Machine Learning 1 - Homework 4

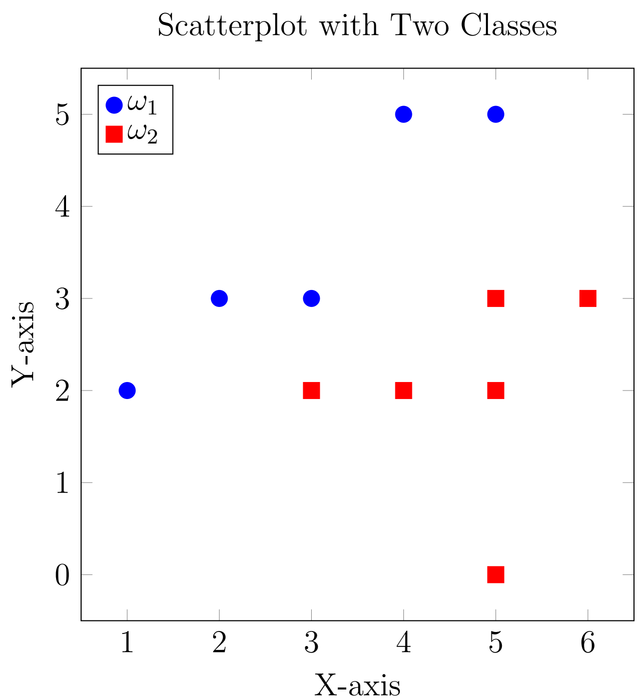
Christopher Williams

October 7, 2025

1 Phase 1

1.1 Two Classes

1.1.1 Scatter Plot



1.1.2 Class Variance

To calculate the between-class variance, S_B, and the within-class indicator, S_W, we need to calculate the mean of each class, and the total mean.

$$\mu_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i^{(1)} = \frac{1}{5}\begin{bmatrix}1+2+3+4+5\\2+3+3+5+5\end{bmatrix} = \frac{1}{5}\begin{bmatrix}15\\18\end{bmatrix} = \begin{bmatrix}3\\3.6\end{bmatrix} \tag{1}$$

$$\mu_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_i^{(2)} = \frac{1}{6}\begin{bmatrix}4+5+5+3+5+6\\2+0+2+2+3+3\end{bmatrix} = \frac{1}{6}\begin{bmatrix}28\\12\end{bmatrix} = \begin{bmatrix}4.67\\2\end{bmatrix} \tag{2}$$

$$\mu_{tot} = \frac{1}{n_{tot}}\sum_{i=1}^{n_1+n_2} x_i^{(1\&2)} = \frac{1}{11}\begin{bmatrix}1+2+3+4+5+4+5+5+3+5+6\\2+3+3+5+5+2+0+2+2+3+3\end{bmatrix} = \frac{1}{11}\begin{bmatrix}43\\30\end{bmatrix} = \begin{bmatrix}3.91\\2.73\end{bmatrix} \tag{3}$$

After finding the means, we can now find the between-class variance with this equation

$$S_B = n_1(\mu_1 - \mu_{tot})(\mu_1 - \mu)^T + n_2(\mu_2 - \mu_{tot})^T \tag{4}$$

$$S_B = 5\Bigg(\begin{bmatrix} 3-3.91 \\ 3.6-2.73 \end{bmatrix}\begin{bmatrix} 3-3.91 & 3.6-2.73 \end{bmatrix}\Bigg) + 6\Bigg(\begin{bmatrix} 4.67-3.91 \\ 2-2.73 \end{bmatrix}\begin{bmatrix} 4.67-3.91 & 2-2.73 \end{bmatrix}\Bigg) \tag{5}$$

$$\boxed{S_B = \begin{bmatrix} 7.63 & -7.25 \\ -7.25 & 6.98 \end{bmatrix}} \tag{6}$$

