CS645 Homework Solutions, Week 11
Chapter 4, PRML

1. Provide some details to get 4.68.

Solution:

4.68: $a_k(x) = w_k^T x + w_{k0}$

Given:

4.62: $p(C_k \mid \bar{x}) = \dfrac{p(\bar{x} \mid C_k)p(C_k)}{\sum_j p(\bar{x} \mid C_j)p(C_j)} = \dfrac{\exp(a_k)}{\sum_j \exp(a_j)}$

4.63: $a_k = \ln p(\bar{x} \mid C_k)p(C_k)$

Assuming that class conditionals are Gaussian we have:

$$p(\bar{x} \mid C_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\bar{x} - \bar{u}_k)^T \Sigma^{-1} (\bar{x} - \bar{u}_k)\right\}$$

The term in the exponential is equivalent to:

$$-\frac{1}{2}\left\{\bar{x}^T \Sigma^{-1} \bar{x} - 2\bar{x}^T \Sigma^{-1} \bar{u}_k + \bar{u}_k \Sigma^{-1} \bar{u}_k\right\} = -\frac{1}{2}\bar{x}^T \Sigma^{-1} \bar{x} + \bar{x}^T \Sigma^{-1} \bar{u}_k - \frac{1}{2}\bar{u}_k \Sigma^{-1} \bar{u}_k$$

Plugging into 4.62, the constants in front will cancel, giving:

$$p(C_k \mid \bar{x}) = \frac{\exp\left\{-\frac{1}{2}\bar{x}^T \Sigma^{-1} \bar{x} + \bar{x}^T \Sigma^{-1} \bar{u}_k - \frac{1}{2}\bar{u}_k \Sigma^{-1} \bar{u}_k\right\} p(C_k)}{\sum_j \exp\left\{-\frac{1}{2}\bar{x}^T \Sigma^{-1} \bar{x} + \bar{x}^T \Sigma^{-1} \bar{u}_j - \frac{1}{2}\bar{u}_j \Sigma^{-1} \bar{u}_j\right\} p(C_j)}$$

$$= \frac{\exp(-\frac{1}{2}\bar{x}^T \Sigma^{-1} \bar{x}) \exp\left\{\bar{x}^T \Sigma^{-1} \bar{u}_k - \frac{1}{2}\bar{u}_k \Sigma^{-1} \bar{u}_k\right\} p(C_k)}{\sum_j \exp(-\frac{1}{2}\bar{x}^T \Sigma^{-1} \bar{x}) \exp\left\{\bar{x}^T \Sigma^{-1} \bar{u}_j - \frac{1}{2}\bar{u}_j \Sigma^{-1} \bar{u}_j\right\} p(C_j)}$$

$$= \frac{\exp\left\{\bar{x}^T \Sigma^{-1} \bar{u}_k - \frac{1}{2}\bar{u}_k \Sigma^{-1} \bar{u}_k + \ln p(C_k)\right\}}{\sum_j \exp\left\{\bar{x}^T \Sigma^{-1} \bar{u}_j - \frac{1}{2}\bar{u}_j \Sigma^{-1} \bar{u}_j + \ln p(C_j)\right\}}$$

$$= \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

Note that the second order terms involving x cancel out in the division.

Therefore, $a_k(\bar{x}) = \bar{x}^T \Sigma^{-1} \bar{u}_k - \frac{1}{2} \bar{u}_k \Sigma^{-1} \bar{u}_k + \ln p(C_k)$

We can define additional terms so that:

$$\bar{w}_k = \Sigma^{-2} \bar{u}_k$$

$$w_{ko} = -\frac{1}{2} \bar{u}_k \Sigma^{-1} \bar{u}_k + \ln p(C_k)$$

$$a_k = \bar{x}^T \bar{w}_k + w_{k0} = \bar{w}_k^T \bar{x} + w_{k0}$$

Which is 4.68.

