02_Classification_using_Bayes_Theory

2.1 Bayes Decision Theory 贝叶斯决策理论

Basic Assumptions

- The decision problem is posed in probabilistic terms.
- **ALL** relevant probability values are known.

2.1.1 Process

- **Given:**
  1. A test sample x.
     - Contains features $x = [x_1, x_2, \dots x_l]^T$.
     - Often reduced, removed some non-discriminative (un-useful) features.
  2. A list of classes/patterns $\omega = \{\omega_1, \omega_2, \dots \omega_c\}$.
     - Defined by human-being.
  3. A classification method M.
     - A **database** storing multiple samples with the same type of x.
     - Each sample is assigned to an arbitrary class $\omega_{any} \in \{\omega_1, \omega_2, \dots \omega_c\}$.
- **Do:**
  - $\{P(\omega_1|x), \dots, P(\omega_c|x)\} \leftarrow classify(M, x, \omega)$
  - That is, for all the possible classes, find:
    - The probability that the given x belongs to that class.
- **Get:**
  - $\omega_{target}(x) = argmax_i[P(\omega_i|x)], i \in [1, c]$.
  - That is, assign x a class/pattern from ω with the **most probable** one.

Example
MNIST database.

- Test sample:
  - $x =$ A 28×28 grayscale image of a hand-written number.
- Set of classes:
  - ω = $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.
- Classification Method:
  - Derived from 10,000 of 28×28 similar gray-scale images.
- Process:

- Given an image, using the classification method, get a list of probabilities $P(\omega) = \{P(\omega_1), P(\omega_2), \dots, P(\omega_c)\}$.
- Select the ω_i with the largest probability $P(\omega_i)$, that is $selected = argmax[P(\omega_i)]$.

2.1.2 Properties of Variables.

- The set of all classes ω :
  - c available classes: $\omega = \{\omega_1, \omega_2, \dots, \omega_c\}$
- Prior Probabilities $P(\omega) = \{P(\omega_1), P(\omega_2), \dots, P(\omega_c)\}$:
  - Probability Distribution of random variable ω_j in the database.
    - The fraction of samples in the database that belongs to class ω_j.
    - $P(\omega)$ is the prior knowledge on $\omega = \{\omega_1, \omega_2, \dots, \omega_c\}$.
  - It is Non-Negative.
    - $\forall i \in [1, c], P(\omega_i) \geq 0$.
    - The probabilities of all classes are greater-or-equal to 0.
  - It is Normalized.
    - $\sum_{i=1}^{c} P(\omega_i) = 1$.
    - The sum of the prior probabilities of all classes is 1.

2.2 Prior & Posterior Probabilities 先验与后验概率

2.2.1 Definition of Prior Probability 先验概率

- Decision **BEFORE** Observation (Naïve Decision Rule).
  - Don't care about test sample x.
  - Given x, always choose the class that:
    - has the most member in the database.
    - i.e., has the highest prior probability.
- Classification Process:
  1. $\omega = \{\omega_1, \omega_2, \dots, \omega_c\}$.
  2. By counting the number of members $Num(\omega_i)$ for each class $\omega_i \in \omega, i \in [1, c]$, we get the prior probabilities $P(\omega) = \{P(\omega_1), P(\omega_2), \dots, P(\omega_c)\}$.
  3. Then, classify x directly into $argmax_i[P(\omega_i)]$.
- The decision is the same all the time obviously, and the prob. of a right guess is $\frac{1}{c}$.

2.2.2 Definition of Posterior Probability 后验概率

- Decision **WITH** Observation.

- Cares about test sample x.
- Considering x, as well as the prior probabilities $P(\omega) = \{P(\omega_1), P(\omega_2), \ldots, P(\omega_c)\}$,
  - and give x the class with the biggest posterior probability.

