Gaussian Discriminant Analysis (GDA)

Introduction

Gaussian Discriminant Analysis (GDA) is a probabilistic classification model based on the assumption that the data within each class follows a Gaussian (normal) distribution. GDA is a type of generative model, which means it models the joint probability distribution of the features and the class labels. It is particularly useful when the underlying data distribution is approximately Gaussian, and it provides a principled way to handle classification problems by leveraging probability theory.

Core Concepts

1. **Generative Model**: Unlike discriminative models (e.g., logistic regression) that model $P(y|x)$, GDA models the joint probability $P(x, y)$ and then uses Bayes' theorem to derive $P(y|x)$.
2. **Gaussian Distribution**: For GDA, each class y is assumed to have its own Gaussian distribution:
   $P(x|y = c) = \frac{1}{(2\pi)^{d/2}|\Sigma_c|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_c)^T \Sigma_c^{-1}(x - \mu_c)\right)$
   where μ_c and Σ_c are the mean vector and covariance matrix for class c, respectively.
3. **Prior Probabilities**: The prior probability $P(y = c)$ represents the likelihood of class c in the dataset.

Theory and Assumptions

Model Assumptions

GDA makes the following assumptions about the data:

1. **Class-conditional distribution**: The feature vectors x given a class y are normally distributed. Specifically,
   $p(x|y = k) = \mathcal{N}(x|\mu_k, \Sigma)$
   where μ_k is the mean vector of the features for class k and Σ is the shared covariance matrix across all classes.
2. **Prior probability**: Each class has a prior probability $\phi_k = p(y = k)$.

Gaussian Discriminant Analysis Model

The GDA model can be summarized as follows:

1. **Prior distribution**: $p(y)$
2. **Likelihood**: $p(x|y)$
3. **Posterior distribution**: Using Bayes' theorem, we can compute the posterior distribution $p(y|x)$ which is used for classification.

For a binary classification problem ($y \in \{0, 1\}$), the likelihoods are:
$p(x|y = 0) = \mathcal{N}(x|\mu_0, \Sigma)$
$p(x|y = 1) = \mathcal{N}(x|\mu_1, \Sigma)$

Using Bayes' theorem, the posterior distribution is given by:
$p(y = 1|x) = \frac{p(x|y=1)p(y=1)}{p(x)}$
$p(y = 0|x) = \frac{p(x|y=0)p(y=0)}{p(x)}$

For classification, we assign the label $y = 1$ if $p(y = 1|x) > p(y = 0|x)$, and $y = 0$ otherwise.

Derivation of Parameters

To fit a GDA model, we need to estimate the parameters ϕ, μ_0, μ_1, and Σ from the training data.

Maximum Likelihood Estimate of ϕ

The parameter ϕ represents the prior probability of the class $y = 1$. It can be estimated as the fraction of the training examples that belong to class 1:

$\hat{\phi} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{y^{(i)} = 1\}$

where $\mathbf{1}\{\cdot\}$ is the indicator function, which is 1 if the argument is true and 0 otherwise, and m is the total number of training examples.

Step-by-step approach to estimate ϕ using Maximum Likelihood Estimation (MLE) involves setting up the likelihood function based on the data and then finding the value of ϕ that maximizes this likelihood.

Define the Likelihood Function

Given a dataset $\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$, where $x^{(i)}$ is the feature vector and $y^{(i)} \in \{0, 1\}$ is the class label, the likelihood function for ϕ is the probability of observing the given labels under the Bernoulli distribution parameterized by ϕ. The likelihood function is:

$L(\phi) = p(y^{(1)}, y^{(2)}, \ldots, y^{(m)}; \phi)$

Express the Likelihood Function in Terms of ϕ

Since the labels are assumed to be independent and identically distributed (i.i.d.), the joint probability of the labels is the product of the individual probabilities. For each $y^{(i)}$:

$p(y^{(i)} = 1; \phi) = \phi$
$p(y^{(i)} = 0; \phi) = 1 - \phi$

Thus, the likelihood function becomes:

$L(\phi) = \prod_{i=1}^{m} p(y^{(i)}; \phi) = \prod_{i=1}^{m} \phi^{y^{(i)}} (1 - \phi)^{1 - y^{(i)}}$

Take the Logarithm of the Likelihood Function

To simplify the maximization process, we take the natural logarithm of the likelihood function to obtain the log-likelihood function:

$\ell(\phi) = \log L(\phi) = \log \left(\prod_{i=1}^{m} \phi^{y^{(i)}} (1 - \phi)^{1 - y^{(i)}} \right)$

Using the properties of logarithms, this simplifies to:

