Self-Notes on [Multivariate Methods, PCA]

[Unat Tekşen] [504241592]

May 2, 2025

1 Multivariate Methods

The continuing part is in Section 4.

1.1 Review

- For each class, we should find μ, Σ. We will estimate them.
- If features are independent $\implies$ Σ is diagonal.

Each $p(x_i)$ will have a multivariate Gaussian distribution. The 2D Gaussian distribution is defined by:

$$P(x_1, x_2) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) = P(x_1)P(x_2) \tag{1}$$

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = -\frac{1}{2}\left[\left[\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right]^T \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}\right] \tag{2}$$

$$= -\frac{1}{2}\begin{bmatrix} \frac{(x_1-\mu_1)}{\sigma_1^2} & \frac{(x_2-\mu_2)}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}$$

$$= -\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2 + \frac{-1}{2\sigma_2^2}(x_2-\mu_2)^2 \tag{3}$$

1.2 Multivariate Parameters

- Mean
$$E[\mathbf{X}] = \boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_p]^T \tag{4}$$
- Covariance
$$\sigma_{ij} = \mathrm{Cov}(X_i, X_j) \tag{5}$$
- Correlation
$$\mathrm{Corr}(X_i, X_j) = \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \tag{6}$$

- Covariance matrix

$$\mathbf{\Sigma} = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] \tag{7}$$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22}^2 & \dots & \vdots \\ \vdots & \vdots & \ddots & \sigma_{dd-1} \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_{dd}^2 \end{pmatrix} \tag{8}$$

1.3 Multivariate Parameter Estimation

- Sample mean $\mathbf{m}$:

$$m_i = \frac{\sum_{j=1}^{N} x_{ij}}{N}, \quad (i = 1, \dots, d) \tag{9}$$

- Covariance matrix $\mathbf{S}$:

$$s_{ij} = \frac{\sum_{t=1}^{N} (x_i^t - m_i)(x_j^t - m_j)}{N} \tag{10}$$

- Correlation matrix $\mathbf{R}$:

$$r_{ij} = \frac{s_{ij}}{s_i s_j} \tag{11}$$

- If features X_i, X_j are:
  - Independent, then $\sigma_{ij} = 0$, diagonals are non-zero.

$$\begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sigma_d^2 \end{pmatrix} \tag{12}$$

  - Positive correlation, $\sigma_{ij} > 0$
  - Negative correlation, $\sigma_{ij} < 0$

1.4 Model Complexity: Bias - Variance

- As we increase complexity, bias decreases and variance increases
- Assume simple models to control variance (regularization)

Key changes ma

Assumption	Covariance matrix	No of parameters
Shared, Hyperspheric	$\mathbf{S}_i = \mathbf{S} = s\mathbf{I}$	1
Shared, Axis-aligned	$\mathbf{S}_i = \mathbf{S}$, with $s_{ij} = 0$	d
Shared, Hyperellipsoidal	$\mathbf{S}_i = \mathbf{S}$	$\frac{d(d+1)}{2}$
Different, Hyperellipsoidal	$\mathbf{S}_i$	$\frac{Kd(d+1)}{2}$

1.5 Discriminant Functions for Classification

(Saved computation of denominator $P(x)$)

- Classifier = m discriminant functions and classification is based on selecting the largest discriminant.
- $f(x) = g_i(x) - g_j(x) \longrightarrow$ depends on μ, Σ

$$g_i(x) = P(C_i|x) \propto P(x|C_i)P(C_i) \tag{13}$$

$$g_i(x) = \log P(x|C_i) + \log P(C_i) \tag{14}$$

$$\log P(x|C_i) = -\frac{1}{2}(x - \mu_i)^T \Sigma_i^T (x - \mu_i) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_i| + \log P(C_i) \tag{15}$$

- $\Sigma_0 = \sigma^2 I$
  $\Sigma_1 = \sigma^2 I$
  $\implies \Sigma = \Sigma_0 = \Sigma_1$, so ignore $-\frac{1}{2}\log|\Sigma_i|$

$$\text{If } \Sigma = \sigma^2 I, \text{ then } \Sigma^{-1} = \frac{1}{\sigma^2} I \tag{16}$$

Case $\Sigma_i = \sigma^2 I$

- Features are independent with different means and equal variances
- $\sigma^2 I = \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma^{-1} (x - \mu_i) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| + \log P(C_i) \tag{17}$$

