Problem 3: Some practice on naïve Bayes classifier (continuous attributes)

Pertinent readings for Problem #3:

Pictorial view of naïve Bayes classifier for 2 attributes

https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote05.html

Kubat: Introduction to machine learning (2nd ed)

Blackboard location: /Resources / Our main textbooks for this class

Ch. 2: Pages 30 - 40 (naïve Bayes classifier, continuous variables)

Page 38: The big example that you should really look at ! =) ★★★

Your homework tasks: (turn in the parts highlighted in yellow

Part 1: Practice native Bayes on our good ol' iris data

Let's revisit Ronald Fisher's iris data, where you have 150 irises that can be classified as 3 iris types (Attribute #5):

$\overrightarrow{x_1}$ **Sepal length**	$\overrightarrow{x_2}$ **Sepal width**	$\overrightarrow{x_3}$ **Petal length**	$\overrightarrow{x_4}$ **Petal width**	$\overrightarrow{x_1}$ **Iris type**
Continuous	Continuous	Continuous	Continuous	3 classes Iris sertosa Iris versicolor Iris virginica

For this problem, I only want to consider attributes $\overrightarrow{x_1}$ *and* $\overrightarrow{x_4}$

3 output classes

C_1, C_2 and C_3

Table 1: The iris data set that we're all familiar with !

And as always, we aim to code a machine learning (ML) box so that:

1) Given a new iris sample with input attributes x_1, and x_4

2) Our ML box can try to predict whether our new iris was either an:

   a) *Iris sertosa* (numerical output value = 1 , Class C_1), or
   b) *Iris versicolor* (numerical output value = 2 , Class C_2), or
   c) *Iris virginica* (numerical output value = 3 , Class C_3)

The overall scheme is depicted in Figure 2 below:

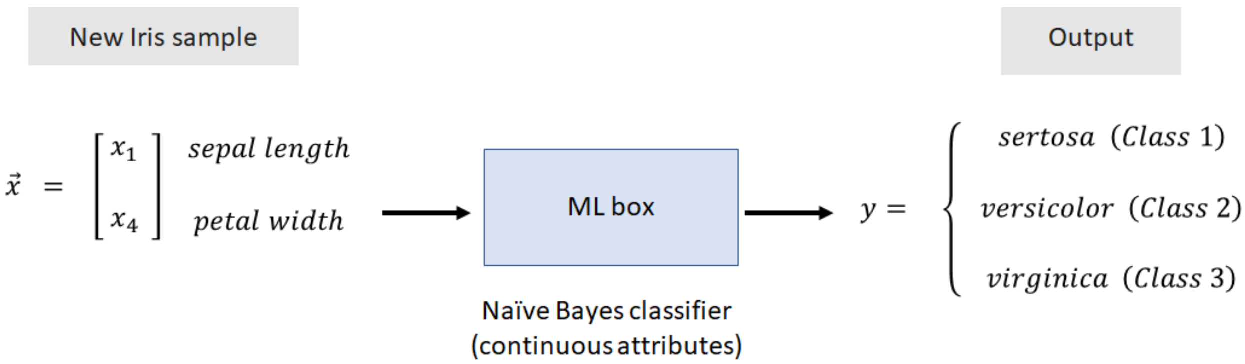

Figure 2: The ML box that we're gonna revisit in Part 1 of this problem =)

Your tasks for Part A: Previewing + reading in data files

1) Using matlab's *preview* and *readtable* functions , read in the iristraining data file called "iris_data.csv."

```
% -- Preview the file using the "detectImportOptions" command

opts = detectImportOptions('iris_dataset.csv', 'NumHeaderLines', 1);
preview(' iris_dataset.csv', opts)

input('This is a preview of the CSV file..  press enter to continue !')


% -- Now, we will read in the table for real !

A = readtable('iris_dataset.csv', 'NumHeaderLines', 1);
```

