Department of Systems Design Engineering
University of Waterloo

Lab 2

Model Estimation and Discriminant Functions

Pattern Recognition
SYDE 372

Dan Hendry, 20207096.
Jenny Lu, 20172834

March 17, 2009

Introduction

This lab examines the areas of statistical model estimation and classifier aggregation. Model estimation is performed by implementing parametric and non-parametric estimators. Aggregation is introduced by combining sequences of simple linear discriminants into a single powerful classifier.

The lab is completed using MATLAB.

Discussion

Model Estimation: 1-D Case

The data sets are defined as follows:
Class A: Gaussian distribution, $\mu = 5$, $\sigma = 1$
Class B: Exponential distribution, $\lambda = 1$

Parametric Estimation: Gaussian

Assuming the unknown density is Gaussian, the Maximum Likelihood estimation of the parameters are derived to be

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_i x_i$$

$$\hat{\sigma}_{ML} = \frac{1}{N}\sum_i (x_i - \hat{\mu}_{ML})^2$$

and the resulting $\hat{\mathrm{p}}(x)$ is given by

$$\hat{\mathrm{p}}(x) = \frac{1}{\sqrt{2\pi}\hat{\sigma}} e^{-\frac{(x_i-\hat{\mu})^2}{2\hat{\sigma}^2}}$$

For class A, the estimated parameter values are calculated to be

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_i x_i = 5.0763$$

$$\hat{\sigma}_{ML} = \frac{1}{N}\sum_i (x_i - \hat{\mu}_{ML})^2 = 1.1274$$

In this case, the estimated parameters are, as expected, very similar to the true parameter values $\mu = 5$ and $\sigma = 1$, since the assumption of the form of the density happens to be correct. The resulting estimated $\hat{\mathrm{p}}(x)$ is superimposed on the true $\mathrm{p}(x)$in Figure 1, and the estimated density is very close to the original. Differences in the densities are due to the variation in the sample data points and would be reduced as the number of sample points increases.

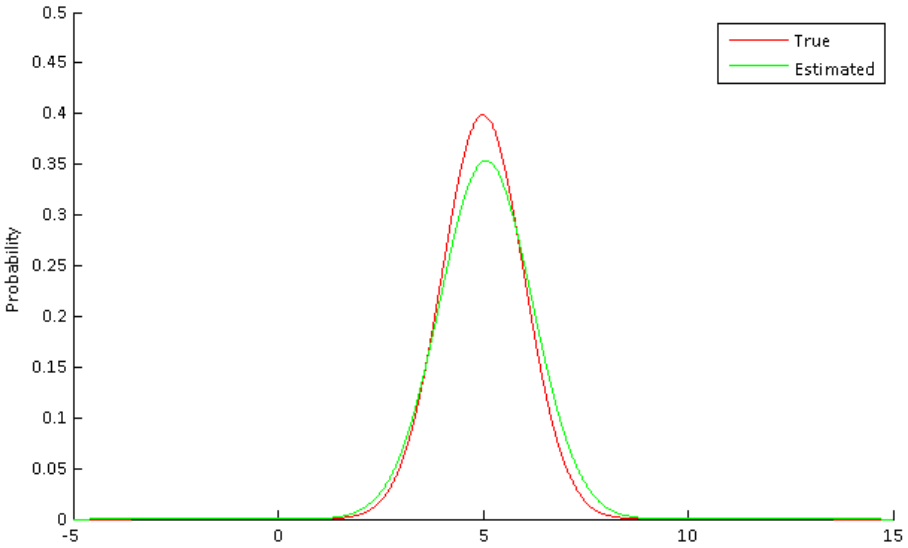

Figure 1: Gaussian parameter estimation of class A

For class B, the estimated parameter values are calculated to be

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{i} x_i = 0.9633$$

$$\hat{\sigma}_{ML} = \frac{1}{N}\sum_{i} (x_i - \hat{\mu}_{ML})^2 = 0.8643$$

The resulting estimated $\hat{\mathrm{p}}(x)$ is superimposed on the true $\mathrm{p}(x)$in Figure 2. Since the assumption of the form of the density happens to be incorrect, the estimated density is not close to the original, especially for small values of x. As x goes to infinity, both the Gaussian (estimated) and exponential (true) densities approach zero, hence for larger values of x the estimated density is closer to the original.

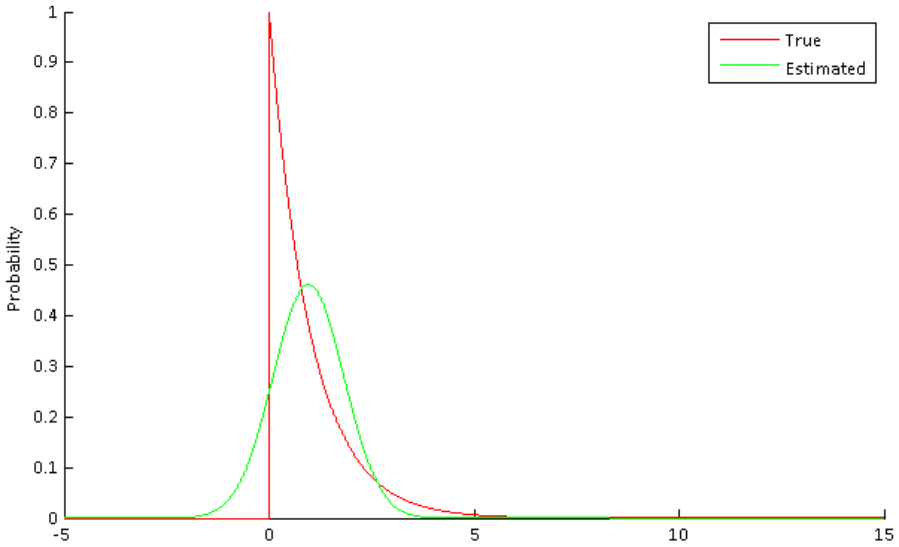

Figure 2: Gaussian parameter estimation of class B

Parametric Estimation: Exponential

Assuming the unknown density is exponential, the Maximum Likelihood estimation of the parameter is derived to be