■

2. PRML 4.5.

Solution:

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

Expanding this above equation by using $m_2 - m_1 = \mathbf{w^T}(\mathbf{m_2} - \mathbf{m_1})$ and $y_n = \mathbf{w^T x_n}$

$$\frac{\mathbf{w^T}(\mathbf{m_2} - \mathbf{m_1})\mathbf{w^T}(\mathbf{m_2} - \mathbf{m_1})}{\sum_{n \varepsilon C_1} (\mathbf{w^T x_n} - \mathbf{w^T m_1})^\mathbf{2} + \sum_{n \varepsilon C_2} (\mathbf{w^T x_n} - \mathbf{w^T m_2})^\mathbf{2}}$$

$$\frac{\mathbf{w^T}(\mathbf{m_2} - \mathbf{m_1})\mathbf{w^T}(\mathbf{m_2} - \mathbf{m_1})}{\sum_{n \varepsilon C_1} \mathbf{w^T}(\mathbf{x_n} - \mathbf{m_1})\mathbf{w^T}(\mathbf{x_n} - \mathbf{m_1}) + \sum_{n \varepsilon C_2} \mathbf{w^T}(\mathbf{x_n} - \mathbf{m_2})\mathbf{w^T}(\mathbf{x_n} - \mathbf{m_2})}$$

Using the fact that for matrices $(AB)^T = B^T A^T$ and rearranging we have

$$\frac{\mathbf{w^T}(\mathbf{m_2} - \mathbf{m_1})(\mathbf{m_2} - \mathbf{m_1})^\mathbf{T}\mathbf{w}}{\sum_{n \varepsilon C_1} \mathbf{w^T}(\mathbf{x_n} - \mathbf{m_1})(\mathbf{x_n} - \mathbf{m_1})^\mathbf{T}\mathbf{w} + \sum_{n \varepsilon C_2} \mathbf{w^T}(\mathbf{x_n} - \mathbf{m_2})(\mathbf{x_n} - \mathbf{m_2})^\mathbf{T}\mathbf{w}}$$

$$J(\mathbf{w}) = \frac{\mathbf{w^T S_B w}}{\mathbf{w^T S_w w}}$$

where

$$\mathbf{S_B} = (\mathbf{m_2} - \mathbf{m_1})(\mathbf{m_2} - \mathbf{m_1})^\mathbf{T}$$

and

$$\mathbf{S}_w = \sum_{n \varepsilon C_1} (\mathbf{x_n} - \mathbf{m_1})(\mathbf{x_n} - \mathbf{m_1})^\mathbf{T} + \sum_{n \varepsilon C_2} (\mathbf{x_n} - \mathbf{m_2})(\mathbf{x_n} - \mathbf{m_2})^\mathbf{T}$$

■

3. Consider the situation in 4.2.1 where the covariances are not the same, but there are only two classes. Derive an equation for the decision boundary. (As suggested by the text, and figure 4.11, the form of the equation should be quadratic).

Solution:

Suppose that the covariance matrices are not the same for the case of continuous inputs modeled as Gaussians in the 2-class generative model for classification. Then the equation for decision boundary can be found by determining the argument to the logistic function a as follows

$$\begin{aligned}
a &= \ln \frac{P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{P(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)} \\
&= \ln P(\mathbf{x}|\mathcal{C}_1) - \ln P(\mathbf{x}|\mathcal{C}_2) + \ln \frac{P(\mathcal{C}_1)}{P(\mathcal{C}_2)} \\
&= \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \ln \frac{P(\mathcal{C}_1)}{P(\mathcal{C}_2)} \\
&= \frac{1}{2}\left[\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}\mathbf{x} - 2\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}\mathbf{x} + \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\mathbf{x} - 2\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\mathbf{x} + \boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1\right] + \ln \frac{P(\mathcal{C}_1)}{P(\mathcal{C}_2)} \\
&= \frac{1}{2}\mathbf{x}^{\mathrm{T}}(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1})\mathbf{x} - \mathbf{x}^{\mathrm{T}}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1) + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 + \frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \ln \frac{P(\mathcal{C}_1)}{P(\mathcal{C}_2)}
\end{aligned}$$

Let

$$\begin{aligned}
\mathbf{S} &= \boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1} \\
\mathbf{w} &= \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 \\
w_0 &= \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 + \frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \ln \frac{P(\mathcal{C}_1)}{P(\mathcal{C}_2)}
\end{aligned}$$

Then the decision function for class 1 (posterior probability) is given by

$$\sigma(\mathbf{x}^{\mathrm{T}}\mathbf{S}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{w} + w_0).$$

