ELE888 Intelligent Systems
Lab 1: Bayesian Decision Theory

Santos Bethany
Department of Electrical, Computer and Biomedical Engineering
Ryerson University
Toronto, ON, Canada

I. Introduction

The fundamental statistical approach towards a problem of classification is called the Bayesian Decision Theory. The theory provides a mathematical structure used for logical reasoning using probability. Bayesian Decision Theory assumes that (1) the decision-making problem can be transformed into probabilistic terms and (2) that all the relevant probabilities are known.

The objective of the lab is to understand and implement the Bayesian decision rule. A training set with known values and classifications will help create a Bayesian Rule algorithm that will then be used by the application to classify a set of new data.

II. Theory

The Bayes formula is described as:

$$P(w_j|x) = \frac{p(x|w_j) * P(w_j)}{p(x)}$$

$$p(x) = \sum_{j=1}^{c} p(x|w_j) * P(w_j)$$

where,

c is the number of classes.
w_j is the class.
x being the observation
P (w_j) is the prior probability of the class
p(x) being the probability density function
P (w_j|x) is the posterior probability of (w_j) given (x).
P (x|w_j) is the class-conditional probability density of (x) given P (x|w_j).
.

III. Methodology

1. To gather some information before starting the lab, we compiled and ran the "runlab1.m" file. Basically, the file preprocessed the IRIS labels to generate a set of numeric labels. The data classified into Iris Setosa and Iris Versicolour. The training sets were obtained from these two classes and then the feature was selected by the input of the user. The prior probabilities were then determined
2. In the next step, the prior probabilities were determined by using the "find and length" in MATLAB
3. Class conditional probabilities for both classes were found by calculating the mean and standard deviations of the data.
4. Posterior probabilities was then determined by computing probability density function (which is the product of the prior probability and the class-conditional probabilities). Finally, the probability density function was used in the Bayes Formula to find the Posterior probabilities.
5. The discriminant function was then computed by taking the difference between the first posterior probability with the second one.
6. As required, x1=[3.3, 4.4, 5.0, 5.7, 6.3] was used as sepal width and sepal length as the feature to determine the posterior probability, class labels and discriminant function values.
7. Optimal Threshold was found eventually in order to classify into two classes.
8. Finally, the code was adjusted to accept the Sepal Length

IV. Results

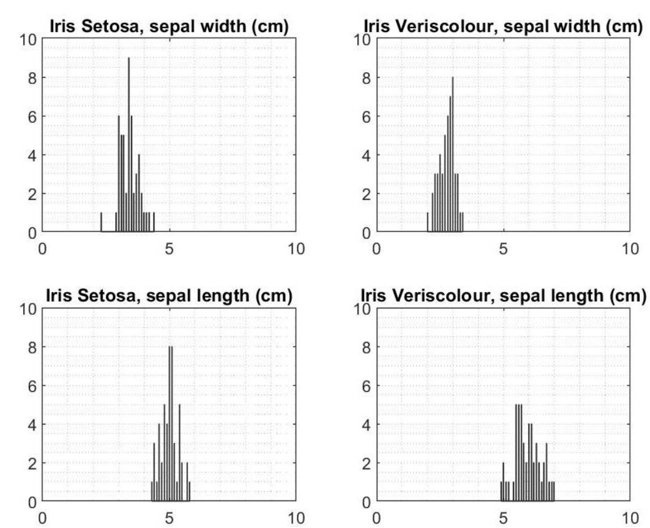

Figure1. Individual Histograms

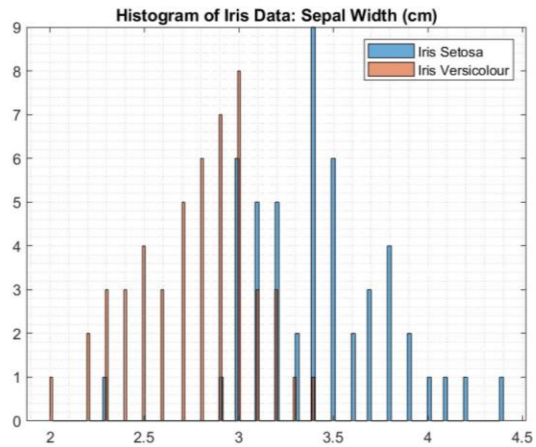

Figure #2: Joint histogram showing sepal width data for Iris Setosa and Iris Versicolour classes