$$\hat{\lambda}_{ML} = \frac{N}{\sum_i x_i}$$

and the resulting $\hat{\mathrm{p}}(x)$ is given by

$$\hat{\mathrm{p}}(x) = \hat{\lambda} e^{-\hat{\lambda} x}$$

For class A, the estimated parameter value is calculated to be

$$\hat{\lambda}_{ML} = \frac{N}{\sum_i x_i} = 0.1970$$

The resulting estimated $\hat{\mathrm{p}}(x)$ is superimposed on the true $\mathrm{p}(x)$in Figure 3. Since the assumption of the form of the density happens to be incorrect, the estimated density is not close to the original, especially for small values of x and x close to $\mu = 5$. As x goes to infinity, both the exponential (estimated) and Gaussian (true) densities approach zero, hence for much larger values of x the estimated density is closer to the original.

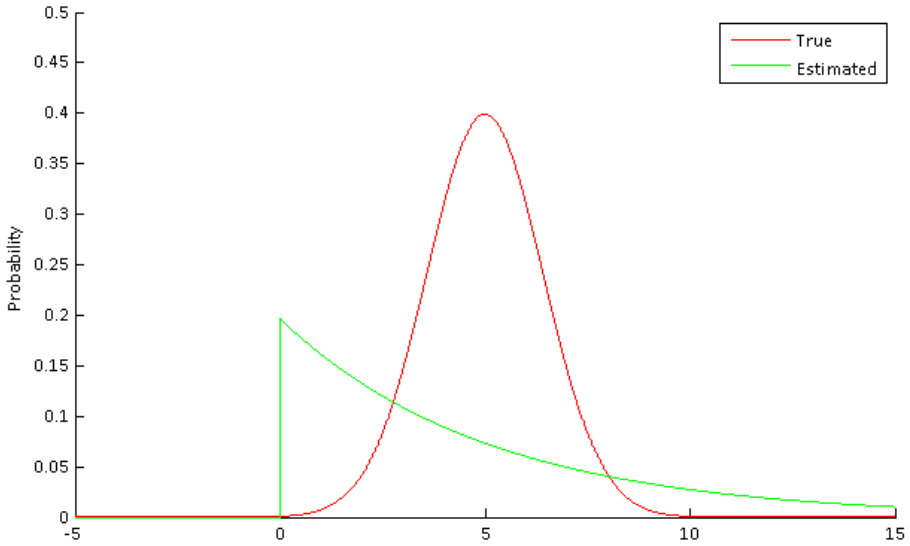

Figure 3: Exponential parameter estimation of class A

For class B, the estimated parameter value is calculated to be

$$\hat{\lambda}_{ML} = \frac{N}{\sum_i x_i} = 1.0381$$

In this case, estimated parameter is, as expected, very similar to the true parameter values $\lambda = 1$, since the assumption of the form of the density happens to be correct. The resulting estimated $\hat{\mathrm{p}}(x)$ is superimposed on the true $\mathrm{p}(x)$in Figure 4, and the estimated density is very close to the original. Differences in the densities are due to the variation in the sample data points and would be reduced as the number of sample points increases.

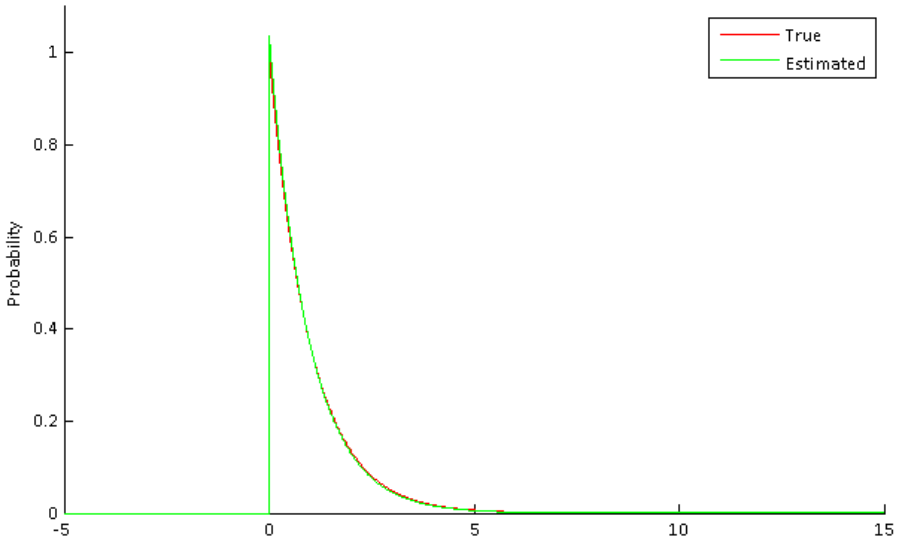

Figure 4: Exponential parameter estimation of class B

Parametric Estimation: Uniform

Assuming the unknown density is uniform, the Maximum Likelihood estimation of the parameter is derived to be

$$\hat{a}_{ML} = \min(x_i)$$
$$\hat{b}_{ML} = \max(x_i)$$

and the resulting $\hat{\mathrm{p}}(x)$ is given by

$$\hat{\mathrm{p}}(x) = \begin{cases} \dfrac{1}{b-a}, x \in [a,b] \\ \\ 0, \text{otherwise} \end{cases}$$

For class A, the estimated parameter values are calculated to be

$$\hat{a}_{ML} = \min(x_i) = 2.7406$$
$$\hat{b}_{ML} = \max(x_i) = 8.3079$$

The resulting estimated $\hat{\mathrm{p}}(x)$ is superimposed on the true $\mathrm{p}(x)$in Figure 5. Since the assumption of the form of the density happens to be incorrect, the estimated density is not close to the original, especially for values of x close to $\mu = 5$. As x goes to infinity, $\hat{\mathrm{p}}(x) = 0$ for $x \notin [a,b]$ for the uniform (estimated) density as the Gaussian (true) density also approaches zero, hence for values of x far away from $\mu = 5$ the estimated density is closer to the original.