- **Posterior Probability:**
  - [DEF] Posterior Probability of a class ω_j on test sample x:
    - Given test sample x, how possible does x could be classified into class ω_j.
  - $P(\omega_j|x) = \dfrac{p(x|\omega_j)P(\omega_j)}{p(x)}$, $Posterior = \dfrac{Likelihood \times Prior}{Evidence}$.
    - $p(x|\omega_j)$: **Likelihood (KNOWN)**
      - The fraction of samples stored in the database that
        - is same to x, and
        - belongs to class ω_j.
    - $P(\omega_j)$: **Prior probability of class ω_j (KNOWN)**
      - The fraction of samples stored in the database that
        - belongs to class ω_j.
    - $p(x)$: **Evidence (IRRELEVANT)**
      - Unconditional density of x.
      - That is, $p(x) = \sum_{j=1}^{c} p(x|\omega_j)P(\omega_j)$.
- **Special Cases:**
  1. Equal Prior Probability.
     - $P(\omega_1) = P(\omega_2) = \ldots = P(\omega_c) = \dfrac{1}{c}$.
     - The amount of members in each class are same.
     - Here, posterior probs. $\forall j \in [1, c], P(\omega_j|x)$ is dependent on the likelihoods $P(x|\omega_j)$ only.
  2. Equal Likelihood.
     - $P(x|\omega_1) = P(x|\omega_2) = \ldots = P(x|\omega_c)$.
     - The amount of members that's same to x in each class are the same.
     - Here, posterior probs. $\forall j \in [1, c], P(\omega_j|x)$ is dependent on the prior probabilities $P(\omega_j)$ only.
     - Back to Naïve Decision Rule.

2.2.3 Classification Examples

Given:

1. Test sample $x \in \{+, -\}$.
2. A list of classes $\omega = \{\omega_1 = cancer, \omega_2 = no\_cancer\}$.
3. Classification Method M, with known probabilities:
   - Prior Probabilities:
   - $P(\omega_1) = 0.008$

- $P(\omega_2) = 1 - P(\omega_1) = 0.992$
- Likelihoods:
- For class $\omega_1 = cancer$: $P(+|\omega_1) = 0.98$, $P(-|\omega_1) = 0.02$
- For class $\omega_2 = no\_cancer$: $P(+|\omega_2) = 0.03$, $P(-|\omega_2) = 0.97$.

Classification:

- Given a test sample $x = +$.
  - The prob. that this person gets cancer is:
    - $P(\omega_1|+) = \frac{P(+|\omega_1) \times P(\omega_1)}{P(+)} = \frac{0.98 \times 0.008}{P(+)} = \frac{0.00784}{P(+)}$.
  - The prob. that this person doesn't gets cancer is:
    - $P(\omega_2|+) = \frac{P(+|\omega_2) \times P(\omega_2)}{P(+)} = \frac{0.03 \times 0.992}{P(+)} = \frac{0.02976}{P(+)}$
  - Therefore, the classification result would be:
    - $\omega_{target} = argmax_i[P(\omega_i|+)]$
      $= argmax_i[\frac{P(+|\omega_i) \times P(\omega_i)}{P(x)}]$
      $= argmax_i[P(+|\omega_i) \times P(\omega_i)]$
      $= \omega_2$, for $0.00784 < 0.02976$
    - That is, no_cancer.

2.3 Loss Functions 决策成本函数

2.3.0 Why do we use loss functions?

- Different selection errors may have differently significant consequences, i.e., "losses" or "costs". 不同决策的成本、后果不同。
  - In pure Naïve Bayes classification, we only consider probability.
  - However,
    - we can tolerate "non-cancer" being classified into "cancer",
    - while it's more lossy to classify "cancer" into "non-cancer".
  - There is a need to consider this kind of "loss" into our decision method.
- We want to know if the Bayes decision rule is optimal.
  - Need a evaluation method
  - calc how many error you make, sum together

2.3.1 Probability of Error

For only two classes:

- If $P(\omega_1|x) > P(\omega_2|x)$, $x \leftarrow \omega_1$. Prob. of error: $P(\omega_2|x)$.
- If $P(\omega_1|x) < P(\omega_2|x)$, $x \leftarrow \omega_2$. Prob. of error: $P(\omega_1|x)$.

2.3.2 Loss Function (i.e., "Cost Function")

Problem

- Take action α_i for a given x.
  - The action α_i: To assign the test pattern x the class ω_i.
- Introduce the loss/cost $\lambda(\alpha_i|\omega_j)$, for the true class ω_j and action α_i on x.
  - That is, $\lambda(\alpha_i|\omega_j)$ is the cost of classifying **any** sample into class ω_i when the true class of that sample is ω_j.
  - For instance, $\lambda(\alpha_{cancer}|\omega_{no\_cancer})$ is the cost of diagnosing a patient that actually doesn't have cancer as "having cancer".
    - (Which by intuition is not as serious as its reverse, therefore the value of this λ should also be lower than its reverse.)
- We don't actually know the true class ω_j for a random sample x, so we use the Expected Loss.
  - That is, we consider the "average loss" of classifying x into ω_i by considering:
    - The loss of classifying x into ω_j for all $\omega_j \in \omega$.
    - The probability that $x \in \omega_j$, i.e., $P(\omega_j|x)$.