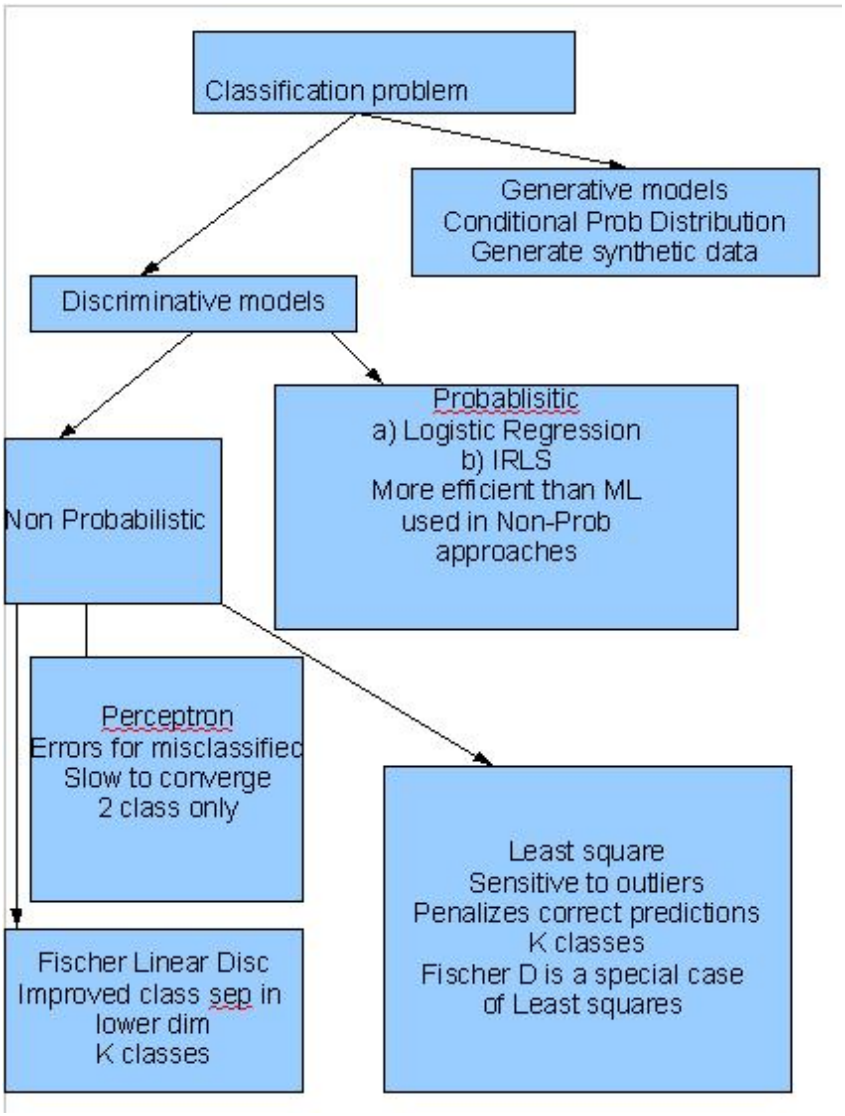
The decision function for class 2 is similarly derived.