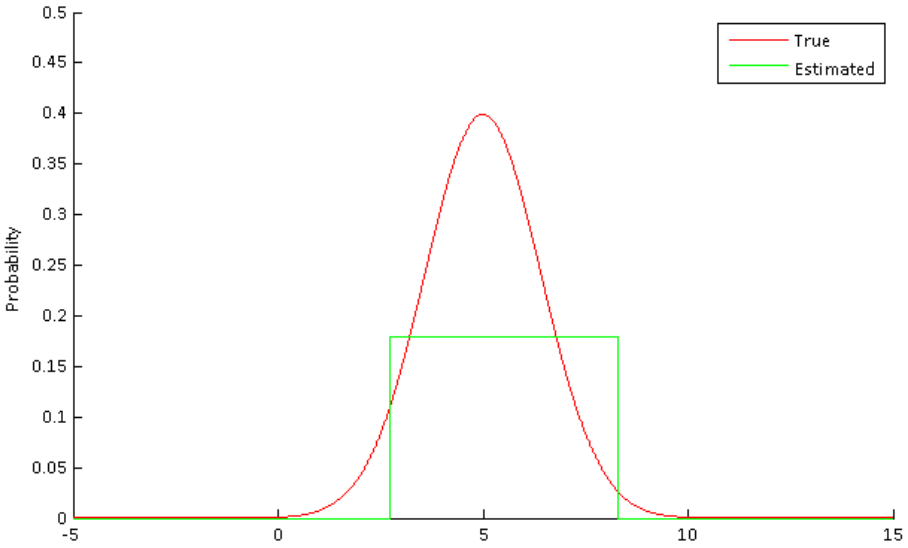

Figure 5: Uniform parameter estimation of class A

For class B, the estimated parameter values are calculated to be

$$\hat{a}_{ML} = \min(x_i) = 0.0143$$
$$\hat{b}_{ML} = \max(x_i) = 4.2802$$

The resulting estimated $\hat{\mathrm{p}}(x)$ is superimposed on the true $\mathrm{p}(x)$in Figure 6. Similarly to class A, since the assumption of the form of the density happens to be incorrect, the estimated density is not close to the original, especially for small values of x. As x goes to infinity for $x > b$, $\hat{\mathrm{p}}(x) = 0$ for $x \notin [a, b]$ for the uniform (estimated) density as the exponential (true) density also approaches zero, hence for larger values of x the estimated density is closer to the original. For $x < a$ and $x < 0$, $\hat{\mathrm{p}}(x) = \mathrm{p}(x) = 0$, and the estimated and true densities are the same.

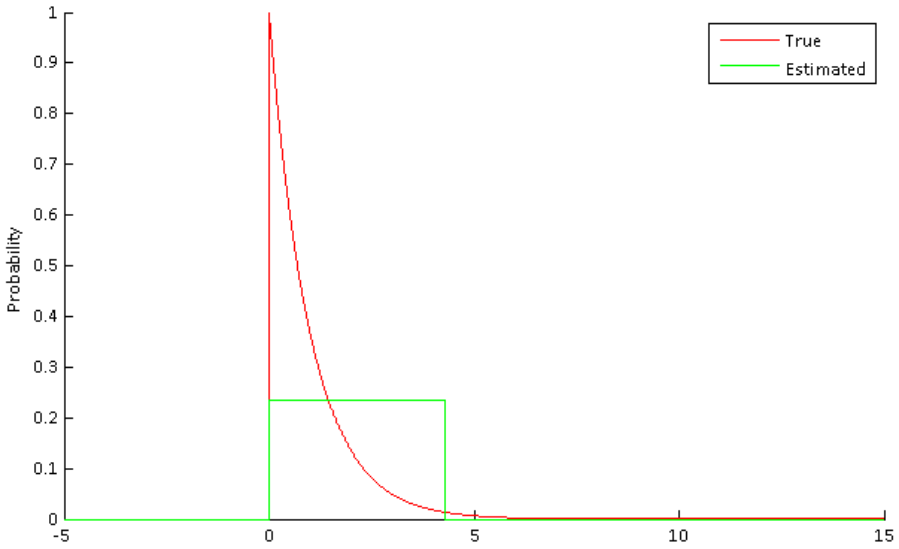

Figure 6: Uniform parameter estimation of class B

Non-Parametric Estimation

Using the parzen window approach, each sample x_i contributes locally to the estimated probability density, and the resulting $\hat{\mathrm{p}}(x)$ is given by

$$\hat{p}(x) = \frac{1}{N}\sum_{i} \phi(x - x_i)$$

For Gaussian parzen windows, the resulting $\hat{p}(x)$ is given by

$$\hat{p}(x) = \frac{1}{N}\sum_{i} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-x_i)^2}{2\sigma^2}}$$

Using Gaussian parzen windows having standard deviations $\sigma = 0.1$ and $\sigma = 0.4$, the density of class A is estimated and superimposed on the true density in Figure 7. The estimated density using $\sigma = 0.4$ is a closer approximation, since the increased standard deviation in effect stretches out each parzen window, hence contributing to a smoother overall result that more closely matches the true Gaussian density. The estimated density using $\sigma = 0.1$ is much spikier, and this is caused by the variation in the sample data points.