[DEF]Expected Loss (Average Loss, Conditional Risk) 期望成本:

- The expected loss of classifying x into ω_i.
- $R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) \times P(\omega_j|x)$, where
  - $\lambda(\alpha_i|\omega_j)$: The cost of classifying x into ω_i under the true class ω_j.
  - $P(\omega_j|x)$: The posterior probability that x belongs to class ω_j.
    - Computed during the Naïve Bayes Classification with $P(\omega_j)$ and $P(x|\omega_j)$.

[DEF]Bayes Risk 贝叶斯风险:

- The modified measurement of the original Bayes Rule.
  - Consider the importance of each error.
  - Consider minimum loss, instead of maximum probability.
- Bayes Risk finds the action that gives the minimum expected loss of x.
  - $\alpha(x) = argmin_{\alpha_i \in A} R(\alpha_i|x)$
    - $= argmin_{\alpha_i \in A} \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|x)$

Derivation: For a 2-class problem

- Known:
  - Test sample x.
  - Classes $\omega = \{\omega_1, \omega_2\}$.
  - The calculated posterior probabilities:
    - $P(\omega_1|x)$, $P(\omega_2|x)$.

- Loss Matrix: $\begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix}$, where $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$.
  - λ_{ij}: The cost of classifying x into ω_i when the true class of x is ω_j.
- $\omega_{target} = argmin_{\alpha_i \in A} R(\alpha_i|x)$
- If we choose ω_1, we have:
  - $R(\alpha_1|x) < R(\alpha_2|x)$
  - $\iff \lambda_{11}P(\omega_1|x) + \lambda_{12}P(\omega_2|x) < \lambda_{21}P(\omega_1|x) + \lambda_{22}P(\omega_2|x)$
  - $\iff (\lambda_{21} - \lambda_{11})P(\omega_1|x) > (\lambda_{12} - \lambda_{22})P(\omega_2|x)$
  - $\iff \dfrac{P(\omega_1|x)}{P(\omega_2|x)} > \dfrac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}}$
  - $\iff \dfrac{P(x|\omega_1)P(\omega_1)}{P(x|\omega_2)P(\omega_2)} > \dfrac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}}$
  - $\iff \dfrac{P(x|\omega_1)}{P(x|\omega_2)} > \dfrac{(\lambda_{12} - \lambda_{22})P(\omega_2)}{(\lambda_{21} - \lambda_{11})P(\omega_1)}$
  - $\iff \dfrac{P(x|\omega_1)}{P(x|\omega_2)} > \theta_t$