Next, we can find the within-class variance with this equation

$$S_W = \sum_{x_1 \in \omega_1} (x_1 - \mu_1)^T (x_1 - \mu_1) + \sum_{x_2 \in \omega_2} (x_2 - \mu_2)^T (x_2 - \mu_2) \tag{7}$$

$$S_W = \begin{bmatrix} 1-3 & 2-3.6 \\ 2-3 & 3-3.6 \\ 3-3 & 3-3.6 \\ 4-3 & 5-3.6 \\ 5-3 & 5-3.6 \end{bmatrix}\begin{bmatrix} 1-3 & 2-3 & 3-3 & 4-3 & 5-3 \\ 2-3.6 & 3-3.6 & 3-3.6 & 5-3.6 & 5-3.6 \end{bmatrix} + \tag{8}$$

$$\begin{bmatrix} 4-4.6 & 2-2 \\ 5-4.6 & 0-2 \\ 5-4.6 & 2-2 \\ 3-4.6 & 2-2 \\ 5-4.6 & 3-2 \\ 6-4.6 & 3-2 \end{bmatrix}\begin{bmatrix} 4-4.6 & 5-4.6 & 5-4.6 & 3-4.6 & 5-4.6 & 6-4.6 \\ 2-2 & 0-2 & 2-2 & 2-2 & 3-2 & 3-2 \end{bmatrix} \tag{9}$$

$$= \begin{bmatrix} 10 & 8 \\ 8 & 7.2 \end{bmatrix} + \begin{bmatrix} 5.33 & 1.00 \\ 1.00 & 6.00 \end{bmatrix} \tag{10}$$

$$S_W = \boxed{\begin{bmatrix} 15.33 & 9.00 \\ 9.00 & 13.20 \end{bmatrix}} \tag{11}$$

1.2 Spectral Decomposition of Fisher Criterion

1.2.1 Eigenvectors

To find the direction that maximizes class separation, we solve the generalized eigenvectors:

$$S_B \mathbf{w} = \lambda S_W \mathbf{w} \tag{12}$$

This is the equivalent to finding eigenvalues and eigenvectors of $S_W^{-1} S_B$.
First, S_W^{-1}:

$$\det(S_W) = 15.33 \times 13.20 - 9.00 \times 9.00 = 121.36 \tag{13}$$

$$S_W^{-1} = \frac{1}{121.36}\begin{bmatrix} 13.20 & -9.00 \\ -9.00 & 15.33 \end{bmatrix} = \begin{bmatrix} 0.1087 & -0.0741 \\ -0.0741 & 0.1263 \end{bmatrix} \tag{14}$$

Then $S_W^{-1} S_B$:

$$S_W^{-1} S_B = \begin{bmatrix} 0.1087 & -0.0741 \\ -0.0741 & 0.1263 \end{bmatrix}\begin{bmatrix} 7.63 & -7.25 \\ -7.25 & 6.98 \end{bmatrix} = \begin{bmatrix} 1.362 & -1.308 \\ -1.480 & 1.421 \end{bmatrix} \tag{15}$$

Solving the characteristic equation $\det(S_W^{-1}S_B - \lambda I) = 0$:

$$\lambda^2 - 2.78\lambda = 0 \tag{16}$$
$$\lambda_1 = 2.78 \tag{17}$$
$$\lambda_2 = 0 \tag{18}$$

For $\lambda_1 = 2.78$, the eigenvector is:

$$\mathbf{w}_1 = \begin{bmatrix} -0.68 \\ 0.73 \end{bmatrix} \tag{19}$$

For $\lambda_2 = 0$, the eigenvector is:

$$\mathbf{w}_2 = \begin{bmatrix} 0.69 \\ 0.72 \end{bmatrix} \tag{20}$$

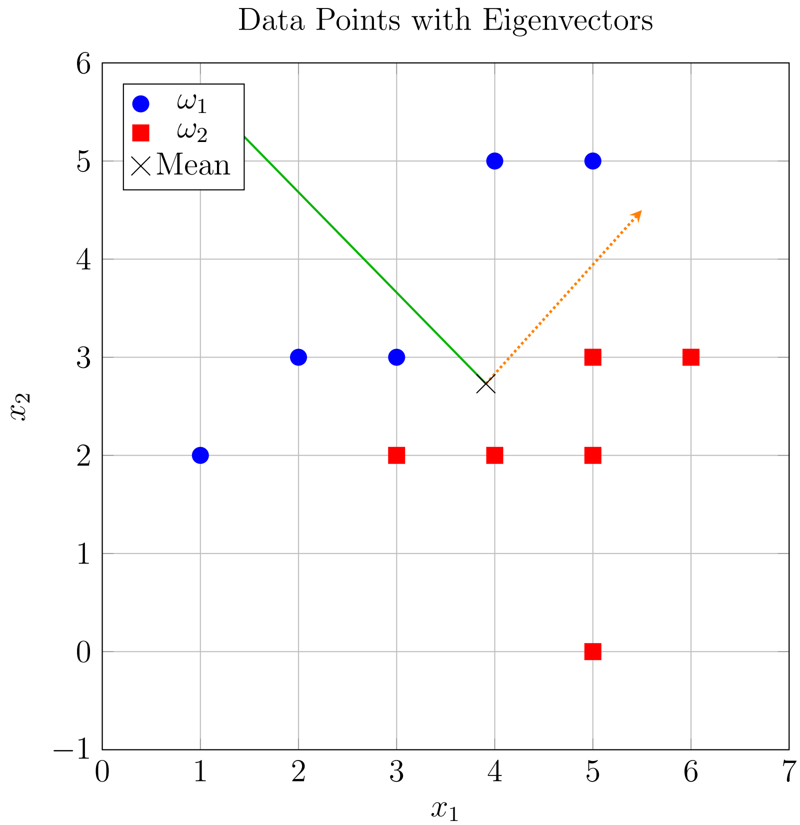


Figure 1: Eigenvectors plotted from the overall mean. $\mathbf{w}_1$ (green) corresponds to $\lambda_1 = 2.783$ and provides maximum class separation. $\mathbf{w}_2$ (orange, dotted) corresponds to $\lambda_2 = 0$ and provides no discriminative information.

1.2.2 Display Data

Projecting the data onto eigenvector $\mathbf{w}_1 = [-0.67, 0.73]^T$:

For each point $\mathbf{x}$, the projection is $y = \mathbf{w}_1^T\mathbf{x} = -0.67x_1 + 0.73x_2$.

Class 1 projections:

$$\begin{aligned}
(1,2):&\quad y = -0.67(1) + 0.73(2) = 0.79\\
(2,3):&\quad y = -0.67(2) + 0.73(3) = 0.85\\
(3,3):&\quad y = -0.67(3) + 0.73(3) = 0.17\\
(4,5):&\quad y = -0.67(4) + 0.73(5) = 0.97\\
(5,5):&\quad y = -0.67(5) + 0.73(5) = 0.29
\end{aligned}$$

Mean: $\bar{y}_1 = 0.61$

Class 2 projections:

$$\begin{aligned}
(4,2):&\quad y = -0.67(4) + 0.73(2) = -1.23\\
(5,0):&\quad y = -0.67(5) + 0.73(0) = -3.38\\
(5,2):&\quad y = -0.67(5) + 0.73(2) = -1.91\\
(3,2):&\quad y = -0.67(3) + 0.73(2) = -0.55\\
(5,3):&\quad y = -0.67(5) + 0.73(3) = -1.17\\
(6,3):&\quad y = -0.67(6) + 0.73(3) = -1.85
\end{aligned}$$