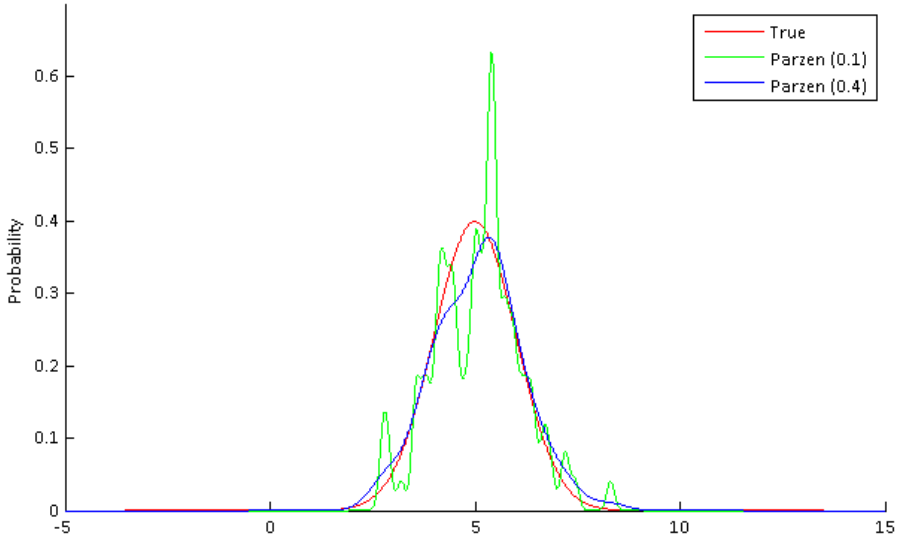

Figure 7: Non-parametric estimation of Class A using Gaussian parzen windows with $\sigma = 0.1$ and $\sigma = 0.4$

Using Gaussian parzen windows having standard deviations $\sigma = 0.1$ and $\sigma = 0.4$, the density of class B is estimated and superimposed on the true density in Figure 8. The estimated density using $\sigma = 0.4$ gives a smoother overall result, but the estimated density using $\sigma = 0.1$ is a closer approximation. Since the increased standard deviation in effect stretches out each parzen window, the narrow peak at $x = 0$ in the true exponential density becomes less defined.

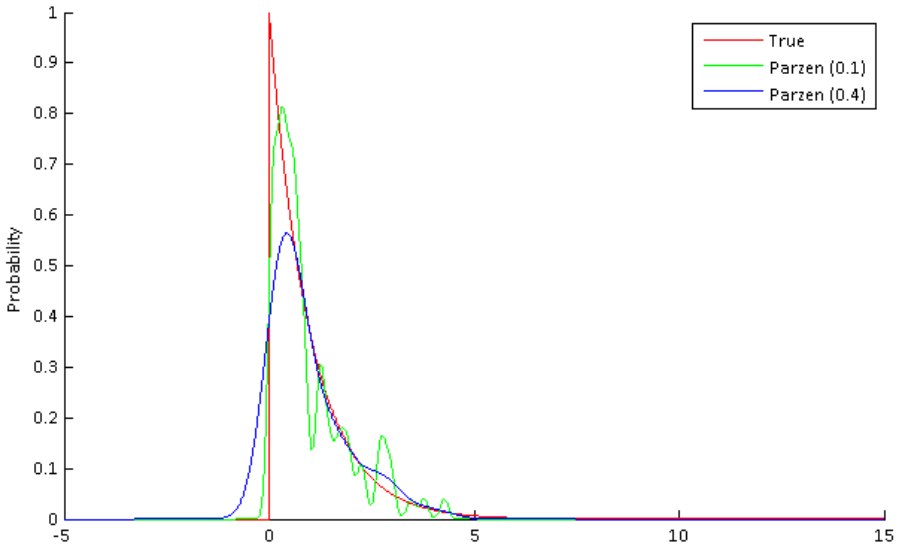

Figure 8: Non-parametric estimation of Class B using Gaussian parzen windows with $\sigma = 0.1$ and $\sigma = 0.4$

General Comments

For class A, the estimated Gaussian density is, as expected, closest to the original, since the assumed form of the density is correct. The next closest is the non-parametric estimated density, especially when a standard deviation of $\sigma = 0.4$ is used for the Gaussian parzen windows to stretch out each window and contribute to a smoother overall result. The estimated exponential and uniform densities are not close to the original, especially for values of x close to $\mu = 5$.

For class B, the estimated exponential density is, as expected, closest to the original, since the assumed form of the density is correct. The next closest is the non-parametric estimated density, especially when a standard deviation of $\sigma = 0.1$ is used for the Gaussian parzen windows to preserve the sharp peak at $x = 0$. The estimated exponential and uniform densities are not close to the original, especially for small values of x.

In general, a parametric estimation is the closest to the original only if the assumed form of the density is correct. A non-parametric estimation is next closest, and works in all cases since no assumptions are made about the form of the density. However, a non-parametric estimation is slower and also more memory-intensive. Therefore, it is better to use a parametric estimation if one is relatively confident about the form of the density. If this is not the case and computation and memory efficiency is not a constraint, it is better to use a non-parametric estimation.

Model Estimation: 2-D Case

Parametric Estimation

Assuming the unknown density is Gaussian, the Maximum Likelihood estimation of the parameters are derived to be

$$\overrightarrow{\hat{\mu}_{ML}} = \frac{1}{N}\sum_i \overrightarrow{x_i}$$

$$\hat{\Sigma}_{ML} = \frac{1}{N}\sum_i \left(\overrightarrow{x_i} - \overrightarrow{\hat{\mu}_{ML}}\right)\left(\overrightarrow{x_i} - \overrightarrow{\hat{\mu}_{ML}}\right)^{\mathrm{T}}$$

and the resulting $\hat{\mathrm{p}}(x)$ is

$$\hat{p}(\vec{x}) = \frac{1}{2\pi|\Sigma_{ML}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\vec{x} - \overrightarrow{\hat{\mu}_{ML}})^T \Sigma_{ML}^{-1}(\vec{x} - \overrightarrow{\hat{\mu}_{ML}})\right)$$