Ignore these parts $\log 2\pi - \frac{1}{2}\log|\Sigma|$:

$$= -\frac{1}{2}(x - \mu_i)^T (\frac{1}{\sigma^2} I)(x - \mu_i) + \log P(C_i) \tag{18}$$

$$= -\frac{1}{2\sigma^2}(x - \mu_i)^T (x - \mu_i) + \log P(C_i) \tag{19}$$

$$= -\frac{1}{2\sigma^2}(x^Tx - x^T\mu_i - \mu_i^Tx + \mu_i^T\mu_i) \tag{20}$$

$$= -\frac{1}{2\sigma^2}(-2\mu_i^Tx + \mu_i^T\mu_i) + \log P(C_i) \tag{21}$$

Discriminant function is linear w.r.t x:

$$g_i(x) = w_i^Tx + w_{i0} \tag{22}$$

$$\text{Assume } \mu_0 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad \mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$x^Tx - \mu_i^Tx + \mu_i^T\mu_i$$

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} - \begin{bmatrix} 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + C \longrightarrow \text{scalar} \tag{23}$$

$$g_0(x) = -\frac{1}{2\sigma^2}[x^Tx - 2x^T\mu_0 + \mu_0^T\mu_0] \tag{24}$$

$$g_1(x) = -\frac{1}{2\sigma^2}[x^Tx - 2x^T\mu_1 + \mu_1^T\mu_1] \tag{25}$$

$$g_0(x) - g_1(x) = -\frac{1}{2\sigma^2}[-2x^T(\mu_0 - \mu_1) + \mu_0^T\mu_0 - \mu_1^T\mu_1] \tag{26}$$

$$= \frac{1}{2\sigma^2}[-2\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 13 - 2] \tag{27}$$

$$= +2x_1 + 4x_2 - 11 \tag{28}$$

Ex: $x^T\Sigma^{-1}\mu_i = \mu_i^T\Sigma^{-1}x$ because Σ is symmetric.

- If each class has its own Σ, quadratic term will be different.
- So, decision boundary $\longrightarrow$ ellipses, paraboloids.
- For example, for class 1 (μ_1, Σ_1) and class 2 (μ_2, Σ_2), the number of parameters are d and $\frac{d(d+1)}{2}$ respectively for each class, representing the mean vector and the symmetric covariance matrix in d dimensions.
- The expected squared error $E[(d - \theta)^2]$ is equal to bias + variance; as power/complexity increases, bias decreases and variance increases.
- From dataset:
  - $\Sigma_1 = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \quad P(C_1) = 0.8$
  - $\Sigma_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}, \quad P(C_2) = 0.2$
  - 80 examples from class 1, 20 examples from class 2

2 Dimensionality Reduction

Increasing features increases performance, but at some point, it won't help.

Curse of dimensionality: The curse of dimensionality refers to the challenges that arise as the number of features increases, including sparse data, less meaningful distance measures, the need for exponentially more data, and a higher risk of overfitting.

2.1 Feature Selection

It is about the best possible features. Consider a dataset with features $X_1, X_2, \ldots, X_d$ and a class label C_i:

$$\begin{bmatrix} 0.9 & 0.2 & \ldots & 1 \\ 0.7 & 0.5 & \ldots & 1 \\ 0.3 & 0.8 & \ldots & 1 \\ 0.1 & 0.4 & \ldots & 0 \\ 0.2 & 0.3 & \ldots & 0 \\ 0.1 & 0.1 & \ldots & 0 \end{bmatrix}$$

We aim to select a subset of features, for example, selecting X_1 as it correlates most with the class label.

There are different approaches to feature selection:

2.1.1 Filter-based Feature Selection

- Example: Filter is correlation.

2.1.2 Wrapper-based Feature Selection

- Example: Use classifier, select one feature, use error, and judge classifiers.
- Get rid of features that have less contribution.
- **Backward Feature Selection**
- A backward feature selection aims to find a subset of features, for instance:

$$\{X_1, X_2, ..., X_n\}, \{X_2, ..., X_n\}, \ldots$$

- **Forward Feature Selection**
- Start with an empty set, try different features, and analyze combinations.

$$F = \{\}, F = \{x_1\}, F = \{x_2\}, F = \{x_1, x_2\}$$

3 PCA

PCA Working Procedure

1. Standardize data (each feature has zero mean and unit variance).
2. Calculate covariance matrix.
3. Find eigenvectors and eigenvalues.
4. Select top eigenvectors (most significant principal components that capture the most variance in the data).
5. Project data onto these components for lower dimensionality.

