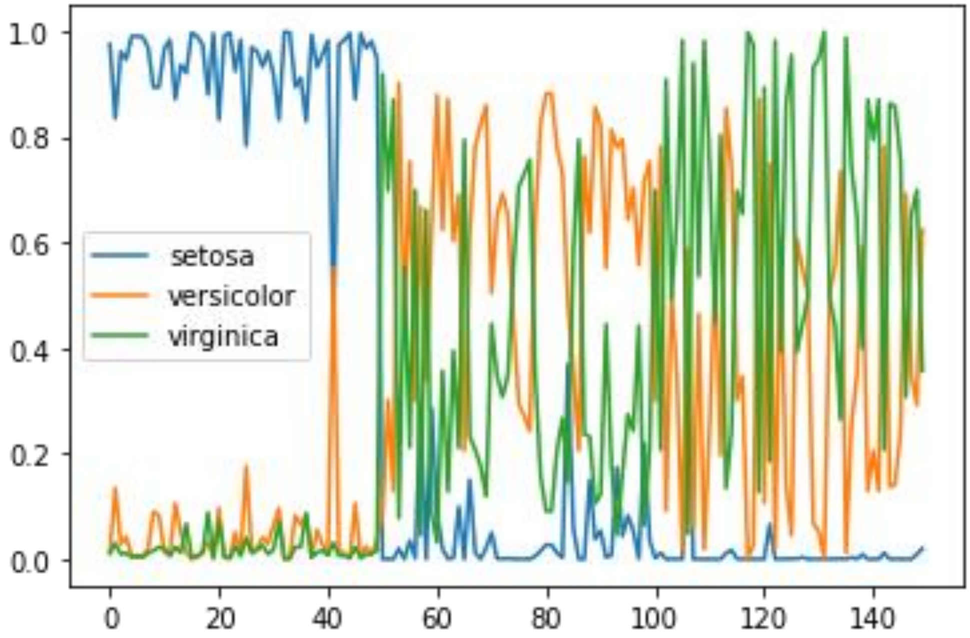


Figure 2: Posterior Probabilities

```
accurecy:  0.78
```

<u>4. Now change the actual values of and to the given values below for each class. Plot the likelihoods of the first feature (Sepal length) with updated and values for each class and apply Naïve Bayes Classifier. Obtain accuracy value for each case and make a comment based on it.</u>

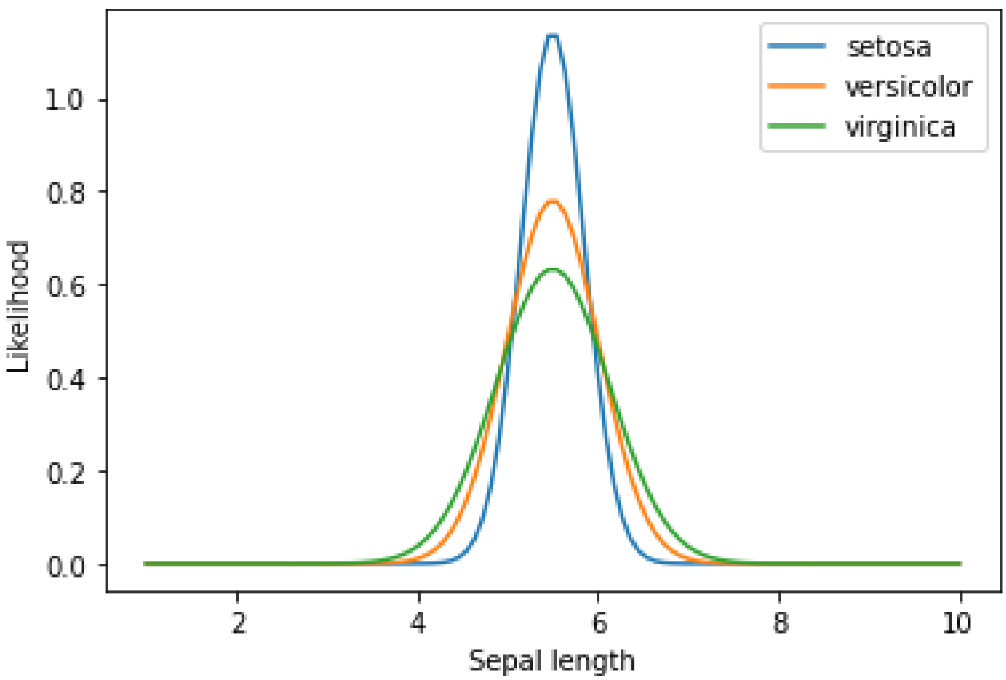
<u>Case 1:</u>

```
cls = GaussianNB()
cls.fit(X, y)
cls.theta_[:, 0] = 5.5
y_pred = cls.predict(X)
print("accurecy: ", accuracy_score(y, y_pred))
x1 = X[:, 0][y == 0]
sigma1 = np.std(x1)
mean1 = np.mean(x1)
x2 = X[:, 0][y == 1]
sigma2 = np.std(x2)
mean2 = np.mean(x2)
x3 = X[:, 0][y == 2]
sigma3 = np.std(x3)
```

```
mean3 = np.mean(x3)
x=np.linspace(1, 10, 100)
plt.plot(x, stats.norm.pdf(x, 5.5, sigma1), label=iris_ds.tar-
get_names[0])
plt.plot(x, stats.norm.pdf(x, 5.5, sigma2), label=iris_ds.tar-
get_names[1])
plt.plot(x, stats.norm.pdf(x, 5.5, sigma3), label=iris_ds.tar-
get_names[2])
plt.xlabel('Sepal length')
plt.ylabel('Likelihood')
plt.legend()
plt.show()

plt.plot(cls.predict_proba(X)[:, 0], label=iris_ds.target_names[0])
plt.plot(cls.predict_proba(X)[:, 1], label=iris_ds.target_names[1])
plt.plot(cls.predict_proba(X)[:, 2], label=iris_ds.target_names[2])
plt.legend()
```