The estimated parameter values for classes A, B, and C are calculated to be

$$\hat{\mu}_A = \begin{bmatrix}347.16\\131.20\end{bmatrix} \quad \hat{\Sigma}_A = \begin{bmatrix}1749 & -1595\\-1595 & 3310\end{bmatrix}$$

$$\hat{\mu}_B = \begin{bmatrix}291.84\\224.02\end{bmatrix} \quad \hat{\Sigma}_B = \begin{bmatrix}3283 & 1164\\1164 & 3380\end{bmatrix}$$

$$\hat{\mu}_C = \begin{bmatrix}119.55\\346.67\end{bmatrix} \quad \hat{\Sigma}_C = \begin{bmatrix}2711 & -1314\\-1314 & 1682\end{bmatrix}$$

The ML classifies a point x as c_i by maximizing the following probability function for classes $i = 1 \dots n$ with Gaussian distributions,

$$P(\vec{x}) = \hat{P}(\vec{x}|c_i) = \frac{1}{2\pi|\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\vec{x} - \overrightarrow{z_i})^T \Sigma_i^{-1}(\vec{x} - \overrightarrow{z_i})\right)$$

The resulting classification boundary is plotted in Figure 9.

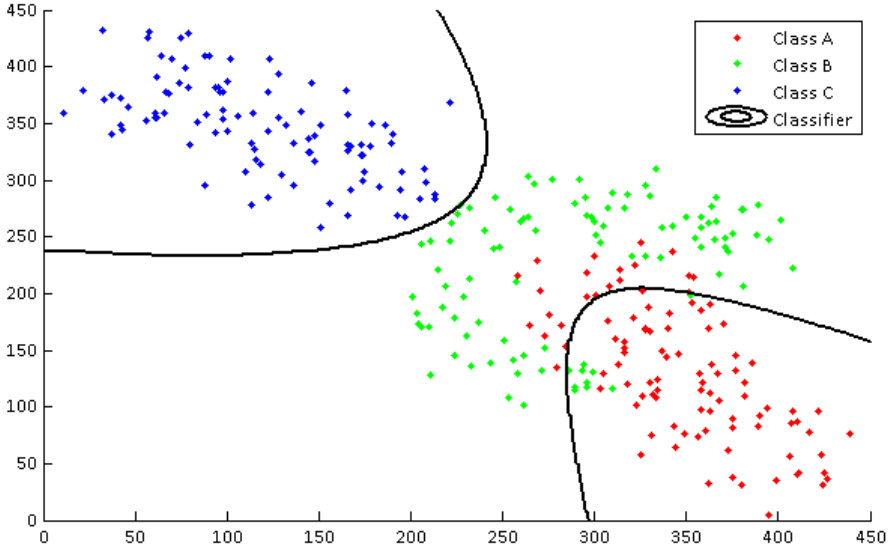

Figure 9: ML classification boundaries of 2-D Gaussian parametric estimation of classes A, B, and C

The classifier does not perform well in classifying class A from class B, because class B deviates from Gaussian behavior in the direction of A. By assuming the form of the density of class B to be Gaussian, the hollow in class B that overlaps with class A cannot be taken into account, and data points on either side of the hollow cause the estimated mean to be shifted towards class A despite the existence of the hollow. Hence many data points in class A that exist in the space of the hollow are instead classified as class B. Since the form of the density of class B does not correspond to any of the standard distributions (Gaussian, exponential, uniform, etc.), it is very difficult to derive an estimated density that is close to the original based on parametric methods.

Non-Parametric Estimation

For 2-D Gaussian parzen windows, the resulting $\hat{p}(x)$ is given by

$$\hat{p}(\vec{x}) = \frac{1}{N}\sum_i \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\vec{x} - \overrightarrow{x_i})^T \Sigma^{-1}(\vec{x} - \overrightarrow{x_i})\right)$$

Using Gaussian parzen windows having $\sigma^2 = 400$, the covariance is given by

$$\Sigma = \begin{bmatrix} 400 & 0 \\ 0 & 400 \end{bmatrix}$$

The ML classifies a point x as c_i by maximizing $\hat{P}(\vec{x}|c_i)$ for classes $i = 1 \ldots n$, and the resulting classification boundary is plotted in Figure 10.

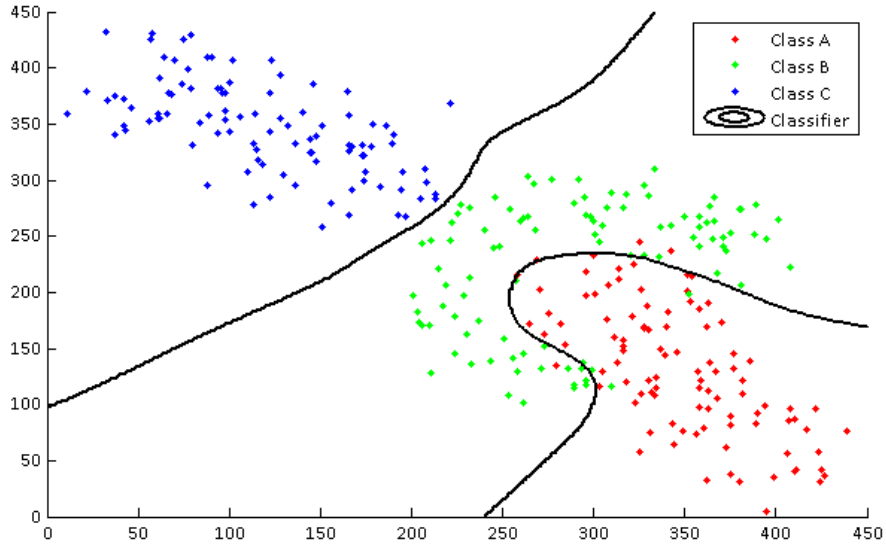

Figure 10: ML classification boundaries of non-parametric estimation of classes A, B, and C using 2-D Gaussian parzen windows with $\sigma^2 = 400$

The classifier performs very well since no assumptions are made about the form of the density for any of the classes. Hence the non-standard form of the density of class B is accounted for, and there are only a few errors in classifying class A from class B near the boundary of the hollow where they overlap.