Mean: $\bar{y}_2 = -1.68$

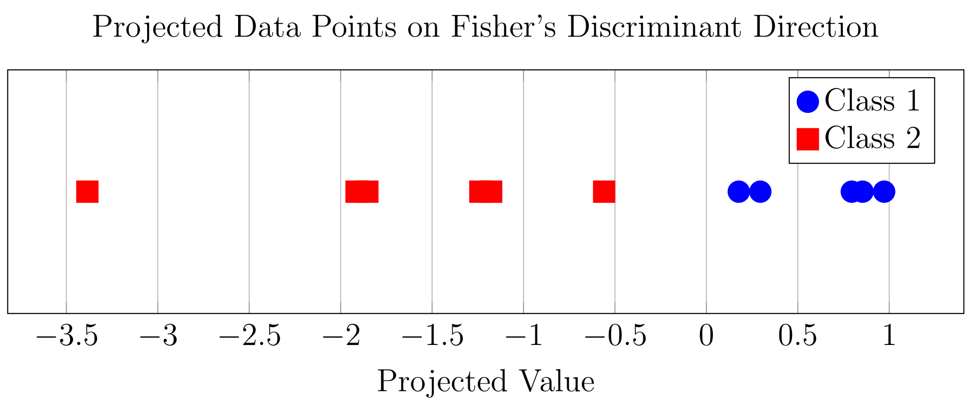


Figure 2: Projected data points showing clear separation between classes along the discriminant direction.

1.3 Reduce Dimensionality

1.3.1 Plot of Probability Density

Choice of projection direction: We choose eigenvector $\mathbf{w}_1 = [-0.67, 0.73]^T$ for the projection because it corresponds to the largest eigenvalue ($\lambda_1 = 2.783$), which maximizes the Fisher criterion. This direction maximizes the ratio of between-class variance to within-class variance, providing optimal class separation.

The second eigenvector $\mathbf{w}_2$ has eigenvalue $\lambda_2 = 0$, meaning it provides no discriminative information and would not help separate the classes.

After projection, Class 1 has mean 0.619 and Class 2 has mean -1.687, showing clear separation in the 1D subspace.

1.4 L2 Distance Calculations

Before LDA (Original 2D space):

$$||\mu_1 - \mu_2||_2 = \sqrt{(3-4.6)^2 + (3.6-2)^2} = \sqrt{2.77 + 2.56} = \sqrt{5.338} = 2.31 \tag{21}$$

After LDA (Projected 1D space):

$$|\bar{y}_1 - \bar{y}_2| = |0.61 - (-1.68)| = |2.30| = 2.30 \tag{22}$$

Comparison:

- Distance before LDA: 2.31
- Distance after LDA: 2.30
- Difference: 0.01 (0.22%)

The L2 distance between class means is nearly preserved after projection (difference of only 0.005). This demonstrates that Fisher's LDA successfully identifies the projection direction that maintains class separation while reducing dimensionality from 2D to 1D. The small reduction is expected and shows that the discriminant direction $\mathbf{w}_1$ captures almost all the information relevant for class separation.