Project x on w:

$$\text{length}(||\text{kw}||) = \|\mathbf{x}\| \cos\theta \tag{29}$$

$$l = \|\mathbf{x}\| \cdot \cos\theta \tag{30}$$

$$\mathbf{x} \cdot \mathbf{w} = \mathbf{x}^T \mathbf{w} = \mathbf{w}^T \mathbf{x} \tag{31}$$

$$l = \frac{\mathbf{w}^T \mathbf{x}}{\|\mathbf{w}\|} \tag{32}$$

Example:

$$\mathbf{w} = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \tag{33}$$

$$\frac{\mathbf{w}^T}{\|\mathbf{w}\|} = \frac{1}{5} \begin{bmatrix} 3 & 4 \end{bmatrix} \tag{34}$$

$$\mathbf{v} = \begin{bmatrix} 3/5 \\ 4/5 \end{bmatrix} \tag{35}$$

$$\|\mathbf{v}\| = 1 \tag{36}$$

In 2D:

- $\mathbf{w}_1^T \mathbf{x}$, $\|\mathbf{w}_1\| = 1$
- $\mathbf{w}_2^T \mathbf{x}$, $\|\mathbf{w}_2\| = 1$
- $\text{Var}(\mathbf{z})$ is maximized.
- $\sigma^2 = E[(x - \mu)^2]$
- $\Sigma = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]$

$$\text{Var}(x) = E[(x - E[x])^2] \tag{37}$$

$$= E[(\mathbf{w}^T \mathbf{x} - E[\mathbf{w}^T \mathbf{x}])^2] \tag{38}$$

- There is no randomness about $\mathbf{w}$, we just don't know it.
- There is randomness about $\mathbf{x}$.

$$E[\mathbf{w}^T\mathbf{x}] = \mathbf{w}^T E[\mathbf{x}], \qquad (E[\mathbf{x}] = \boldsymbol{\mu})$$

$$E[\mathbf{w}^T\mathbf{x}] = \mathbf{w}^T\boldsymbol{\mu}$$

Expanding expectation:

$$= E[(\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\boldsymbol{\mu})(\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\boldsymbol{\mu})] \tag{39}$$

Rewriting end term:

$$(\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\boldsymbol{\mu}) = (\mathbf{x}^T\mathbf{w} - \boldsymbol{\mu}^T\mathbf{w})$$

$$= E[\mathbf{w}^T(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\mathbf{w}] \tag{40}$$

$$= \mathbf{w}^T E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]\mathbf{w} \tag{41}$$

$$\Sigma = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]$$

$$\mathrm{Var}(z) = \mathbf{w}^T\Sigma\mathbf{w} \tag{42}$$

$$\max(\mathbf{w}^T\Sigma\mathbf{w}) \quad \text{such that} \quad \|\mathbf{w}\| = 1$$

$$\underset{\mathbf{w}}{\mathrm{argmax}}\,\mathbf{w}^T\Sigma\mathbf{w}$$

We will use Lagrange multipliers:

- **Goal:** Find extrema of $f(x)$.
- **Constraint:** Subject to $g(x) = c$.
- **Optimal Point:** Level curve of $f(x)$ tangent to $g(x) = c$.
- **Parallel Gradients:** $\nabla f(x)$ and $\nabla g(x)$ are parallel.
- **Gradient Direction:** Points to steepest increase.
- **Lagrange Multiplier:** $\nabla f(x) = \lambda \nabla g(x)$.
- **Proportional Gradients:** At optimum, gradients are scalar multiples.