General Comments

In general, it is not always possible to use a parametric approach, because a parametric approach cannot yield good results if the one of the densities to be estimated is not of a standard form. In such cases, assuming a standard form and estimating the parameters by maximum likelihood does not account for the ways in which the shape of the density deviates from the assumed standard form, and the probability of error in such regions is very high.

It is better to use a parametric approach if one is relatively confident that all densities are of a standard form and about the form of each density. In such cases the parametric approach would yield good results and is also memory and computation efficient. A non-parametric approach is slower and more memory-intensive, but it is better to use a non-parametric estimation if it is known that one or more of the densities is not of a standard form and/or one cannot be confident about the form of each density.

Sequential Discriminants

Three sequential classifiers are learned for classes A and B, and the classification boundaries are plotted with data points in Figure 11, Figure 12, and Figure 13.

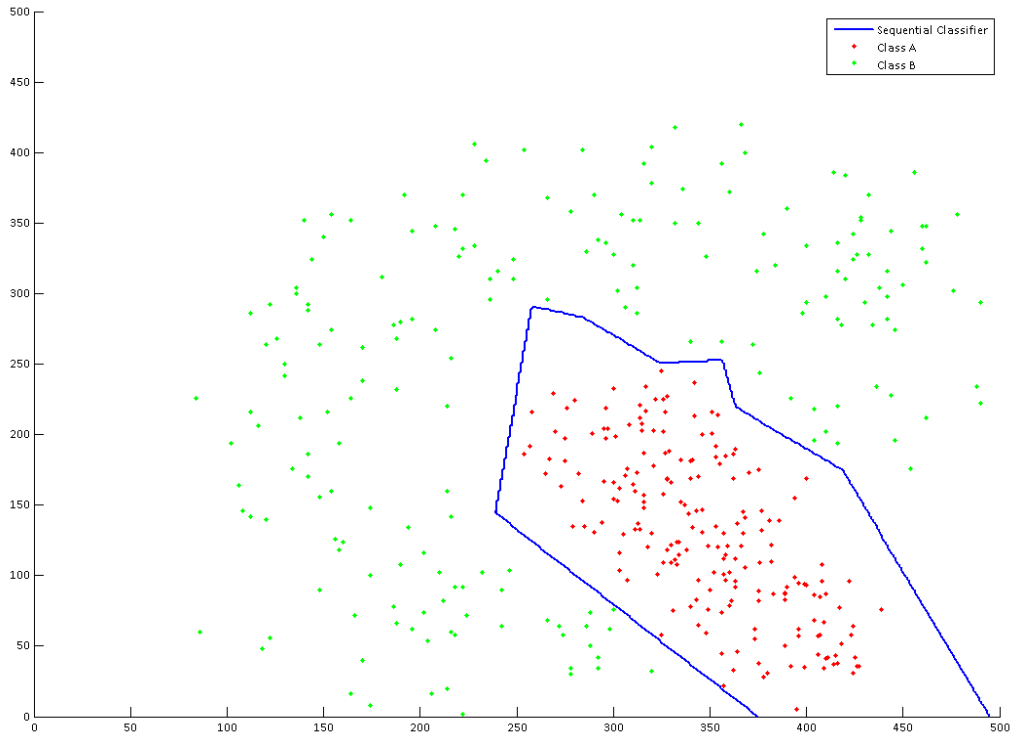

Figure 11: Sequential classifier #1

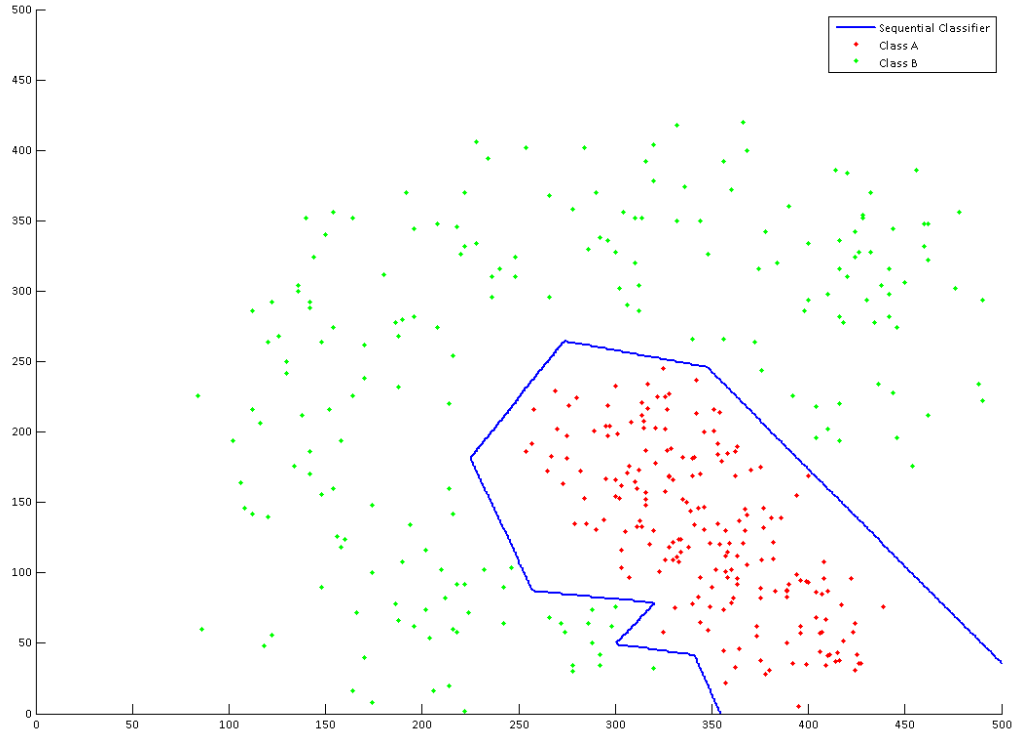

Figure 12: Sequential classifier #2

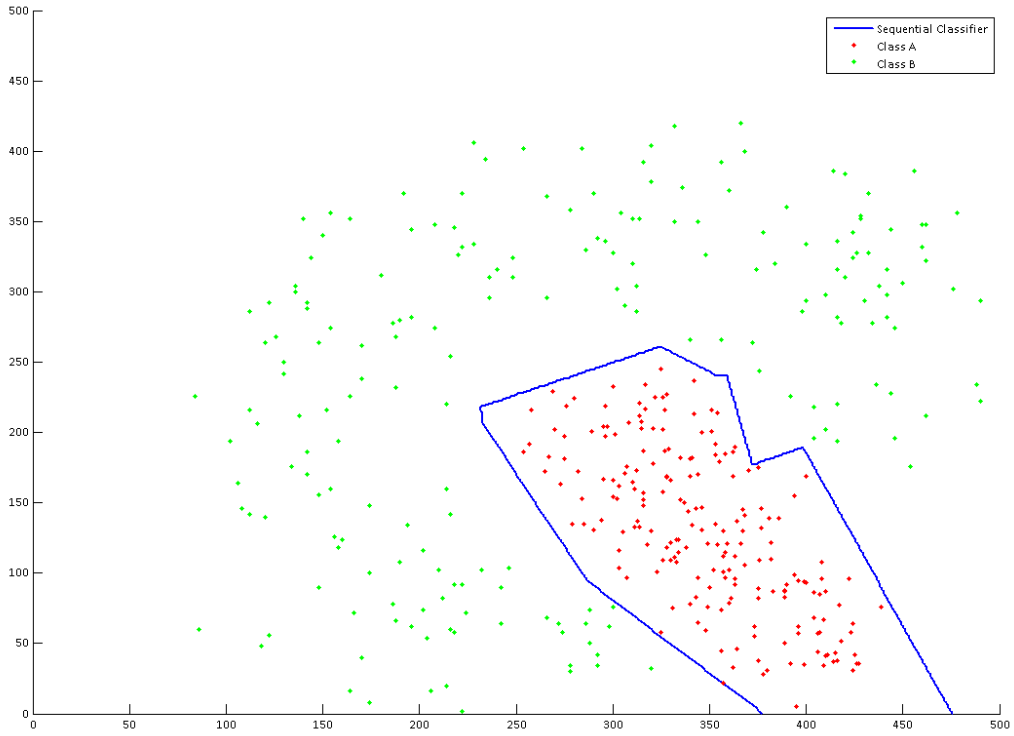

Figure 13: Sequential classifier #3

The sequential classifiers are each composed of a sequence of discriminants, and each individual discriminant classifies some part of the problem perfectly. For any potential discriminant, if it misclassifies any point in some part of the training data, it would be discarded during training. Therefore, according to the rules used to compose the overall sequential classifier, it is impossible for the classifier to classify any point used during training incorrectly. Hence if the classifiers are tested on the training data, the probability of error would be zero.

