CSE 405: Machine Learning

Multivariate Methods

Dr Muhammad Abul Hasan

Department of Computer Science and Engineering
Green University of Bangladesh
muhammad.hasan@cse.green.edu.bd

Fall 2023

Outline

"The noblest pleasure is the joy of understanding.
– Leonardo da Vinci"

Multivariate Data

- Multiple measurements (sensors)
- d inputs/features/attributes: d-variate
- N instances/observations/examples

$$\mathbf{X} = \begin{bmatrix} X_1^1 & X_2^1 & \cdots & X_d^1 \\ X_1^2 & X_2^2 & \cdots & X_d^2 \\ \vdots & \vdots & \ddots & \vdots \\ X_1^N & X_2^N & \cdots & X_d^N \end{bmatrix}$$

Multivariate Parameters

- Mean: $E[\mathbf{x}] = \mu = [\mu_1, \ldots, \mu_d]^T$
- Covariance: $\sigma_{ij} \equiv \text{Cov}(X_i, X_j)$
- Correlation: $\text{Corr}(X_i, X_j) \equiv \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$

$$\Sigma \equiv \text{Cov}(\mathbf{X}) = E[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{22} & \sigma_2^3 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

Parameter Estimation

- Sample mean $\mathbf{m}$: $m_i = \frac{\sum_{t=1}^{N} x_i^t}{N}, i = 1, \ldots, d$
- Covariance Matrix $\mathbf{S}$: $s_{ij} = \frac{\sum_{t=1}^{N} (x_i^t - m_i)(x_j^t - m_j)}{N}$
- Correlation matrix $\mathbf{R}$: $r_{ij} = \frac{s_{ij}}{s_i s_j}$

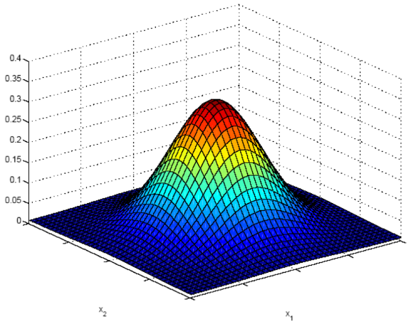
Estimation of Missing Values

- What to do if certain instances have missing attributes?
- Ignore those instances: not a good idea if the sample is small
- Use 'missing' as an attribute: may give information
- Imputation: Fill in the missing value
    - Mean imputation: Use the most likely value (e.g., mean)
    - Imputation by regression: Predict based on other attributes

Multivariate Normal Distribution

$$\mathbf{x} \sim \mathcal{N}_d(\mu, \Sigma)$$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right]$$



Multivariate Normal Distribution

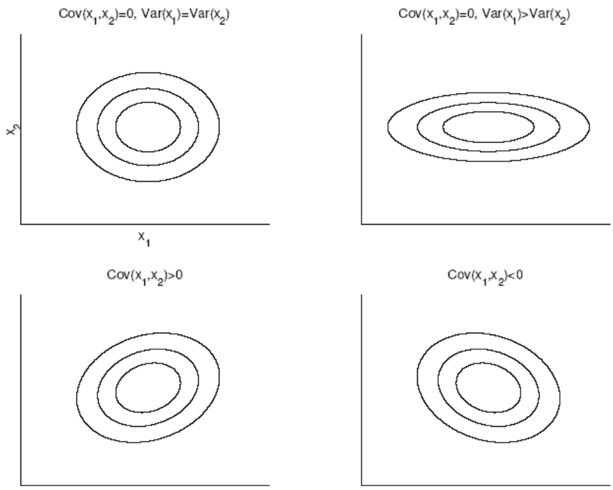
- Mahalanobis distance: $(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)$ measures the distance from x to μ in terms of Σ (normalizes for differences in variances and correlations).
- Bivariate: $d = 2$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

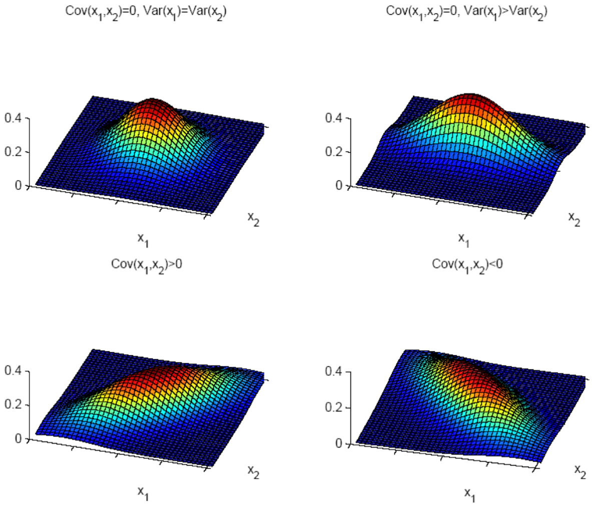
$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp[-\frac{1}{2(1-\rho^2)}(z_1^2 - 2\rho z_1 z_2 + z_2^2)]$$