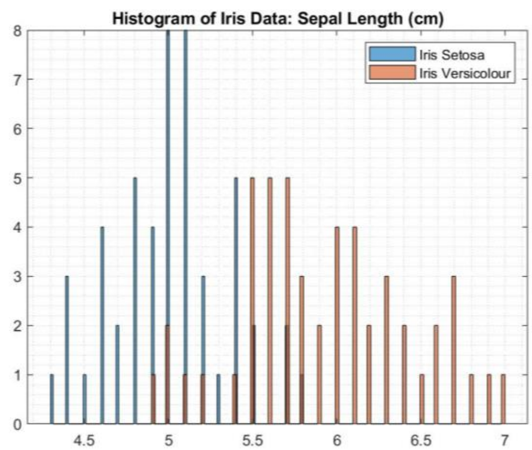

Figure #3: Joint histogram showing sepal length data for Iris Setosa and Iris Versicolour classes

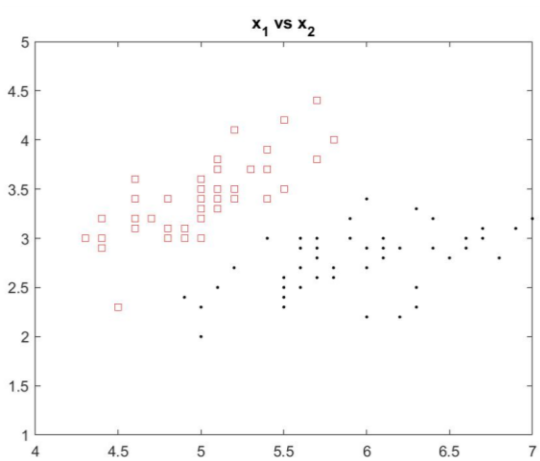

Figure #4: Distribution of feature x1 and x2 for Iris Setosa and Iris Versicolour classes

Identifying the classes and their respective posteriors probabilities and discriminant function values with the input vector x = [3.3, 4.4, 5.0, 5.7, 6.3]

Input value (x)	Posterior g1(x1)	Posterior g2(x1)	Discriminant value g(x1)	Class label Decision
3.3	0.2753e-05	0.0730e-05	2.0232e-06	Iris Setosa
4.4	0.0766	0.0040	0.0726	Iris Setosa
5.0	0.3359	0.0649	0.2710	Iris Setosa
5.7	0.0484	0.3026	-0.2543	Iris Versicolour
6.3	0.0004	0.2620	-0.2616	Iris Versicolour

Input value (x)	Posterior g1(x2)	Posterior g2(x2)	Discriminant value g(x2)	Class label Decision
3.3	0.3080	0.0776	0.2304	Iris Setosa
4.4	0.0117	0.0000	0.0117	Iris Setosa
5.0	0.5835e-04	0.0000	5.8350e-05	Iris Setosa
5.7	0.5253e-08	0.0000	5.2533e-09	Iris Setosa
6.3	0.1219e-12	0.0000	1.2193e-13	Iris Setosa

Decision boundary threshold (Th1) for sepal length feature x1 is approximately 5.38332 cm.
Decision boundary threshold (Th2) for sepal width feature x2 is approximately 3.062

V. Discussion

The decision boundary threshold values were obtained through trial and error until posterior g1(x) is approximately equal to posterior g2(x). This is justified through Bayes Decision rule and a two-case category dichotomizer.

Bayes Decision Rule states:
Decide w1 if g(x) > 0; otherwise decide w2

Since the computation of the single discriminant function g(x) is as follows:

$$g(x) = g1(x) - g2(x)$$

The decision boundary will be a value at which any data greater than or less than will result to a class decision. Therefore theoretically, the threshold value should produce posteriors g1(x) and g2(x) that are equal which should then result in g(x) = 0.