Part B: Plotting the probability *contour maps* for all 3 iris classes

Since there are 3 classes of irises, you know your ML box will ask "3 big questions" regarding the probability of occurrence for each class:

$$\textit{Class 1 question:}\quad P(\ C_1\ |\ \vec{x}\ )\ =\ \frac{P(\ x_1\ |\ C_1)\ \cdot\ \ P(\ x_4\ |\ C_1)}{P(\vec{x})}\cdot P(C_1)$$

$$=\ \frac{\left(\sum\ \begin{matrix}\textit{Gaussians}\\ \textit{along } x_1\ \textit{axis}\\ \textit{for class 1 points}\end{matrix}\right)\cdot\ \left(\sum\ \begin{matrix}\textit{Gaussians}\\ \textit{along } x_4\ \textit{axis}\\ \textit{for class 1 points}\end{matrix}\right)}{P(\vec{x})}\cdot P(C_1)$$

$$P(\ C_1\ |\ \vec{x}\ )\ \ =\ \frac{\begin{pmatrix}\textit{meshgrid worth of Gaussians}\\ \textit{for class 1 points}\\ \\ \end{pmatrix}}{P(\vec{x})}\cdot P(C_1)$$

$$\textit{Class 2 question:}\ P(\ C_2\ |\ \vec{x}\ )\ =\ \frac{P(\ x_1\ |\ C_2)\ \cdot\ P(\ x_4\ |\ C_2)}{P(\vec{x})}\cdot P(C_2)$$

$$P(\ C_2\ |\ \vec{x}\ )\ \ =\ \frac{\left(\sum\ \begin{matrix}\textit{Gaussians}\\ \textit{along } x_1\ \textit{axis}\\ \textit{for class 2 points}\end{matrix}\right)\cdot\ \left(\sum\ \begin{matrix}\textit{Gaussians}\\ \textit{along } x_4\ \textit{axis}\\ \textit{for class 2 points}\end{matrix}\right)}{P(\vec{x})}\cdot P(C_2)$$

$$\textit{Class 3 question:}\ P(\ C_3\ |\ \vec{x}\ )\ =\ \frac{P(\ x_1\ |\ C_3)\ \cdot\ P(\ x_4\ |\ C_3)}{P(\vec{x})}\cdot P(C_3)$$

$$:\ P(\ C_3\ |\ \vec{x}\ )\ \ =\ \frac{\left(\sum\ \begin{matrix}\textit{Gaussians}\\ \textit{along } x_1\ \textit{axis}\\ \textit{for class 3 points}\end{matrix}\right)\cdot\ \left(\sum\ \begin{matrix}\textit{Gaussians}\\ \textit{along } x_4\ \textit{axis}\\ \textit{for class 3 points}\end{matrix}\right)}{P(\vec{x})}\cdot P(C_3)$$

Your next task in Part B

1) Using matlab, on the $\vec{x_4}$ versus $\vec{x_1}$ plane, plot all 150 iris data points, where you:

   a) Split up the 3 iris classes, and
   b) Plot them in different colors

Have your axes limits as:

$$axis([\ xmin\ xmax\ \ ymin\ \ ymax])\ ,\ \ where\ \begin{cases} xmin\ =\ \ 0,\ \ xmax = 10 \\ ymin\ =\ \ 0,\ \ ymax = 10 \end{cases}$$

2) Then, build the probabilities for each of your "3 big questions" ! Remember: We only need to calculate the gray-shaded numerator terms (and don't have to worry about the denominator terms)

$$Class\ 1\ question:\ \ P(\ C_1\ |\ \vec{x}\ )\ =\ \frac{P(\ x_1\,|\ C_1)\ \cdot\ P(\ x_4\,|\ C_1)}{P(\vec{x})}\cdot P(C_1)\ \ \rightarrow\ \ Will\ generate\ 1\ set\ of\ cotnours$$

$$Class\ 2\ question:\ P(\ C_2\ |\ \vec{x}\ )\ =\ \frac{P(\ x_1\,|\ C_2)\ \cdot\ P(\ x_4\,|\ C_2)}{P(\vec{x})}\cdot P(C_2)\ \ \rightarrow\ \ Will\ generate\ the\ 2nd\ set\ of\ cotnours$$

$$Class\ 3\ question:\ P(\ C_3\ |\ \vec{x}\ )\ =\ \frac{P(\ x_1\,|\ C_3)\ \cdot\ P(\ x_4\,|\ C_3)}{P(\vec{x})}\cdot P(C_3)\ \ \rightarrow\ \ Will\ generate\ the\ 3rd\ set\ of\ cotnours$$

** For your σ value in your Gaussians, you can use (one for each attribute $\vec{x_1}$ and $\vec{x_4}$) :

$$\sigma_1\ =\ \sigma_4\ \ =\ \ 0.2$$

3) Finally, overlay the 3 sets of probability contours for each 3 classes of irises ! Your plot should look something like what we did in class… as in Figure 3 below ! =)