$\ell(\phi) = \sum_{i=1}^{m} \left(y^{(i)} \log \phi + (1 - y^{(i)}) \log(1 - \phi) \right)$

Differentiate the Log-Likelihood Function

To find the maximum likelihood estimate, we differentiate the log-likelihood function with respect to ϕ:

$\frac{\partial \ell(\phi)}{\partial \phi} = \sum_{i=1}^{m} \left(\frac{y^{(i)}}{\phi} - \frac{1 - y^{(i)}}{1 - \phi} \right)$

Set the Derivative to Zero

To find the maximum, set the derivative to zero and solve for ϕ:

$\sum_{i=1}^{m} \left(\frac{y^{(i)}}{\phi} - \frac{1 - y^{(i)}}{1 - \phi} \right) = 0$

Simplify the Equation

Rearrange the terms to isolate ϕ:

$\sum_{i=1}^{m} \frac{y^{(i)}}{\phi} = \sum_{i=1}^{m} \frac{1 - y^{(i)}}{1 - \phi}$

$\frac{1}{\phi} \sum_{i=1}^{m} y^{(i)} = \frac{1}{1 - \phi} \sum_{i=1}^{m} (1 - y^{(i)})$

Let m_1 be the number of examples where $y = 1$, i.e., $m_1 = \sum_{i=1}^{m} y^{(i)}$, and let m_0 be the number of examples where $y = 0$, i.e., $m_0 = \sum_{i=1}^{m} (1 - y^{(i)})$. Note that $m_0 + m_1 = m$.

Substituting these into the equation, we get:

$\frac{m_1}{\phi} = \frac{m_0}{1 - \phi}$

Solve for ϕ

Solve for ϕ:

$m_1(1-\phi) = (m-m_1)\phi$

$m_1 - m_1\phi = m\phi - m_1\phi$

$m_1 = m\phi$

$\phi = \frac{m_1}{m}$

Conclusion

The maximum likelihood estimate of ϕ is:

$\hat{\phi} = \frac{m_1}{m} = \frac{1}{m}\sum_{i=1}^{m} y^{(i)}$

In the context of GDA, using the indicator function $\mathbf{1}\{y^{(i)} = 1\}$, we can express this as:

$\hat{\phi} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{y^{(i)} = 1\}$

This result indicates that the MLE of ϕ is simply the proportion of the training examples that belong to class 1.

Maximum Likelihood Estimate of μ_0 and μ_1

The means μ_0 and μ_1 are the average feature vectors for the examples in each class:

$\hat{\mu}_0 = \frac{\sum_{i=1}^{m} \mathbf{1}\{y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^{m} \mathbf{1}\{y^{(i)}=0\}}$
$\hat{\mu}_1 = \frac{\sum_{i=1}^{m} \mathbf{1}\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{m} \mathbf{1}\{y^{(i)}=1\}}$

To derive the Maximum Likelihood Estimate (MLE) of the means μ_0 and μ_1 we follow a systematic approach. Let's derive $\hat{\mu}_0$ in detail. The derivation of $\hat{\mu}_1$ follows the same logic.

1. **Formulate the Likelihood Function:**
   Given m training examples $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})$, we assume that x given y follows a Gaussian distribution. For class $y = 0$:

   $p(x|y=0) = \mathcal{N}(x|\mu_0, \Sigma)$

   The likelihood of observing all the training data under the class $y = 0$ is:

   $L(\mu_0, \Sigma|\{(x^{(i)}, y^{(i)}) \text{ where } y^{(i)} = 0\}) = \prod_{i:y^{(i)}=0} \mathcal{N}(x^{(i)}|\mu_0, \Sigma)$
2. **Log-Likelihood Function:**
   Since the logarithm of the likelihood function is easier to maximize, we consider the log-likelihood. For the subset of data points where $y^{(i)} = 0$:

   $\log L(\mu_0, \Sigma|\{(x^{(i)}, y^{(i)}) \text{ where } y^{(i)} = 0\}) = \sum_{i:y^{(i)}=0} \log \mathcal{N}(x^{(i)}|\mu_0, \Sigma)$

   The multivariate Gaussian (or normal) distribution for a random vector x with mean vector μ and covariance matrix Σ is given by:

   $\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$

   where:

   - n is the dimensionality of x.
   - $|\Sigma|$ is the determinant of the covariance matrix Σ.
   - Σ^{-1} is the inverse of the covariance matrix Σ.
   - $(x-\mu)^T$ is the transpose of the vector $(x-\mu)$.

   Let's break down the transformation of the Gaussian density function to its log-likelihood step by step.