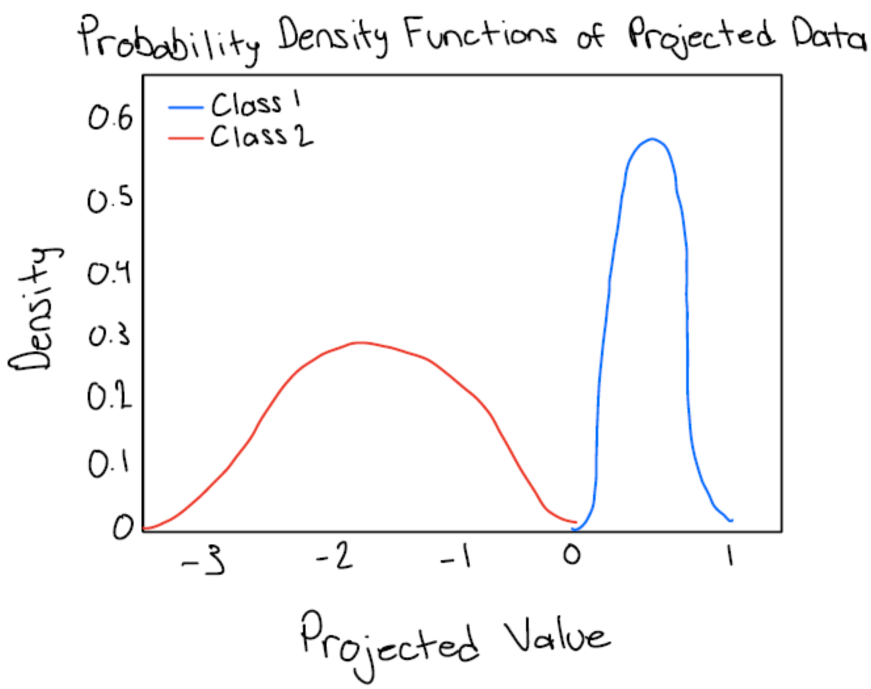


Figure 3: Sketch of Probabiltiy Dense Function

2 Phase 2

2.1 Politician Face's Dataset

The number of unique politicians (classes) are in the dataset is $\boxed{82}$.

The number of observations (images) in the dataset is $\boxed{3382}$

The dimensionality of the dataset is $\boxed{2914}$

The names of all the politicians in this dataset are as follows:

1. Abdullah Gul
2. Alejandro Toledo
3. Alvaro Uribe
4. Amelie Mauresmo
5. Andre Agassi
6. Angelina Jolie
7. Ariel Sharon
8. Arnold Schwarzenegger
9. Atal Bihari Vajpayee
10. Bill Clinton
11. Bill Gates
12. Carlos Menem
13. Carlos Moya
14. Colin Powell
15. David Beckham
16. Donald Rumsfeld
17. Fidel Castro
18. George Robertson
19. George W Bush
20. Gerhard Schroeder
21. Gloria Macapagal Arroyo
22. Gray Davis
23. Guillermo Coria
24. Hamid Karzai
25. Hans Blix
26. Hugo Chavez
27. Igor Ivanov
28. Jack Straw
29. Jacques Chirac
30. Jean Charest
31. Jean Chretien
32. Jennifer Aniston
33. Jennifer Capriati
34. Jennifer Lopez
35. Jeremy Greenstock
36. Jiang Zemin
37. John Ashcroft
38. John Bolton
39. John Howard
40. John Kerry
41. John Negroponte
42. John Snow
43. Joschka Fischer
44. Jose Maria Aznar
45. Juan Carlos Ferrero
46. Julianne Moore
47. Junichiro Koizumi
48. Kofi Annan
49. Lance Armstrong
50. Laura Bush
51. Lindsay Davenport
52. Lleyton Hewitt
53. Luiz Inacio Lula da Silva
54. Mahmoud Abbas
55. Megawati Sukarnoputri
56. Michael Bloomberg
57. Michael Schumacher
58. Naomi Watts
59. Nestor Kirchner
60. Nicole Kidman
61. Paul Bremer
62. Pervez Musharraf
63. Pete Sampras
64. Recep Tayyip Erdogan
65. Renee Zellweger
66. Ricardo Lagos
67. Richard Myers
68. Roh Moo-hyun
69. Rudolph Giuliani
70. Saddam Hussein
71. Serena Williams
72. Silvio Berlusconi
73. Spencer Abraham
74. Tiger Woods
75. Tim Henman
76. Tom Daschle
77. Tom Ridge
78. Tony Blair
79. Venus Williams
80. Vicente Fox
81. Vladimir Putin
82. Winona Ryder

2.2 First Twenty Politicians

Here is the subplot containing the images of the first 20 politicians.