Figure 3: Likelihoods plot



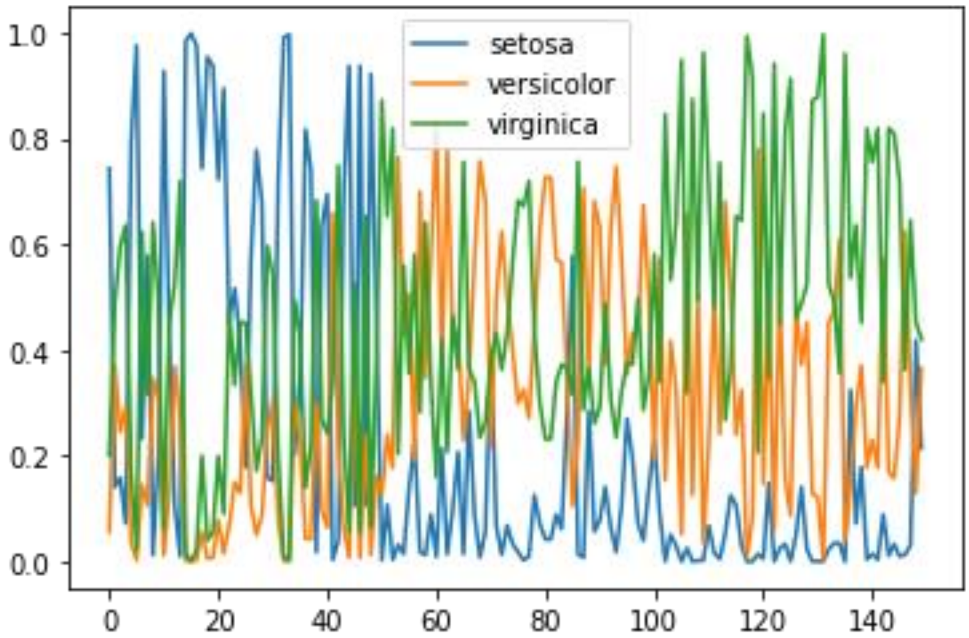


Figure 4: Posterior Probabilities

accurecy: 0.6266666666666667

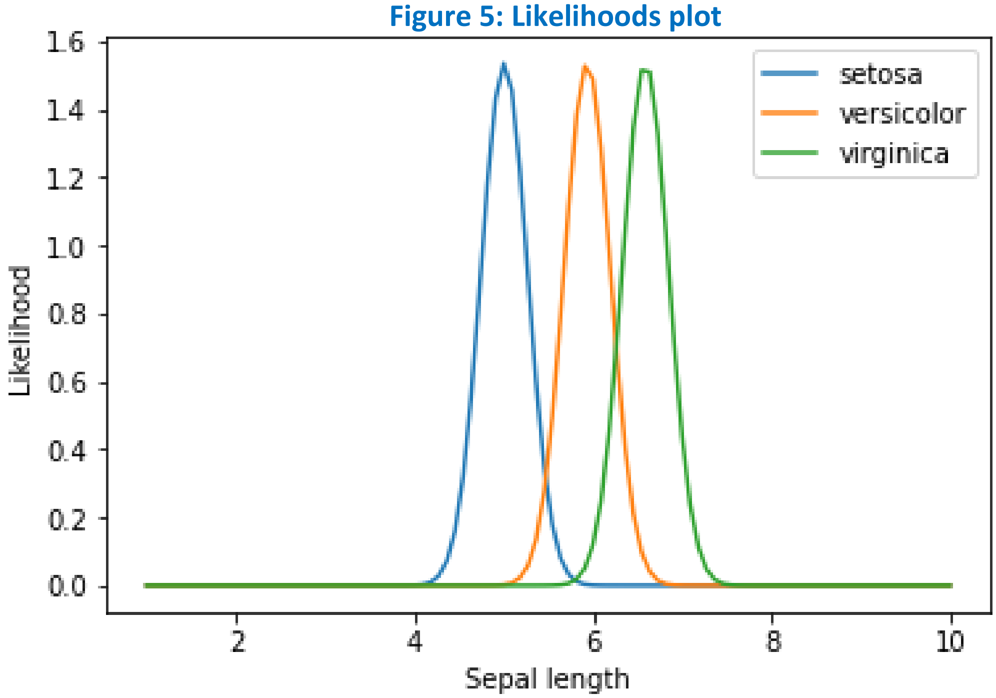
Case 2:

```
cls = GaussianNB()
cls.fit(X, y)
cls.sigma_[:, 0] = 0.26
y_pred = cls.predict(X)
print("accurecy: ", accuracy_score(y, y_pred))
x1 = X[:, 0][y == 0]
sigma1 = np.std(x1)
mean1 = np.mean(x1)
x2 = X[:, 0][y == 1]
sigma2 = np.std(x2)
```

```
mean2 = np.mean(x2)
x3 = X[:, 0][y == 2]
sigma3 = np.std(x3)
mean3 = np.mean(x3)
x=np.linspace(1, 10, 100)
plt.plot(x, stats.norm.pdf(x, mean1, 0.26), label=iris_ds.tar-
get_names[0])
plt.plot(x, stats.norm.pdf(x, mean2, 0.26), label=iris_ds.tar-
get_names[1])
plt.plot(x, stats.norm.pdf(x, mean3, 0.26), label=iris_ds.tar-
get_names[2])
plt.xlabel('Sepal length')
plt.ylabel('Likelihood')
plt.legend()
plt.show()

plt.plot(cls.predict_proba(X)[:, 0], label=iris_ds.target_names[0])
plt.plot(cls.predict_proba(X)[:, 1], label=iris_ds.target_names[1])
plt.plot(cls.predict_proba(X)[:, 2], label=iris_ds.target_names[2])