   Log-Likelihood Function

   To derive the log-likelihood, we take the natural logarithm of the Gaussian density function. Let's denote the Gaussian density function by $p(x|\mu, \Sigma)$:

$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$

Taking the natural logarithm of both sides, we get:

$\log p(x|\mu, \Sigma) = \log\left(\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)\right)$

We can split the logarithm of a product into the sum of the logarithms:

$\log p(x|\mu, \Sigma) = \log\left(\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\right) + \log\left(\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)\right)$

Simplify the Logarithms

1. **Logarithm of the first term**:
   $\log\left(\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\right) = \log\left((2\pi)^{-n/2}\right) + \log\left(|\Sigma|^{-1/2}\right)$

   Using properties of logarithms:
   $\log\left((2\pi)^{-n/2}\right) = -\frac{n}{2}\log(2\pi)$
   $\log\left(|\Sigma|^{-1/2}\right) = -\frac{1}{2}\log|\Sigma|$

   So, combining these results:
   $\log\left(\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\right) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma|$
2. **Logarithm of the second term**:
   $\log\left(\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)\right)$

   Since the logarithm and the exponential functions are inverses:
   $\log\left(\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)\right) = -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$

Combine the Results

Putting it all together, the log-likelihood of the Gaussian density function is:

$\log p(x|\mu, \Sigma) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$

This is the expression we use for the log-likelihood of a single data point x. For multiple data points, we sum this expression over all the data points in our dataset.

3. **Simplifying the Log-Likelihood:**
   Substituting the expression for the Gaussian log-density into the log-likelihood function for class $y = 0$:
   $\log L(\mu_0, \Sigma|\{(x^{(i)}, y^{(i)}) \text{ where } y^{(i)} = 0\}) = \sum_{i:y^{(i)}=0}\left(-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right)$
   Ignoring the constants $-\frac{n}{2}\log(2\pi)$ and $-\frac{1}{2}\log|\Sigma|$ which do not depend on μ_0:
   $\log L(\mu_0|\{(x^{(i)}, y^{(i)}) \text{ where } y^{(i)} = 0\}) = -\frac{1}{2}\sum_{i:y^{(i)}=0}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)$
4. **Maximize the Log-Likelihood:**
   To find $\hat{\mu}_0$, we take the derivative of the log-likelihood with respect to μ_0 and set it to zero:
   $\frac{\partial}{\partial\mu_0}\left(-\frac{1}{2}\sum_{i:y^{(i)}=0}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right) = 0$
   Simplifying the derivative:
   $-\frac{1}{2}\sum_{i:y^{(i)}=0}\left(\Sigma^{-1}(x^{(i)}-\mu_0) + \Sigma^{-1}(x^{(i)}-\mu_0)^T\right) = 0$
   Since $(x^{(i)}-\mu_0)^T\Sigma^{-1}$ is a scalar and its transpose is itself:
   $\sum_{i:y^{(i)}=0}\Sigma^{-1}(x^{(i)}-\mu_0) = 0$
   Multiplying both sides by Σ:
   $\sum_{i:y^{(i)}=0}(x^{(i)}-\mu_0) = 0$
   This implies:
   $\sum_{i:y^{(i)}=0}x^{(i)} = \sum_{i:y^{(i)}=0}\mu_0$
   Let m_0 be the number of examples where $y^{(i)} = 0$. Then:
   $\sum_{i:y^{(i)}=0}x^{(i)} = m_0\mu_0$
   Solving for μ_0:
   $\mu_0 = \frac{1}{m_0}\sum_{i:y^{(i)}=0}x^{(i)}$
5. **Generalize to All Classes:**
   The same derivation applies for class $y = 1$:

$\hat{\mu}_1 = \frac{1}{m_1} \sum_{i:y^{(i)}=1} x^{(i)}$

where m_1 is the number of examples where $y^{(i)} = 1$.

Final Estimates

Combining both results, we have the MLE for μ_0 and μ_1:

$$\hat{\mu}_0 = \frac{\sum_{i=1}^{m} \mathbf{1}\{y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^{m} \mathbf{1}\{y^{(i)}=0\}}$$

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{m} \mathbf{1}\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{m} \mathbf{1}\{y^{(i)}=1\}}$$

These estimates give us the mean vectors for each class based on the training data, derived through the principle of maximum likelihood estimation.

Estimating Σ

The covariance matrix Σ is shared across both classes and can be estimated as:

$$\hat{\Sigma} = \frac{1}{m} \sum_{i=1}^{m} \left(x^{(i)} - \hat{\mu}_{y^{(i)}}\right)\left(x^{(i)} - \hat{\mu}_{y^{(i)}}\right)^T$$

where $\hat{\mu}_{y^{(i)}}$ is the mean vector corresponding to the class of the i-th example.