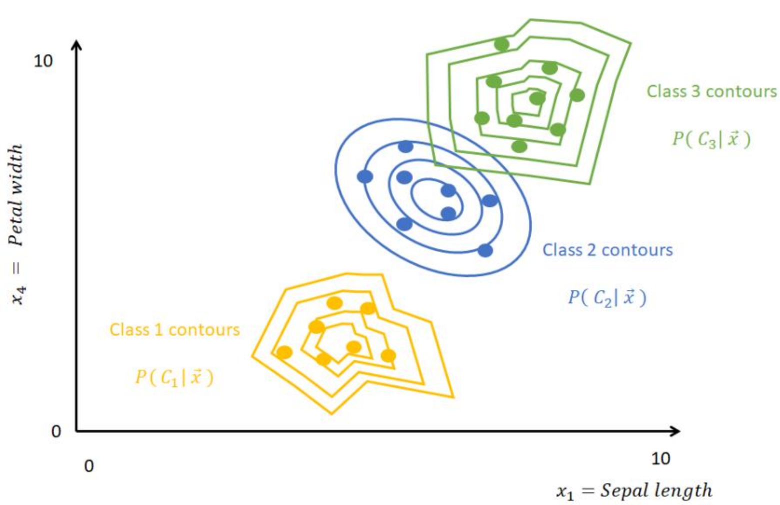

Figure 3: The contour maps that you will generate in task B !

Hints:

1) For your contour plots, you might want to set your contour levels to be something like:

Contour levels = [0 : 0.03 : 0.15]

2) When you're building the contour plot for the 3 big probabilities, you definitely want to use meshgrid variables ! For instance, I started with:

x1 = 0 : 0.05 : 10;
x4 = 0: 0.05 :10;
[X1, X2 = meshgrid(x1, x4)

Then, you can start building the sums / products of your "mini-Gaussians" along x1, x4 axes from this…

… and then build $P(\ C_1 \mid \vec{x}\ )$, $P(\ C_2 \mid \vec{x}\ )$, and $P(\ C_3 \mid \vec{x}\ )$

Part C: Classifying new irises

Now, we're ready to code a super-simple naïve Bayes classifier for your iris ML box !

New Iris sample

$$\vec{x} = \begin{bmatrix} x_1 \\ x_4 \end{bmatrix} \begin{matrix} sepal\ length \\ petal\ width \end{matrix} \longrightarrow \boxed{\text{ML box}} \longrightarrow y = \begin{cases} sertosa\ \ (Class\ 1) \\ versicolor\ \ (Class\ 2) \\ virginica\ \ (Class\ 3) \end{cases}$$

Naïve Bayes classifier
(continuous attributes)

Output

Suppose you were presented with 4 new iris samples:

Attribute	New sample #1	New sample #2	New sample #3	New sample #4
$\overrightarrow{x_1}$ $(sepal\ length)$	5.5	7.0	6.5	6.2
$\overrightarrow{x_4}$ $(petal\ width)$	0.5	1.8	1.5	1.7

Your next task in Part C

1) Using your naïve Bayes classifier, classify these 4 new samples ! For *<u>each of the 4 new irises</u>*, please echo the following:

a) The probabilities of the 3 big question… without the denominator term's contribution (only report the product of the gray-shaded boxes)

$$Class\ 1\ question: \quad P(\ C_1\ |\ \vec{x}\ )\ =\ \frac{P(\ x_1\ |\ C_1)\ \cdot\ P(\ x_4\ |\ C_1)}{P(\vec{x})} \cdot P(C_1)$$

$$Class\ 2\ question: \ P(\ C_2\ |\ \vec{x}\ )\ =\ \frac{P(\ x_1\ |\ C_2)\ \cdot\ P(\ x_4\ |\ C_2)}{P(\vec{x})} \cdot P(C_2)$$

$$Class\ 3\ question: \ P(\ C_3\ |\ \vec{x}\ )\ =\ \frac{P(\ x_1\ |\ C_3)\ \cdot\ P(\ x_4\ |\ C_3)}{P(\vec{x})} \cdot P(C_3)$$

b) Then, I want to know the final classifications for each of the 5 biopsy samples: They're either gonna be malignant (Class C_1)…… or benign (Class C_2) !

c) Overlay the 4 new iris data onto your existing contour plot (and use it as a sanity check for your answers) ! =)