$\nabla g(x)$ and $\nabla f(x)$ are in the same direction, proportionally.

$$\nabla f(x) = \lambda \nabla g(x) \tag{43}$$

$$\nabla f(x) - \lambda \nabla g(x) = 0 \tag{44}$$

Lagrange multiplier (optimize):

$$\mathcal{L} = f(x) - \lambda(g(x) - c) \tag{45}$$

$$\mathcal{L}(\mathbf{w}, \lambda) = \mathbf{w}^T \Sigma \mathbf{w} - \lambda(\mathbf{w}^T \mathbf{w} - 1) \tag{46}$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, \lambda)}{\partial \mathbf{w}} = 0 \tag{47}$$

$$\frac{\partial}{\partial \mathbf{w}}(\mathbf{w}^T \Sigma \mathbf{w} - \lambda(\mathbf{w}^T \mathbf{w} - 1)) = 2\Sigma \mathbf{w} - 2\lambda \mathbf{w} = 0 \tag{48}$$

$$\Sigma \mathbf{w} = \lambda \mathbf{w} \tag{49}$$

Note:

$$\frac{\partial \theta^T A \theta}{\partial \theta} = (A + A^T)\theta \quad (A = A^T) \tag{50}$$

$$= 2A\theta \tag{51}$$

$$\frac{\partial \theta^T \theta}{\partial \theta} = 2\theta \tag{52}$$

$w_1 \longrightarrow \lambda_1,$ largest eigenvector corresponds to largest eigenvalue

$$Z_2 = w_2^T x \quad \text{s.t.} \quad \|w_2\| = 1, \quad w_1^T w_2 = 0 \tag{53}$$

$$\underset{w_2}{\operatorname{argmax}} \quad w_2^T \Sigma w_2 - \lambda(w_2^T w_2 - 1) - \beta(w_2^T w_1) \tag{54}$$

Taking the difference and setting it to 0:

$$2\Sigma w_2 - 2\lambda w_2 - \beta w_1 = 0 \tag{55}$$

$$2w_1^T \Sigma w_2 - 2\lambda w_1^T w_2 - \beta w_1^T w_1 = 0 \tag{56}$$

$$w_1^T w_2 = 0, \quad w_1^T w_1 = 1$$

$$2w_1^T \Sigma w_2 = \beta \tag{57}$$

$$2w_2^T \Sigma w_1 = \beta \tag{58}$$

$$2\lambda w_2^T w_1 = \beta \tag{59}$$

$$xw = z \tag{60}$$

$$\begin{bmatrix} | & | \\ & \\ & \\ | & | \end{bmatrix}_{N\times 100} \begin{bmatrix} | & | \\ w_1 & w_2 \\ | & | \end{bmatrix}_{100\times 2} = \begin{bmatrix} z_1^{(1)} & z_2^{(1)} \\ z_1^{(2)} & z_2^{(2)} \\ \vdots & \vdots \\ z_1^{(N)} & z_2^{(N)} \end{bmatrix}_{N\times 2} \tag{61}$$

3.1 How to choose k?

$$\frac{|\lambda_1| + |\lambda_2| + \cdots + |\lambda_d|}{|\lambda_1| + |\lambda_2| + \cdots + |\lambda_d|} > 0.9 \tag{62}$$

- PCA is unsupervised; it does not consider class labels.
- LDA considers class labels.
- LDA finds a projection **w** that maximizes class mean separation while minimizing within-class variance. The data is then projected onto this lower-dimensional space for classification.
- $m_1 = w^T m_1$
- $m_2 = w^T m_2$
- $\max(m, -m_2)^2$
- $\min(s_1^2 + s_2^2) \Rightarrow \max \frac{1}{s_1^2+s_2^2}$

$$\max \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2} \tag{63}$$

4 Multivariate Methods (from last week)

We have d-dimensional dataset:

$$X \in \mathbb{R}^d \tag{64}$$

$$\mu \in \mathbb{R}^d \quad \text{(mean vector)} \tag{65}$$

$$\Sigma \in \mathbb{R}^{d\times d} \quad \text{(covariance matrix)} \tag{66}$$

Mean vector:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix} = \begin{bmatrix} E[x_1] \\ E[x_2] \\ \vdots \\ E[x_d] \end{bmatrix} \tag{67}$$

Covariance matrix (transpose of covariance matrix is equal to itself):

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1d}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{d1}^2 & \cdots & \sigma_{dd}^2 \end{bmatrix} \tag{68}$$

Data X:

$$X = \begin{bmatrix} x_1^{(1)} & x_d^{(1)} \\ x_1^{(2)} & x_d^{(2)} \\ \vdots & \vdots \\ x_1^{(N)} & x_d^{(N)} \end{bmatrix} \tag{69}$$

The probability density function (PDF) is given by:

$$p(x) = \frac{1}{\sqrt{2\pi}|\Sigma|^{1/2}} \exp\left(-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}\right) \tag{70}$$