2.3.3 Examples

Minimum Prob. Error and Minimum Risk

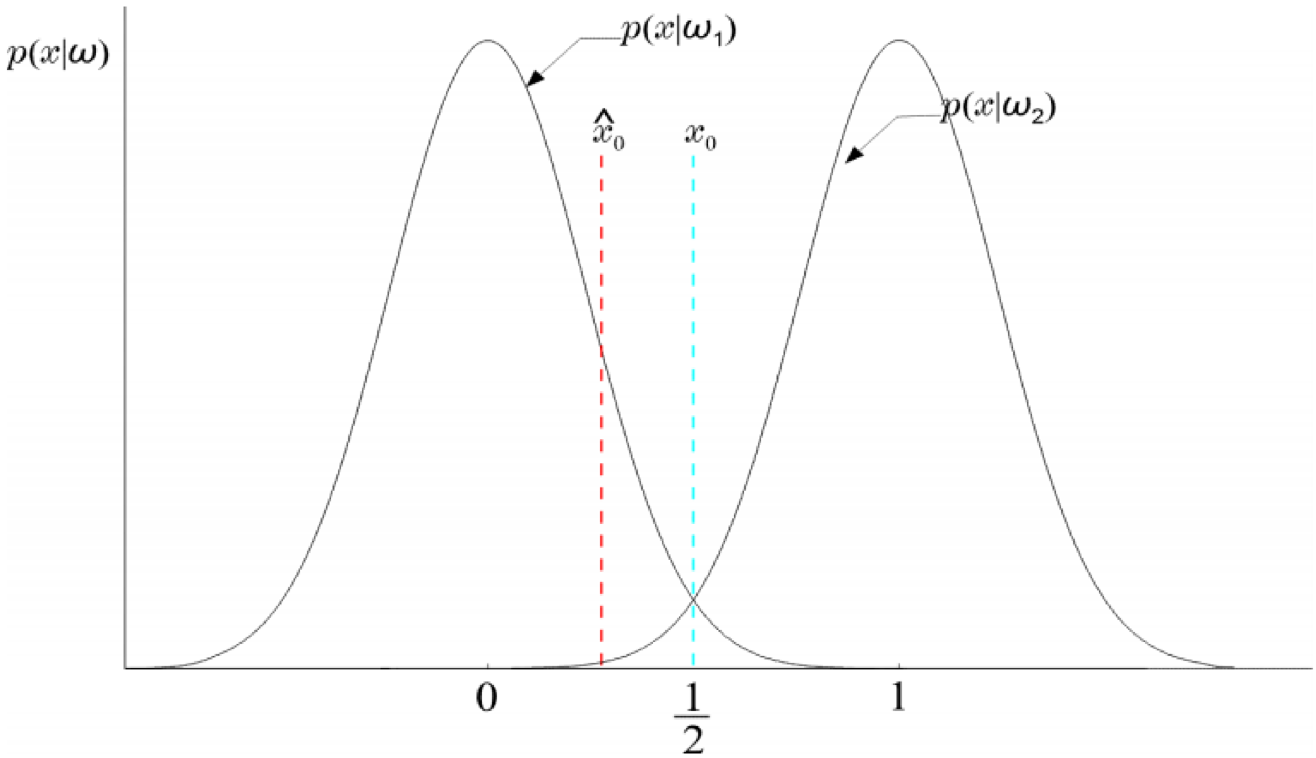
Remark: Gaussian Distribution

- $GD(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}$

Given:

- Two probability distributions of evidence $P(x|\omega_j)$ regarding $j \in \{1, 2\}$.
  - $P(x|\omega_1) = \dfrac{1}{\sqrt{\pi}} e^{-x^2}$, where $\mu = 0, \sigma = \dfrac{1}{\sqrt{2}}$.
  - $P(x|\omega_2) = \dfrac{1}{\sqrt{\pi}} e^{-(x-1)^2}$, where $\mu = 1, \sigma = \dfrac{1}{\sqrt{2}}$.
- Loss matrix:
  - $\begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} = \begin{bmatrix} 0 & 1.0 \\ 0.5 & 0 \end{bmatrix}$
  Do:
- The threshold x_0 for minimum P_e.
  - $P(x_0|\omega_1) = P(x_0|\omega_2)$
    - $\implies \dfrac{1}{\sqrt{\pi}} e^{-x_0^2} = \dfrac{1}{\sqrt{\pi}} e^{-(x_0-1)^2}$
    - $\implies x_0 = -x_0 + 1$, omitting $x_0 = x_0 - 1$ which is impossible;
    - $\implies x_0 = \dfrac{1}{2}$
- The threshold $\hat{x_0}$ for minimum $R(\alpha_i|x)$.

- $R(\alpha_1|x) = R(\alpha_2|x)$
  - $\implies \frac{P(\hat{x_0}|\omega_1)}{P(\hat{x_0}|\omega_2)} = \frac{(\lambda_{12} - \lambda_{22})P(\omega_2)}{(\lambda_{21} - \lambda_{11})P(\omega_1)}$
  - $\implies \frac{P(\hat{x_0}|\omega_1)}{P(\hat{x_0}|\omega_2)} = \frac{(1-0)\times\frac{1}{2}}{(0.5-0)\times\frac{1}{2}}$
  - $\implies P(\hat{x_0}|\omega_1) = 2P(\hat{x_0}|\omega_2)$
  - $\implies \frac{1}{\sqrt{\pi}}e^{-\hat{x_0}^2} = 2\frac{1}{\sqrt{\pi}}e^{-(\hat{x_0}-1)^2}$
  - $\implies -\hat{x_0}^2 = \ln 2 - \hat{x_0}^2 + 2\hat{x_0} - 1$
  - $\implies \hat{x_0} = \frac{1-\ln 2}{2} < \frac{1}{2}$