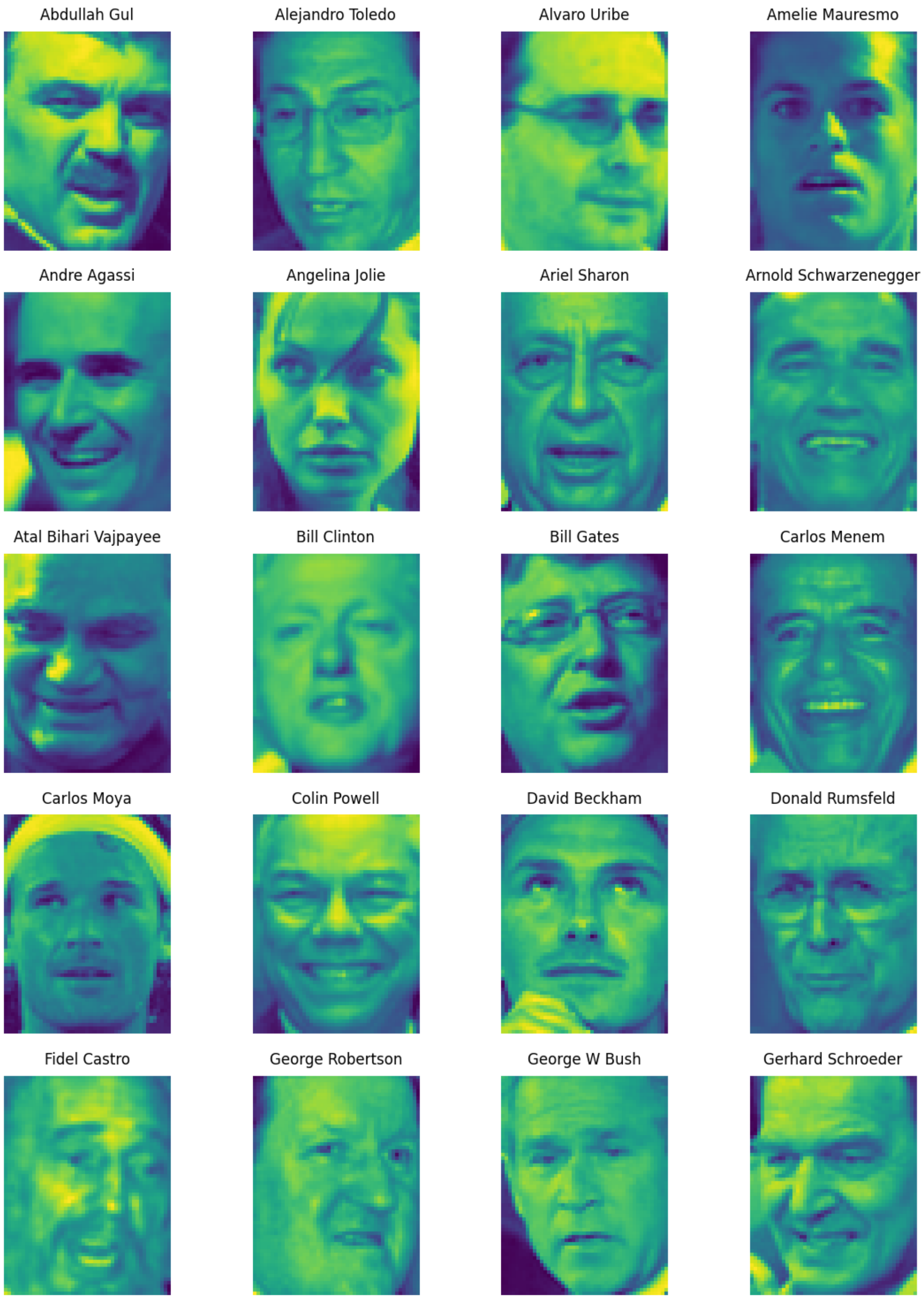


Figure 4: Scatterplot of x vs. y with Eigenvectors

2.3 Image Classification

2.3.1 Required Classes

As determined in Question 5 (Section 2.1), it would be equal to the number of unique politicians in the dataset, which is $\boxed{82}$.

2.3.2 Balanced Dataset?

The dataset is very unbalanced. The minimum number of images per politician is 17, but the maximum for one politician is 530. These numbers are very unbalanced.

Here is a plot show casing the disparity.