We can express the mean vector:

$$\mu_i = \frac{1}{N} \sum_{j=1}^{N} x_i^j \tag{71}$$

$$\Sigma = E[(x-\mu)(x-\mu)^T] \tag{72}$$

$\hat{\mu}$ represents parameters estimated from data:

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (x^{(i)} - \mu)(x^{(i)} - \mu)^T \tag{73}$$

Example:

$$X = \begin{bmatrix} 2 & 9 \\ 4 & 6 \\ 6 & 3 \\ 8 & 1 \\ 10 & 1 \end{bmatrix} \tag{74}$$

1) Mean vector

$$\mu_1 = \frac{1}{5} \sum_{i=1}^{5} x_1^i = 6 \tag{75}$$

$$\mu_2 = \frac{1}{5} \sum_{i=1}^{5} x_2^i = 5 \tag{76}$$

$$\mu^T = \begin{bmatrix} 6 & 5 \end{bmatrix} \tag{77}$$

2) Covariance matrix:

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x^{(i)} - \mu)(x^{(i)} - \mu)^T \tag{78}$$

$$\Sigma = \frac{1}{5}\left(\begin{bmatrix}2-6\\9-5\end{bmatrix}\begin{bmatrix}-4 & 4\end{bmatrix} + \begin{bmatrix}4-6\\7-5\end{bmatrix}\begin{bmatrix}-2 & 2\end{bmatrix} + \cdots + \begin{bmatrix}10-6\\1-5\end{bmatrix}\begin{bmatrix}4 & -4\end{bmatrix}\right) \tag{79}$$

$$\Sigma = \begin{bmatrix}8-8\\-8 \quad 8\end{bmatrix} \tag{80}$$

Also, we can use:

$$\sigma_1^2 = \frac{1}{N}\sum_i (x_1^i - \mu_1) \tag{81}$$

$$\sigma_{12}^2 = \frac{1}{N}\sum_i (x_1^i - \mu_1)(x_2^i - \mu_2) \tag{82}$$

4.0.1 Covariance

$$\rho = \mathrm{Cov}(x_i, x_j) = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \tag{83}$$

Example:

$$\text{Ex} \quad \mathrm{Cov} = \frac{-8}{\sqrt{8}\sqrt{8}} = -1 \tag{84}$$

- If features are independent, then correlation coefficient is 0 $\rho = 0$.
- If x_i, x_j are independent, diagonals are non-zero.

4.0.2 Estimating Curves

When we divide the plot into squares, count samples, create a histogram and we fit a curve to the histogram, it will be bell-shaped.

- If x_1 increases, x_2 increases or decreases. (Left-top plot in the Figure 1)

$$\Sigma = \begin{bmatrix}\sigma^2 & 0\\0 & \sigma^2\end{bmatrix}$$

- If x_1 increases, x_2 increases. Figure (Left-bottom plot in the 1)

$$\Sigma = \begin{bmatrix}\sigma^2 & +\\+ & \sigma^2\end{bmatrix}$$

- If x_1 increases, x_2 decreases. Figure (Right-bottom plot in the 1)

$$\Sigma = \begin{bmatrix}\sigma^2 & -\\\text{-} & \sigma^2\end{bmatrix}$$

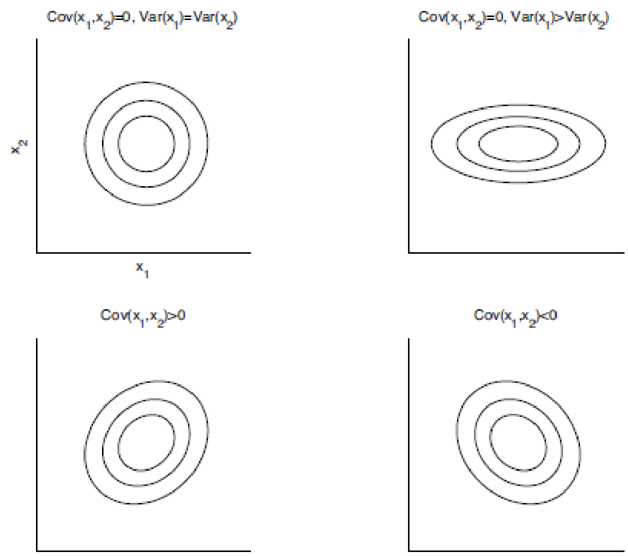

Figure 1: Isoprobability contour plots. Mean is the center, their shapes depend on cov. matrix. Retrieved from [1], page 92.