■

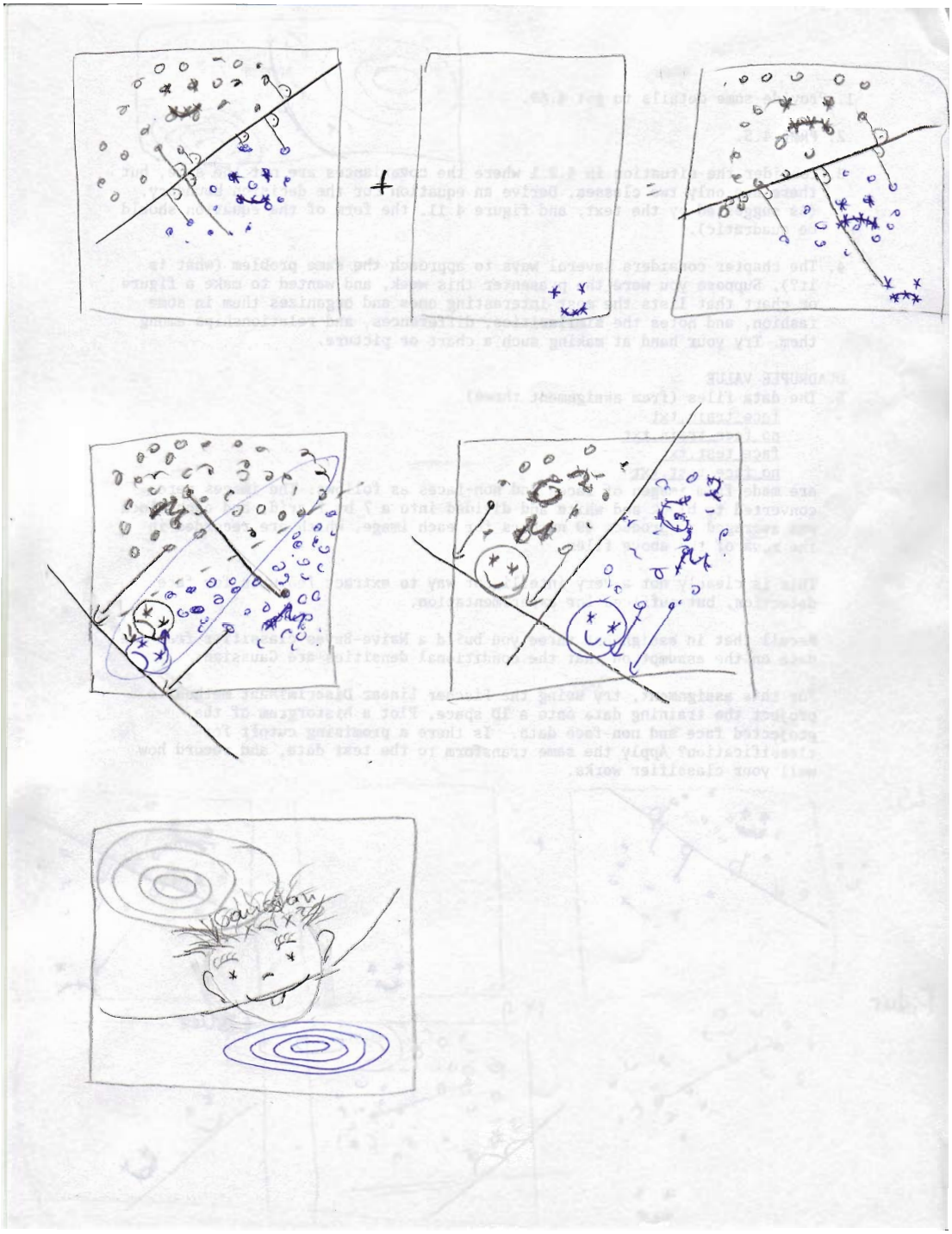
4. The chapter considers several ways to approach the same problem (what are they?). Suppose you were the presenter this week, and wanted to make a figure or chart that lists the most interesting ones and organizes them in some fashion, and notes the similarities, differences, and relationships among them. Try your hand at making such a chart or picture.

   Solution: Consider the following illustrations:

Method	Type	Assumptions	Multiple Classes	Outliers
Least Squares	Deterministic	Error measure	Yes	Bad
Fischer	Deterministic	-	Yes	Good
Perceptron	Deterministic	-	No	Good
Generative	Probabilistic	Likelihood form	Yes	Good
Discriminative	Probabilistic	-	Yes	Good



...or if you prefer, this more colorful solution:



■

5. The data files (from assignment three) are made from images of faces and non-faces as follows. The images were converted to black and white and divided into a 7 by 7 grid, and each block was averaged to produce 49 numbers for each image, which are recorded in the rows of the above files.

   This is clearly not a very intelligent way to extract features for face detection, but suffices for experimentation.

   Recall that in assignment three you build a Naive-Bayes classifier from the data on the assumption that the conditional densities are Gaussian.

   For this assignment, try using the Fischer Linear Discriminant method to project the training data onto a 1D space. Plot a historgram of the projected face and non-face data. Is there a promising cutoff for classification? Apply the same transform to the test data, and record how well your classifier works.

   Solution:

   I trained the fisher discriminant model with the training face data and plotted the unscaled discriminant for each of the 100 face and 100 non-face points. Figure 1 shows this plot. It is pretty clear from the plot that there is a nice value for w that separates the two sets. It is not a histogram, but I thought this plot was very nice in illustrating the division in the two sets. I choose $w_0 = -0.0291$ as the dividing value. Using this value can correctly discriminate between all the training data.