Summary of Parameter Estimates

To summarize, the estimated parameters for GDA are:

1. **Prior probability**:
   $\hat{\phi} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{y^{(i)} = 1\}$
2. **Class means**:
   $\hat{\mu}_0 = \frac{\sum_{i=1}^{m} \mathbf{1}\{y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^{m} \mathbf{1}\{y^{(i)}=0\}}$
   $\hat{\mu}_1 = \frac{\sum_{i=1}^{m} \mathbf{1}\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{m} \mathbf{1}\{y^{(i)}=1\}}$
3. **Covariance matrix**:
   $\hat{\Sigma} = \frac{1}{m} \sum_{i=1}^{m} \left(x^{(i)} - \hat{\mu}_{y^{(i)}}\right)\left(x^{(i)} - \hat{\mu}_{y^{(i)}}\right)^T$

Decision Rule and Boundary

The decision boundary in GDA is the surface (or line in 2D) that separates different classes based on the model. For a binary classification problem with classes $y = 0$ and $y = 1$, the decision boundary is derived as follows:

Posterior Probability

Using Bayes' theorem, the posterior probability $P(y = 1|x)$ is:
$P(y = 1|x) = \frac{P(x|y=1)P(y=1)}{P(x)}$
Since $P(x) = P(x|y = 0)P(y = 0) + P(x|y = 1)P(y = 1)$, we have:
$P(y = 1|x) = \frac{P(x|y=1)P(y=1)}{P(x|y=0)P(y=0)+P(x|y=1)P(y=1)}$

Decision Rule

To decide the class for a new data point x, compare the posterior probabilities:

$$P(y = 1|x) \underset{y=0}{\overset{y=1}{\gtrless}} P(y = 0|x)$$

Taking the logarithm of the ratio, we get:

$$\log \frac{P(y=1|x)}{P(y=0|x)} \underset{y=0}{\overset{y=1}{\gtrless}} 0$$

Substituting the posterior probabilities:

$$\log \frac{P(x|y=1)P(y=1)}{P(x|y=0)P(y=0)} \underset{y=0}{\overset{y=1}{\gtrless}} 0$$

Simplifying further:

$$\log P(x|y = 1) + \log P(y = 1) \underset{y=0}{\overset{y=1}{\gtrless}} \log P(x|y = 0) + \log P(y = 0)$$

This leads to:

$$\log \frac{P(x|y=1)}{P(x|y=0)} \underset{y=0}{\overset{y=1}{\gtrless}} \log \frac{P(y=0)}{P(y=1)}$$

Substitute Gaussian Distribution

Substitute the Gaussian distributions for $P(x|y=0)$ and $P(x|y=1)$:

$$\log \frac{\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right)}{\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right)} \underset{y=0}{\overset{y=1}{\gtrless}} \log \frac{P(y=0)}{P(y=1)}$$

This simplifies to:

$$-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) \underset{y=0}{\overset{y=1}{\gtrless}} \log \frac{P(y=0)}{P(y=1)}$$

Simplify the Exponentials

The exponentials simplify to:

$$-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) \underset{y=0}{\overset{y=1}{\gtrless}} \log \frac{P(y=0)}{P(y=1)}$$

Expand the Quadratic Forms

Expand the quadratic forms:

$$-\frac{1}{2}\left[x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu_1 + \mu_1^T\Sigma^{-1}\mu_1\right] + \frac{1}{2}\left[x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu_0 + \mu_0^T\Sigma^{-1}\mu_0\right] \underset{y=0}{\overset{y=1}{\gtrless}} \log \frac{P(y=0)}{P(y=1)}$$

Cancel the Common Terms

Notice that the terms $x^T\Sigma^{-1}x$ cancel each other out:

$$-\frac{1}{2}\left[-2x^T\Sigma^{-1}\mu_1 + \mu_1^T\Sigma^{-1}\mu_1\right] + \frac{1}{2}\left[-2x^T\,\Sigma^{-1}\mu_0 + \mu_0^T\Sigma^{-1}\mu_0\right] \underset{y=0}{\overset{y=1}{\gtrless}} \log \frac{P(y=0)}{P(y=1)}$$

Combine the Terms

Combine the remaining terms:

$$-x^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + x^T\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 \underset{y=0}{\overset{y=1}{\gtrless}} \log \frac{P(y=0)}{P(y=1)}$$

Rearrange to isolate x:

$$x^T\Sigma^{-1}(\mu_0-\mu_1) \underset{y=0}{\overset{y=1}{\gtrless}} \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0) + \log \frac{P(y=0)}{P(y=1)}$$

Linear Form (When $\Sigma_0 = \Sigma_1$)

This is the linear form of the decision boundary:

$$x^T\Sigma^{-1}(\mu_1-\mu_0) \underset{y=0}{\overset{y=1}{\gtrless}} \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0) + \log \frac{P(y=0)}{P(y=1)}$$

- The left side $x^T\Sigma^{-1}(\mu_1-\mu_0)$ is a linear function of x.
- The right side is a constant.