Limiting the sequential classifiers to J discriminants and, for each limitation, learning a classifier 20 times, the average, standard deviation, maximum, and minimum probability of error is calculated and tabulated in Table 1 for values of J between 1 and 5.

Table 1: Probability of error for sequential classifiers subject to limit $J = 1 \ldots 5$

J	Probability of Error			
	Average	Standard Deviation	Maximum	Minimum
1	0.2984	0.0386	0.3700	0.2400
2	0.1452	0.0441	0.2375	0.0725
3	0.0565	0.0363	0.1775	0.0125
4	0.0764	0.0971	0.3475	0.0050
5	0.0919	0.1216	0.3700	0.0025

The results are plotted in Figure 14.

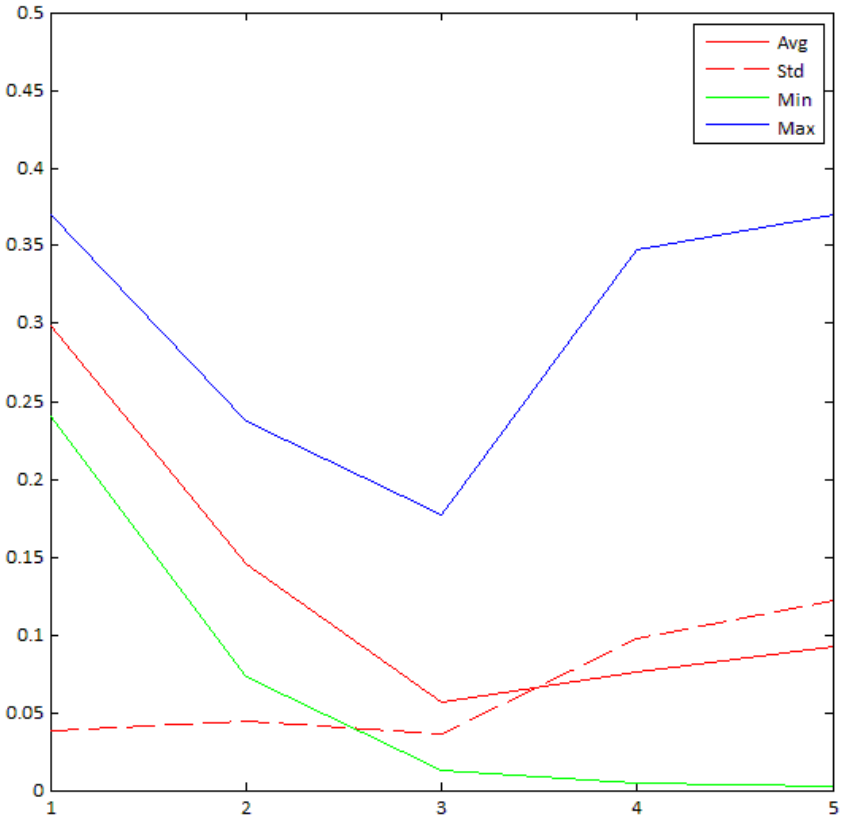

Figure 14: Probability of error for sequential classifiers subject to limit $J = 1 \ldots 5$

Intuitively, the probability of error is expected to decrease as the limit on the value of J increases, since for $J = 1$, only a single discriminant can be used, and the data is not linearly separable, resulting in a relatively high probability of error, but for unconstrained values of J, the probability of error is zero as previously discussed. Indeed, the minimum probability of error decreases as expected; in fact, the minimum probability of error for $J = 5$ is approaching zero at only 0.25%. However, the maximum probability of error is minimized at $J = 3$ but increases again for $J = 4$ and $J = 5$. This in turn causes a slight increase in the average probability of error from $J = 3$ to $J = 5$, and a more significant increase in the standard deviation of the probability of error as well.