4.0.3 Multivariate Classification

$$P(c|x) = \frac{P(x|c)P(c)}{P(c)} \tag{85}$$

Numerator $P(x|c=1,\mu_1,\Sigma_1)P(c=1)$ is $g_1(x)$:

$$P(c=1|x) = \frac{P(x|c=1,\mu_1,\Sigma_1)P(c=1)}{P(x)} \tag{86}$$

Numerator $P(x|c=0,\mu_0,\Sigma_0)P(c=0)$ is $g_1(x)$:

$$P(c=0|x) = \frac{P(x|c=0,\mu_0,\Sigma_0)P(c=0)}{P(x)} \tag{87}$$

$$P(c=1|x) \overset{?}{\geq} P(c=0|x) \tag{88}$$

Classifying based on numerators:

$$\text{Class} = \begin{cases} 1 & \text{if } g_1(x) > g_0(x) \\ 0 & \text{if } g_1(x) < g_0(x) \end{cases}$$

Expanding $g_0(x)$ and $g_1(x)$:

$$g_1(x) = \frac{1}{\sqrt{2\pi}|\Sigma_0|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right) P(c=0) \tag{89}$$

$$g_0(x) = \frac{1}{\sqrt{2\pi}|\Sigma_1|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right) P(c=1) \tag{90}$$

We can represent $\frac{1}{\sqrt{2\pi}|\Sigma_1|^{1/2}}$ as C.

For example, if we assume that:

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad x = \begin{bmatrix} 6 \\ 3 \end{bmatrix}, \quad \mu_1 = \begin{bmatrix} 6 \\ 5 \end{bmatrix}, \quad \mu_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$g_1(x) = C\exp\left(-\frac{1}{2}\begin{bmatrix} 0 & -2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ -2 \end{bmatrix}\right) = C\exp(-2) \tag{91}$$

$$g_0(x) = C\exp\left(-\frac{1}{2}\begin{bmatrix} 0 & 3 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ 3 \end{bmatrix}\right) = C\exp\left(-\frac{4.5}{2}\right) \tag{92}$$

Taking the logarithm of both sides:

$$\ln g_0(x) = \ln C - 2 \tag{93}$$

$$\ln g_1(x) = \ln C - \frac{4.5}{2} \tag{94}$$

And it concludes that:

$$\ln g_0(x) > \ln g_1(x) \implies \text{class } 0 \tag{95}$$

If we take the logarithm:

$$g_1(x) = \ln\frac{1}{\sqrt{2\pi}} - \frac{1}{2}\ln|\Sigma_1|^{(1/2)} - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) \tag{96}$$

If we use the distributive property

$$g_1 = -\frac{1}{2}x^T\Sigma_1^{-1}x - x^T\Sigma_1^{-1}\mu_1 - \mu_1\Sigma_1^{-1}x + \mu_1^T\Sigma_1^{-1}\mu_1 \tag{97}$$

$$g_0 = -\frac{1}{2}x^T\Sigma_0^{-1}x - x^T\Sigma_0^{-1}\mu_0 - \mu_0\Sigma_0^{-1}x + \mu_0^T\Sigma_0^{-1}\mu_0 \tag{98}$$

If we assume these, and try to express $g_1(x) - g_0(x)$ in terms of x_1 and x_2:

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad x = \begin{bmatrix} x1 \\ x2 \end{bmatrix}, \quad \mu_1 = \begin{bmatrix} 6 \\ 5 \end{bmatrix}, \quad \mu_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$g_0(x) = -\frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{99}$$

$$g_1(x) = -\frac{1}{2}\begin{bmatrix}x_1 & x_2\end{bmatrix}\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} - \begin{bmatrix}x_1 & x_2\end{bmatrix}\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}6\\5\end{bmatrix} - \begin{bmatrix}6 & 5\end{bmatrix}\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} + \begin{bmatrix}6 & 5\end{bmatrix}\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}6\\5\end{bmatrix} \tag{100}$$

The difference:

$$g_1(x) - g_0(x) = 6x_1 + 5x_2 - \frac{61}{2} \tag{101}$$

References

[1] Ethem Alpaydın, *Introduction to Machine Learning, 2nd Edition, MIT Press* 2010.