   I used the value of w_0 to try and discriminate between the face and non-face test data. I was surprised that the accuracy was perfect—I could correctly classify face or non-face for all 26 samples in the test set.

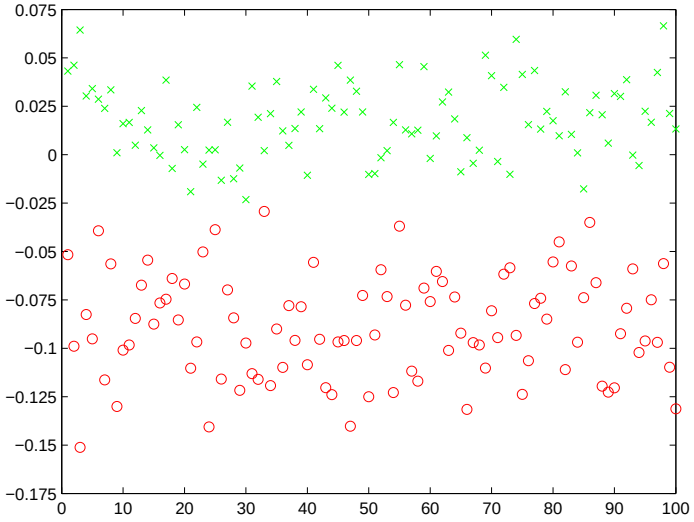
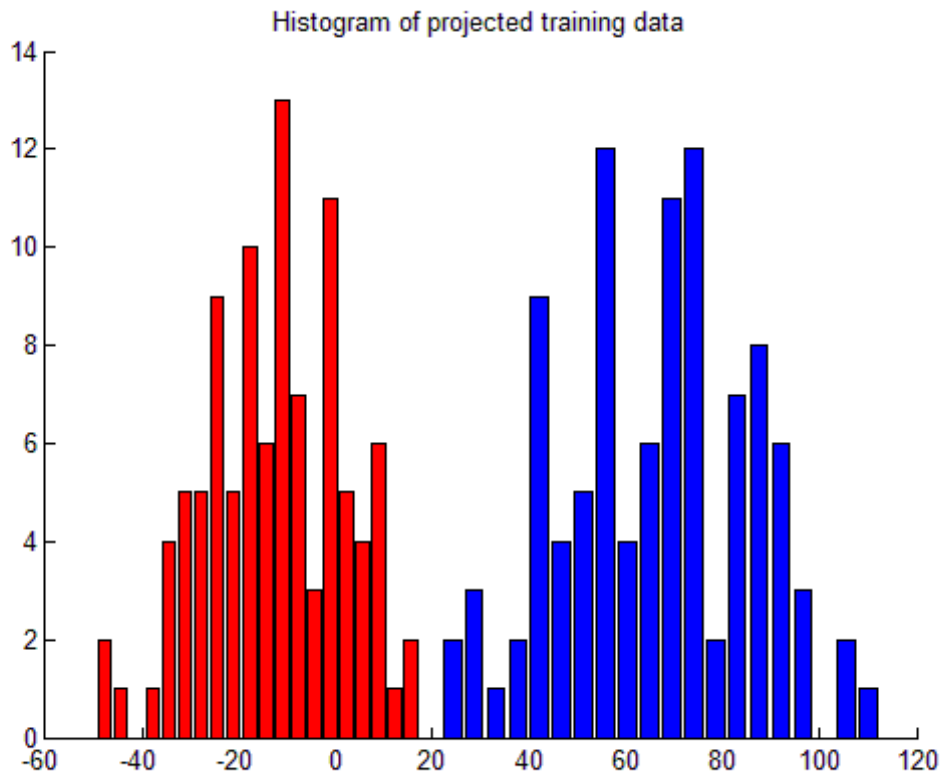


Figure 1: Plot of the fisher discriminate for the training data in the face and non-face sets. The x-axis represents the index of 100 points in each of the sets and the y-axis is the w. The red circles are the face points and the green crosses are the non-faces. Clearly there is a nice discriminate around $w_0 = -0.25$.


Histogram of projected training data
14
12
10
8
6
4
2
0
-60
-40
-20
0
20
40
60
80
100
120