$$z_i = \frac{(x_i - \mu_i)}{\sigma_i}$$

Bivariate Normal



Bivariate Normal



Independent Inputs: Naive Bayes

- If x_i are independent, offdiagonals of Σ are 0, Mahalanobis distance reduces to weighted (by $1/\sigma_i$) Euclidean distance:

$$p(\mathbf{x}) = \Pi_{i=1}^{d} p_i(x_i) = \frac{1}{(2\pi)^{\frac{3}{2}} \Pi_{i=1}^{d} \sigma_i} \exp[-\frac{1}{2} \Sigma_{i=1}^{d} (\frac{x_i - \mu_i}{\sigma_i})^2]$$

- If variances are also equal, reduces to Euclidean distance.

Parametric Classification

- If $p(\mathbf{x}|C_i) \sim \mathcal{N}(\mu_i, \Sigma_i)$

$$p(\mathbf{x}|C_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{\frac{1}{2}}} \exp[-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i)]$$

- Discriminant functions:

$$\begin{aligned} g_i(\mathbf{x}) &= \log p(\mathbf{x}|C_i) + \log p(C_i) \\ &= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i) + \log P(C_i) \end{aligned}$$

Estimation of Parameters

$$\hat{p}(C_i) = \frac{\Sigma_t r_i^t}{N}$$

$$\mathbf{m}_i = \frac{\Sigma_t r_i^t \mathbf{x}^t}{\Sigma_t r_i^t}$$

$$\mathbf{S}_i = \frac{\Sigma_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^T}{\Sigma_t r_i^t}$$

$$g_i(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{S_i}| - \frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \mathbf{S}_i^{-1}(\mathbf{x} - \mathbf{m}_i) + \log\hat{p}(C_i)$$

Different S_i

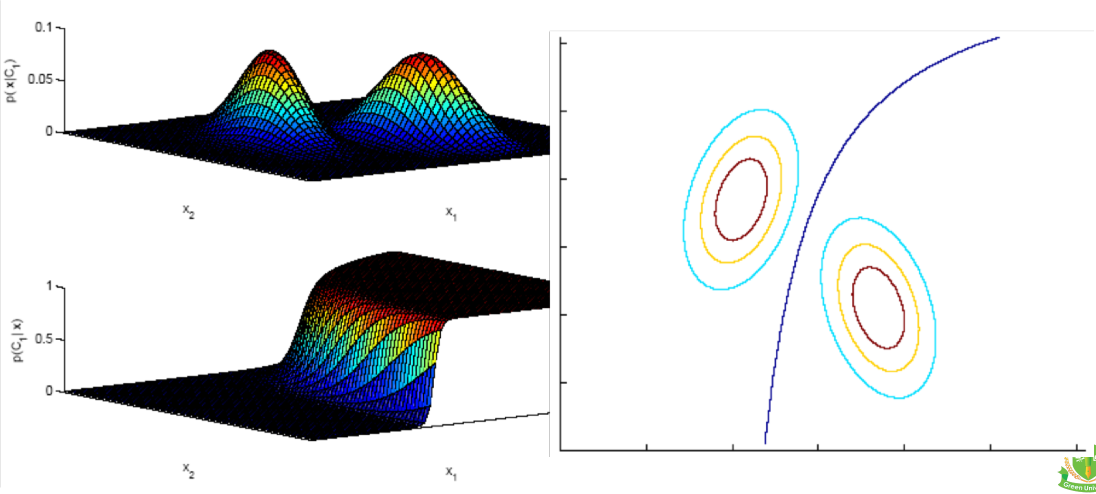
- Quadratic discriminant:

$$
\begin{aligned}
g_i(\mathbf{x}) &= -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}^T\mathbf{S}_i^{-1}\mathbf{x} - 2\mathbf{x}^T\mathbf{S}_i^{-1}\mathbf{m}_i + \mathbf{m}_i^T\mathbf{S}_i^{-1}\mathbf{m}_i) + \log\hat{p}(C_i) \\
&= \mathbf{x}^T\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^T\mathbf{x} + \omega_i0
\end{aligned}
$$

where,

$$
\begin{aligned}
\mathbf{W}_i &= -\frac{1}{2}\mathbf{S}_i^{-1} \\
\mathbf{w}_i &= \mathbf{S}_i^{-1}\mathbf{m}_i \\
\omega_{i0} &= -\frac{1}{2}\mathbf{m}_i^T\mathbf{S}_i^{-1}\mathbf{m}_i - \frac{1}{2}\log|\mathbf{S}_i| + \log\hat{p}(C_i)
\end{aligned}
$$

Different S_i



Common Covariance Matrix S

- Shared common sample covariance S

$$\mathbf{S} = \Sigma_i \hat{p}(C_i)\mathbf{S}_i$$

- Discriminant reduces to

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \mathbf{S}^{-1}(\mathbf{x} - \mathbf{m}_i) + \log \hat{p}(C_i)$$