Minimum Error Rate Classification

- A zero-one loss function
  - $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
  - All errors are equally costly.
- Conditional Risk:
  - $R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|x)P(\omega_j|x)$
  - $= \lambda(\alpha_i|\omega_i)P(\omega_i|x) + \sum_{j\neq i} \lambda(\alpha_i|\omega_j)P(\omega_j|x)$
  - $= 0 + \sum_{j\neq i} 1 \times P(\omega_j|x)$
  - $= \sum_{j\neq i} P(\omega_j|x)$
  - $= 1 - P(\omega_i|x)$

2.4 Discriminant Functions 判别函数

2.4.1 Definition of Discriminant Function

- If a function f satisfies:
  - If $f(\cdot)$ monotonically increases, and
  - $\forall i \neq j, f(P(\omega_i|x)) > f(P(\omega_j|x))$, then
  - $x \rightarrow \omega_i$
- Then, $g_i(x) = f(P(\omega_i|x))$ is a discriminant function.
- That is, this function is able to "tell" a certain one ω_i from others on any input x. 给定一个测试样本x，判别函数能够从所有其它分类中挑选一个最可能的ω_j。
  - i.e., it separates ω_i and $\neg\omega_i$.

2.4.2 Property of Discriminant Function

1. One function per class.
   1. A discriminant function is able to "tell" a certain one ω_i specifically for any input x.
2. Various discriminant functions $\rightarrow$ Identical classification results. 样式各异，结果相同。
   1. It is correct to say, the discriminant functions:
      1. **Preserves** the original monotonical-increase of its inputs.
      2. But changes the changing rate by **processing** the inputs.
   2. i.e.,
      1. "$\forall i \neq j, f(g_i(x)) > f(g_j(x)) \wedge f \nearrow$ "and "$\forall i \neq j, g_i(x) > g_j(x)$" are equivalent in decision.
      2. Changing growth rate of input:
         1. $f(g_i(x)) = k \cdot g_i(x)$, a linear change.
         2. $f(g_i(x)) = \ln g_i(x)$, a log change, i.e., it grows, but slower as it proceed.
      3. Therefore, the discriminant function may vary, but the output is always the same.
3. Examples of discriminant functions:
   1. Minimum Risk: $g_i(x) = -R(\alpha_i|x) = -\lambda(\alpha_i|x) \times P(\omega_i|x)$, for $i \in [1, c]$
   2. Minimum Error Rate: $g_i(x) = P(\omega_i|x)$, for $i \in [1, c]$