plt.legend()
```

Figure 5: Likelihoods plot



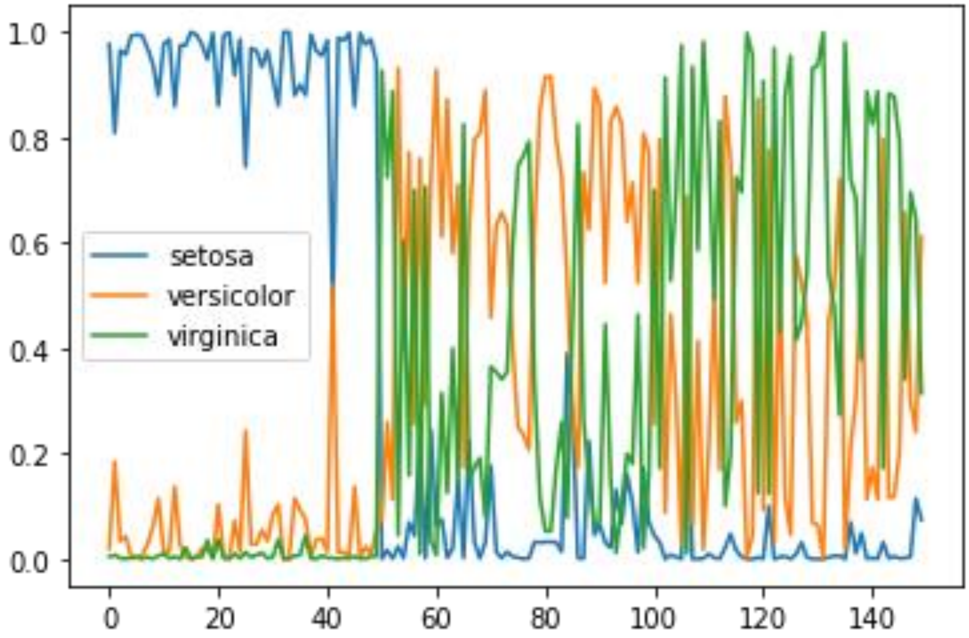


Figure 6: Posterior Probabilities

`accurecy:  0.8`

- In the first case where the means are equal, we obtained a low accuracy value of 62.6%.
  - The reason behind this can be inferred from the posterior probability plot as the three classes appear to be highly overlapping.
  - Also, as we can see from the likelihoods plot that the overlapping of the three classes makes prediction very difficult and increases the error since the three classes have the same mean.

- In contrast to the first case, in the second case where the variances are equal, we got a high accuracy value of 80%.
  - The reason behind that can be inferred from the posterior probability plot as the three classes appear to be well separated.
  - Also, as we can see from the likelihoods plot that the overlapping of the three classes is very small and they are well separated which makes the error much lower than the first case since the three classes have different means.