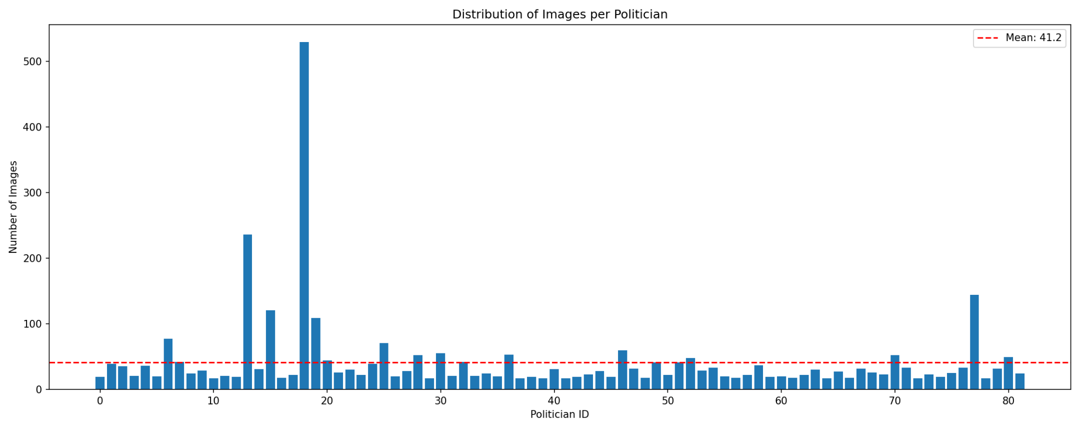


Figure 5: Distribution of Images per Politician

2.4 Standardize and LDA

Before computing the S_B and S_W, we must standardize the data matrix

```
X_mean = X.mean(axis=0)  # mean of each feature
X_std = np.sqrt(((X - X_mean) ** 2).mean(axis=0))  # std using mean

X_standardized = (X - X_mean) / X_std
```

And now we can compute the within and between class variances.

2.4.1 Compute SW

The output of the code in the console of the first 5x5 block of the S_W matrix is

```
[[2322.4009409  2169.98031604 1882.10101306 1591.92321476 1387.70772718]
 [2169.98031604 2283.24829829 2123.59060502 1797.15796739 1535.31690915]
 [1882.10101306 2123.59060502 2275.36009753 2107.61102676 1804.45945579]
 [1591.92321476 1797.15796739 2107.61102676 2293.39020145 2147.38287556]
 [1387.70772718 1535.31690915 1804.45945579 2147.38287556 2324.52218747]]
```

2.4.2 Compute SB

The output of the code in the console of the first 5x5 block of the S_B matrix is

```
[[1059.59459009 1072.23407495 1050.89820118  994.5277415   908.67901927]
 [1072.23407495 1098.75804608 1091.22989577 1044.25127986  962.54941031]
 [1050.89820118 1091.22989577 1106.63791166 1081.28099694 1013.88409876]
 [ 994.5277415  1044.25127986 1081.28099694 1088.61423538 1052.09610274]
 [ 908.67901927  962.54941031 1013.88409876 1052.09610274 1057.47941533]]
```

2.5 Spectral Decomposition of Fisher Criterion

2.5.1 Five Largest Eigenvalues

The largest five eigenvalues with their corresopnding eigenvectors:

```
Top 5 eigenvalues: [36.43 34.79 26.89 26.65 24.02]

Top 5 eigenvectors: [[ 0.01  0.    -0.    -0.01  0.  ]
 [-0.02 -0.    0.    0.01 -0.  ]
 [ 0.02 -0.02 -0.   -0.   -0.01]
 ...
 [-0.    0.   -0.   -0.01 -0.01]
 [ 0.01  0.02 -0.01  0.01  0.01]
 [-0.   -0.01  0.   -0.   -0.  ]]
```

I can't display the entire eigenvectors of this very multidimensional eigenvector.

2.5.2 Five Smallest Eigenvalues

Now the bottom 5:

```
Bottom 5 eigenvalues: [-0. -0. -0. -0. -0.]

Bottom 5 eigenvectors: [[ 0.    0.    0.    0.01 -0.  ]
 [-0.02 -0.   -0.01 -0.02 -0.01]
 [ 0.02 -0.01  0.    0.02  0.  ]
 ...
 [ 0.01  0.01 -0.01 -0.01 -0.  ]
 [-0.   -0.    0.01  0.    0.  ]
 [-0.   -0.   -0.   -0.   -0.  ]]
```