The decision boundary is most often task and cost specific and can change depending on the penalty associated with misclassification. Looking at the joint histogram of sepal width in Figure #2, if a higher penalty is associated with misclassifying w2 as w1 (red area), where w1 = Iris Setosa and w2 = Iris Versicolour, then the decision boundary can be adjusted (ie Th2 = 3.5 cm) such that there is little error with misclassification.

From Figure #3, the feature sepal length has an error probability with approximately equal weight on either side of the threshold. On the other hand, the feature sepal width has an uneven error probability, where it is more likely to misclassify w2 as w1. Therefore, in deciding which feature to use for classification would depend on the task or cost at hand. In a more general sense, it would be preferred to use sepal length as a feature to classify due to equal distribution of error probability however, if let us say a situation is presented where there is a higher penalty incurred when w2 is misclassified as w1, then the sepal width is a more desirable feature for classification.

One of the goals with pattern recognition is to minimize error as much as possible—this can be done by gathering and processing data that can clearly discriminate between classes. Using two features for classification can give a clearer and more discriminating data compared to a single feature. Figure #4 displays better data when classifying w1 and w2 since there is little to no overlapping datapoints and shows a clear area of separation.

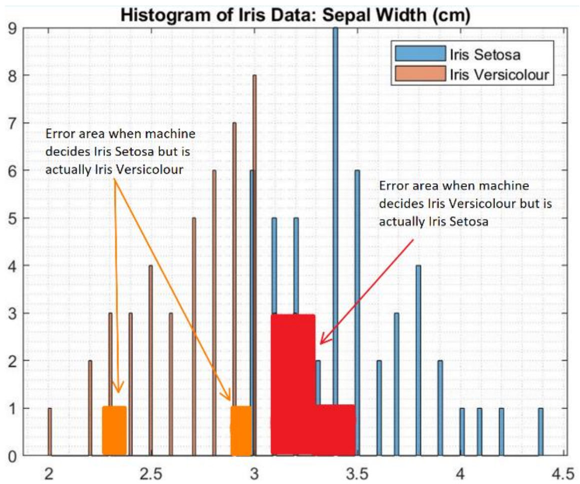

Figure 5. Joint Histograms showing error area

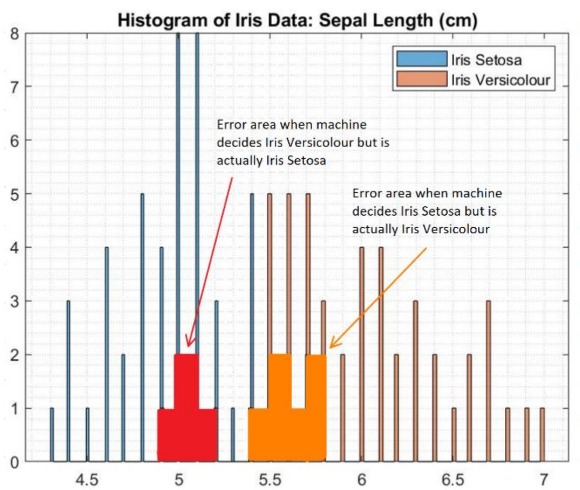

Figure 6. Joint Histograms showing error area

Identifying the classes and their respective posteriors probabilities and discriminant function values with the input vector x = [3.3, 4.4, 5.0, 5.7, 6.3]

VI. CONCLUSION

In conclusion, it was found that the threshold Th1 for feature x1 is approximately 5.38332 cm, and the threshold Th2 for feature x2 is approximately 3.06266 cm. These decision boundaries were theoretically derived using trial and error from Bayes Decision rule and a two-case category dichotomizer. These threshold values are most often task or cost specific and thus allowed to be adjusted depending on the heaviness of the penalties incurred from misclassifying. Lastly, using a single feature to classify can be inadequate and most often hold larger error probabilities, thus it was found that using two features for classification can produce a data set that can be used to better discriminate between the classes.