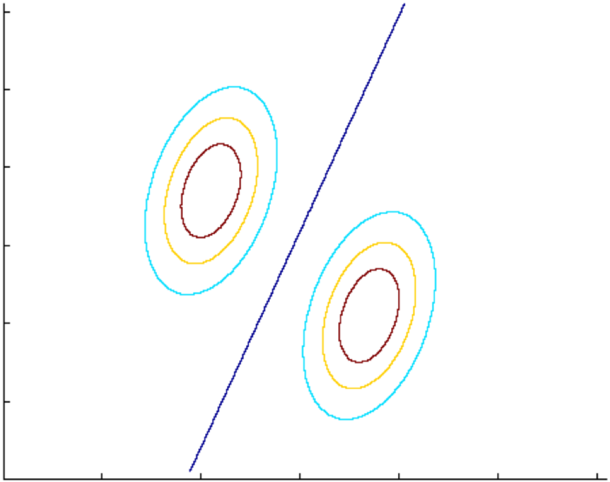
which is a linear discriminant

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + \omega_{i0}$$

where,

$$\mathbf{w}_i = \mathbf{S}^{-1}\mathbf{m}_i \quad \omega_{i0} = -\frac{1}{2}\mathbf{m}_i^T \mathbf{S}^{-1}\mathbf{m}_i + \log \hat{p}(C_i)$$

Common Covariance Matrix S



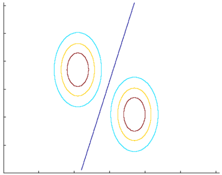
Diagonal S

- When $x_j \quad j = 1, \ldots, d$, are independent, Σ is diagonal $p(\mathbf{x}|C_i) = \Sigma_j p(x_j|C_i)$ (Naive Bayes' assumption)

$$g_i(\mathbf{x}) = -\frac{1}{2}\Sigma_{j=1}^{d}(\frac{x_j^t - m_{ij}}{s_j})^2 + \log \hat{p}(C_i)$$

Classify based on weighted Euclidean distance (in s_j units) to the nearest mean.

Diagonal S



variances may be different

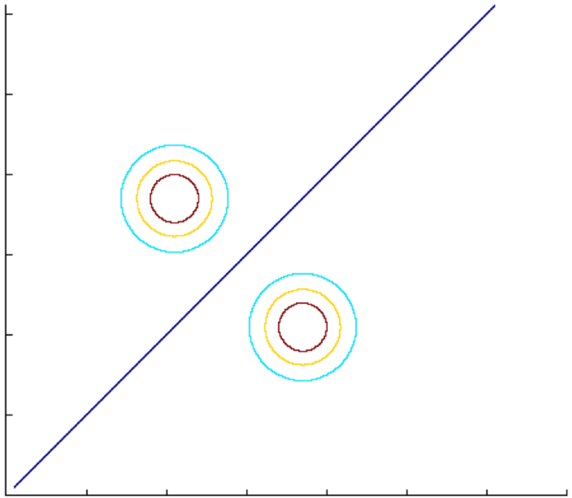
Diagonal S, equal variances

- Nearest mean classifier: Classify based on Euclidean distance to the nearest mean.

$$
\begin{aligned}
g_i(\mathbf{x}) &= -\frac{||\mathbf{x} - \mathbf{m}_i||^2}{2s^2} + \log \hat{p}(C_i) \\
&= -\frac{1}{2s^2} \Sigma_{j=1}^{d} (x_j^t - m_{ij})^2 + \log \hat{p}(C_i)
\end{aligned}
$$

- Each mean can be considered a prototype or template and this is template matching.

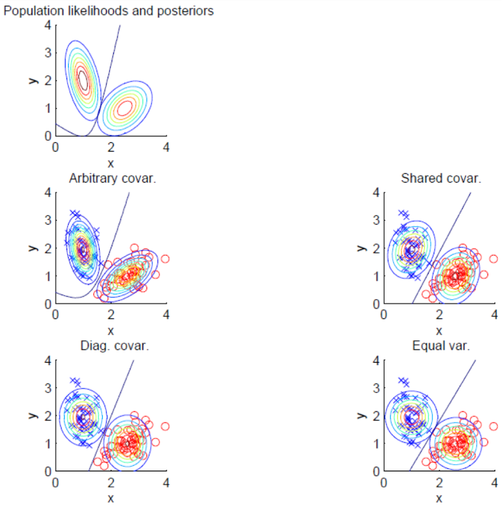
Diagonal S, equal variances



Model Selection

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

- As we increase complexity (less restricted S), bias decreases, and variance increases
- Assume simple models (allow some bias) to control variance (regularization)


Population likelihoods and posteriors
Arbitrary covar.
Shared covar.
Diag. covar.
Equal var.
x
y
0
1
2
3
4

Thank You!