This linear function forms the decision boundary that separates the two classes. When the covariance matrices are the same, the quadratic terms cancel out, resulting in a linear decision boundary.

Quadratic Decision Boundary (When $\Sigma_0 \neq \Sigma_1$)

Let's dive deeper into the explanation of the Quadratic Decision Boundary for Gaussian Discriminant Analysis (GDA) when the covariance matrices for the two classes are different ($\Sigma_0 \neq \Sigma_1$).

The decision rule for classifying a new data point x can be expressed as comparing the log-odds of the two classes:

$$\log \frac{P(y=1|x)}{P(y=0|x)} \underset{y=0}{\overset{y=1}{\gtrless}} 0$$

Using Bayes' theorem and the Gaussian distributions, this can be rewritten as:

$$\log P(y=1) + \log P(x|y=1) \underset{y=0}{\overset{y=1}{\gtrless}} \log P(y=0) + \log P(x|y=0)$$

Remember;

$$P(x|y=c) = \frac{1}{(2\pi)^{d/2}|\Sigma_c|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_c)^T\Sigma_c^{-1}(x-\mu_c)\right)$$

Substituting the Gaussian density functions, we get:

$$\log P(y=1) - \frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) \underset{y=0}{\overset{y=1}{\gtrless}} \log P(y=0) - \frac{1}{2}\log|\Sigma_0| - \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)$$

Rearranging terms, we have:

$$-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) + \log P(y=1) - \log P(y=0) - \frac{1}{2}\log|\Sigma_1| + \frac{1}{2}\log|\Sigma_0| \underset{y=0}{\overset{y=1}{\gtrless}} 0$$

Identifying Linear Terms and Constant

To better understand the components, let's expand and rearrange the quadratic terms:

1. **Quadratic Terms**:
   $-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)$ and $\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)$
2. **Linear Terms**:
   When we expand the quadratic terms, some terms involve x linearly. Let's expand both quadratic expressions:
   $(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) = x^T\Sigma_1^{-1}x - 2\mu_1^T\Sigma_1^{-1}x + \mu_1^T\Sigma_1^{-1}\mu_1$
   $(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) = x^T\Sigma_0^{-1}x - 2\mu_0^T\Sigma_0^{-1}x + \mu_0^T\Sigma_0^{-1}\mu_0$

   Therefore, the linear terms are:
   $\mu_1^T\Sigma_1^{-1}x - \mu_0^T\Sigma_0^{-1}x$
3. **Constant Terms**:
   The remaining terms are constants because they do not depend on x:
   $\frac{1}{2}\left(\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1\right) + \log P(y=1) - \log P(y=0) - \frac{1}{2}\log|\Sigma_1| + \frac{1}{2}\log|\Sigma_0|$

Putting it all together, the decision boundary involves quadratic terms in x, linear terms, and a constant. Specifically, the decision boundary can be written as:

$$-\frac{1}{2}x^T\Sigma_1^{-1}x + \mu_1^T\Sigma_1^{-1}x - \frac{1}{2}x^T\Sigma_0^{-1}x + \mu_0^T\Sigma_0^{-1}x \underset{y=0}{\overset{y=1}{\gtrless}} \text{constant}$$

Where the **linear terms** are:
$\mu_1^T\Sigma_1^{-1}x - \mu_0^T\Sigma_0^{-1}x$

And the **constant** is:
$\frac{1}{2}(\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1) + \log P(y=1) - \log P(y=0) - \frac{1}{2}\log|\Sigma_1| + \frac{1}{2}\log|\Sigma_0|$

Conclusion

Gaussian Discriminant Analysis is a powerful classification technique based on probabilistic principles and the assumption that the features follow a Gaussian distribution within each class. By estimating the parameters ϕ, μ_0, μ_1, and Σ, we can construct a model that effectively separates classes based on their statistical properties. GDA is particularly useful when the Gaussian assumption holds true and provides a solid foundation for understanding more complex generative models.