The increase in the worst case probability of error for $J = 4$ and $J = 5$ may be explained by considering class shape and examining the error rate for each class. Class A is surrounded by class B and as such, the first few discriminants will always classify class A correctly. By the algorithm used to train the classifier, portions of class B which are correctly classified are removed. Hence the number of points which have the potential to be misclassified is reduced resulting in lower error rates for $J = 1$ to $J = 3$ when additional discriminants are added. Throughout this process, the number of points in class A remains the same. As the limit on the value of J increases, new classifiers may be able to classify a portion of class B perfectly but could misclassify part of class A, an example of which is shown in Figure 15. Several cases of this worst case scenario causes the rise in the maximum, average, and standard deviation of the probability of error observed for $J = 4$ and $J = 5$. This rise is temporary and is expected to fall as additional discriminants learn more about the shape of class A.

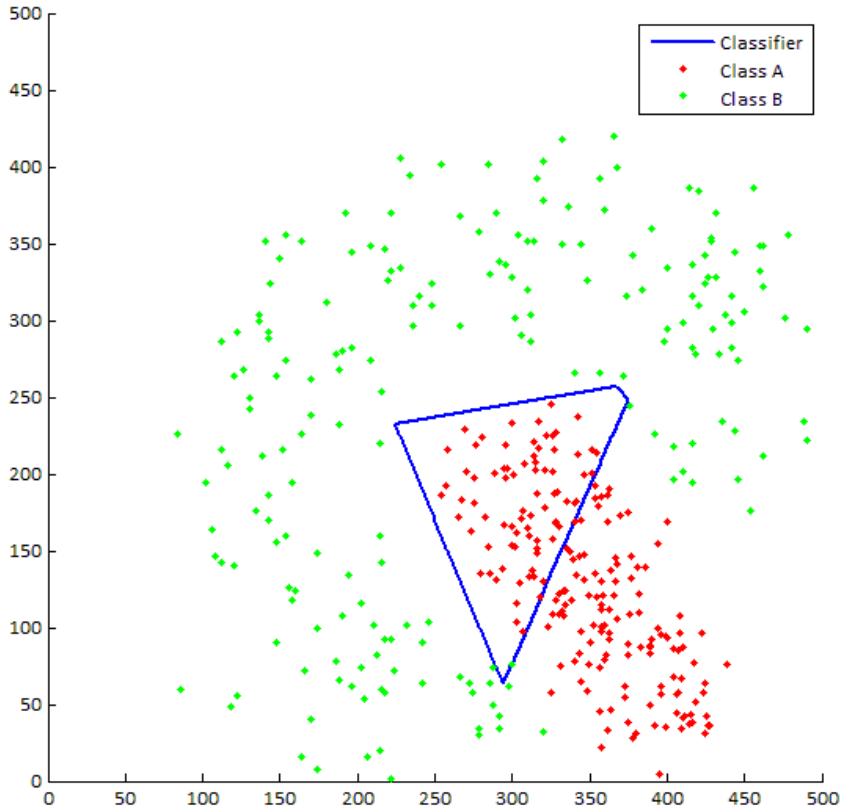

Figure 15: Sequential classifier subject to $J = 4, \mathrm{P}(\epsilon) = 0.3025$

Unconstrained point testing allows a perfect discriminant to be found for a particular class. Placing any limitation on the number of point pairs that can be tested, where the pairs of points are chosen randomly, means that some percentage of discriminants found during training would not be perfect. Use of an imperfect discriminant would result in the misclassification of training data if the same set of data were used for testing.

For certain types of data, it can be assumed that, on average, a classifier which performs better with respect to training data will perform better when classifying other data. For the purposes of this discussion it must be assumed that training data does not contain outliers which belonging to one class yet fall in the true shape of another. In such a case, a perfect discriminant would be undesirable. It may be concluded that a classifier that has a higher probability of error when tested on its training data would tend to have a higher rate of error when used with other sets of testing data, subject to the above mentioned assumptions. Since limiting the number of point pairs that can be tested would increase the probability of error with respect to training data, it would generally also increase the probability of error.

Conclusions

In general, a parametric estimation is the closest to the original only if the assumed form of the density is correct. A non-parametric estimation is next closest, and works in all cases since no assumptions are made about the form of the density. However, a non-parametric estimation is slower and also more memory-intensive. Therefore, it is better to use a parametric estimation if one is relatively confident about the form of the density. However, it is not always possible to use a parametric approach, because a parametric approach cannot yield good results if the one of the densities to be estimated is not of a

standard form. In such cases, assuming a standard form and estimating the parameters by maximum likelihood does not account for the ways in which the shape of the density deviates from the assumed standard form, and the probability of error in such regions is very high.

Sequential classifiers composed of sequences of discriminants, where each individual discriminant classifies some part of the problem perfectly, achieve a probability of error of zero when tested on training data, by definition of the algorithm used to generate the classifiers. However, when the number of discriminants are limited, the probability of error is no longer zero and in general becomes inversely proportional to the number of disciminants allowed. Limiting the number of point pairs from the training data that can be tested would also in general increase the probability of error.