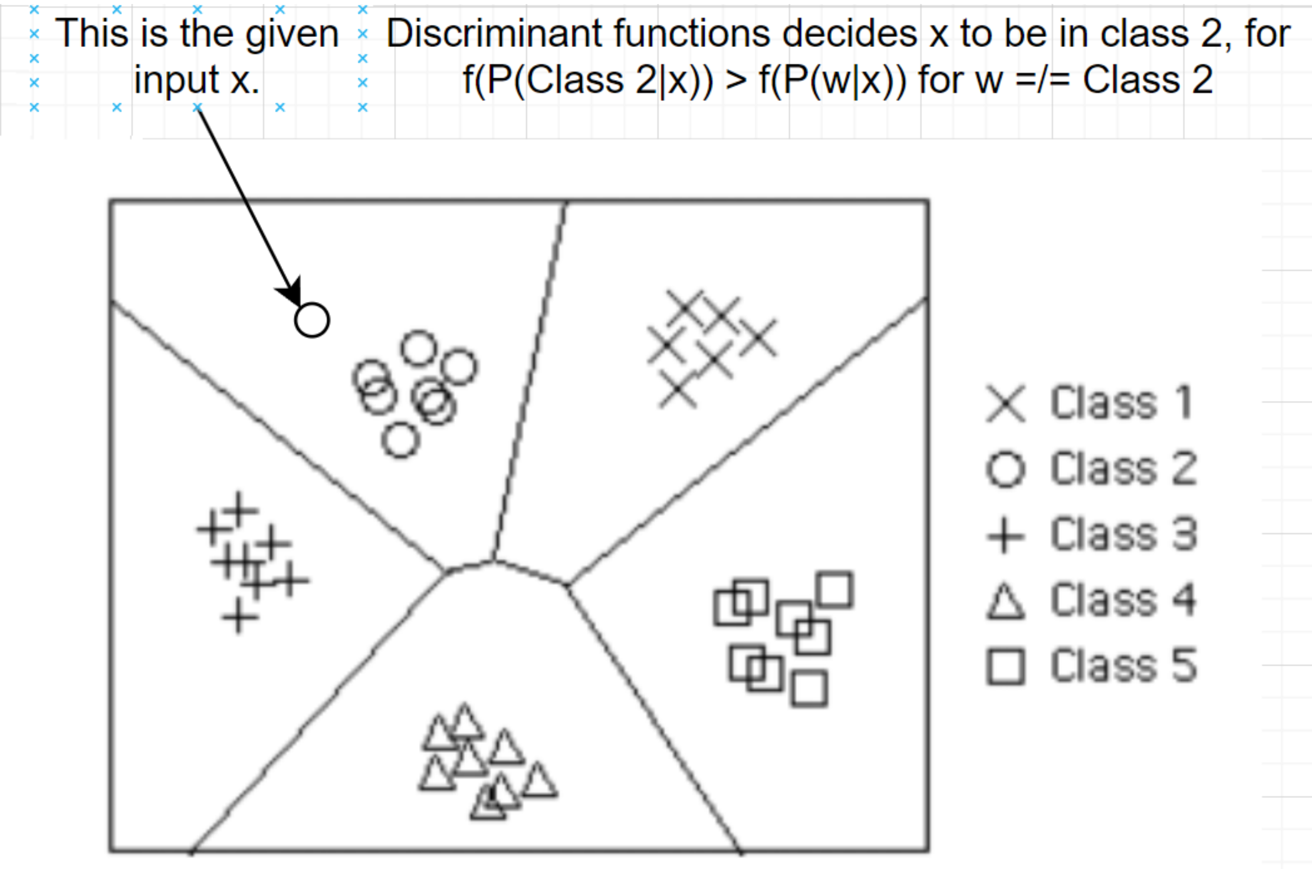
2.4.3 Decision Region 决策区域

- c discriminant functions $\implies$ c decision regions
  - $g_i(x) \implies R_i \subset R^d, i \in [1, c]$
- One function per decision region that is distinct and mutual-exclusive.
  - $R_i = \{x | x \in R^d : \forall i \neq j, g_i(x) > g_j(x)\}$, where

- $\forall i \neq j, R_i \cap R_j = \emptyset$, and $\cap_{i=1}^{c} R_i = R^d$

2.4.4 Decision Boundaries 决策边界

- "Surface" in feature space, where ties occur among 2 or more largest discriminant functions.
- x_0 is on the decision boundary/surface if and only if
  - $\exists \omega_i, \omega_j \in \omega, g_i(x_0) = g_j(x_0)$.



2.5 Bayesian Classification for Normal Distributions

2.5.1 Multi-Dimensional Normal Distribution 高维正态分布

1-D Case 多类别，一维数据

- There are several classes:
  - Each class has its own distribution of data samples, i.e., each class has its own μ and σ.
- For a specific class, there are plenty of data samples:

- Each sample is a **scalar**, that is a 1×1 "matrix", which is a "plain number".
- The samples follows a **Normal Distribution**.

For a specific class ω_i, suppose the data conforms a normal distribution. Here:

- $x \sim N(\mu_i, \sigma_i) : P(x|\omega_i) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$, where
  - μ is the mean value.
    - $\mu_i = E(x)$
  - σ^2 is the variance.
    - $\sigma_i = E[(x-\mu)^2]$

Multivariate Case 多类别，高维数据

- There are several classes:
  - Each class has its own distribution of data samples, i.e., each class has its own μ and σ.
- For a specific class, there are plenty of data samples:
  - Each sample is a **vector**, that is a $d \times 1$ matrix, where d is the dimension of data.
  - The samples follow a d**-dimensional Normal Distribution**.

Here, for a specific class ω_i, suppose the multi-dimensional data X conforms a normal distribution.

- $X \sim N(\mu_i, \Sigma_i) : P(X|\omega_i) = \frac{1}{|\Sigma_i|^{\frac{1}{2}} \times (2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}(X-\mu_i)^\top \Sigma_i^{-1}(X-\mu_i)}$
  - Regular Variables:
    - d-dimensional random variables: $X = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_d \end{bmatrix}$;
    - d-dimensional mean vector: $\mu_i = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \dots \\ \mu_{id} \end{bmatrix} = \begin{bmatrix} E(x_{i1}) \\ E(x_{i2}) \\ \dots \\ E(x_{id}) \end{bmatrix}$, specifically for class ω_i;
    - $d \times d$ covariance matrix: $\Sigma_i = \begin{pmatrix} \sigma_{i11} & \sigma_{i12} & \cdots & \sigma_{i1d} \\ \sigma_{i21} & \sigma_{i22} & \cdots & \sigma_{i2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{id1} & \sigma_{id2} & \cdots & \sigma_{idd} \end{pmatrix} = E[(X-\mu_i)(X-\mu_i)^\top]$, specifically for class ω_i.

- Explanations on $-\frac{1}{2}(X-\mu_i)^\top \Sigma_i^{-1}(X-\mu_i)$
  - Parts:
    - $(X-\mu_i)^\top = \begin{bmatrix} x_1-\mu_{i1} \\ x_2-\mu_{i2} \\ \cdots \\ x_d-\mu_{id} \end{bmatrix}^\top = [(x_1-\mu_{i1}) \quad (x_2-\mu_{i2}) \quad \cdots \quad (x_d-\mu_{id})]$
    - $\Sigma_i^{-1} = \begin{pmatrix} \sigma'_{i11} & \sigma'_{i12} & \cdots & \sigma'_{i1d} \\ \sigma'_{i21} & \sigma'_{i22} & \cdots & \sigma'_{i2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma'_{id1} & \sigma'_{id2} & \cdots & \sigma'_{idd} \end{pmatrix}$, the inverse of the covariance matrix.
    - $(X-\mu_i) = \begin{bmatrix} x_1-\mu_{i1} \\ x_2-\mu_{i2} \\ \cdots \\ x_d-\mu_{id} \end{bmatrix}$
  - Whole:
    - $-\frac{1}{2}(X-\mu_i)^\top \Sigma_i^{-1}(X-\mu_i)$
    - $= -\frac{1}{2}[(x_1-\mu_{i1}) \quad (x_2-\mu_{i2}) \quad \cdots \quad (x_d-\mu_{id})] \begin{pmatrix} \sigma'_{i11} & \sigma'_{i12} & \cdots & \sigma'_{i1d} \\ \sigma'_{i21} & \sigma'_{i22} & \cdots & \sigma'_{i2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma'_{id1} & \sigma'_{id2} & \cdots & \sigma'_{idd} \end{pmatrix} \begin{bmatrix} x_1 - \\ x_2 - \\ \cdot \\ x_d - \end{bmatrix}$
    - $= -\frac{1}{2}[a_1 \quad a_2 \quad \cdots a_d] \begin{bmatrix} x_1-\mu_{i1} \\ x_2-\mu_{i2} \\ \cdots \\ x_d-\mu_{id} \end{bmatrix}$
    - $= y \geq 0$

Example: 2-D Case

- $X \sim N(\mu, \Sigma) : P(X) = \frac{1}{|\Sigma_i|^{\frac{1}{2}} \times (2\pi)} e^{-\frac{1}{2}[(x_1-\mu_{i1}) \quad (x_2-\mu_{i2})]\Sigma_i^{-1}\begin{bmatrix}(x_1-\mu_{i1}) \\ (x_2-\mu_2)\end{bmatrix}}$
  - 2 - dimensional random variable X: $X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
  - 2 - dimensional mean vector: $\mu_i = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \end{bmatrix} = \begin{bmatrix} E(x_{i1}) \\ E(x_{i2}) \end{bmatrix}$
  - 2×2 covariant matrix Σ_i:
    - $\Sigma_i = E[(X-\mu_i)(X-\mu_i)^\top]$
    - $= E(\begin{bmatrix} x_1-\mu_{i1} \\ x_2-\mu_{i2} \end{bmatrix} [x_1-\mu_{i1} \quad x_2-\mu_{i2}])$
    - $= \begin{bmatrix} (x_1-\mu_{i1})^2 & (x_1-\mu_{i1})(x_2-\mu_{i2}) \\ (x_2-\mu_{i2})(x_1-\mu_{i1}) & (x_2-\mu_{i2})^2 \end{bmatrix}$
    - $= \begin{bmatrix} \sigma_1^2 & \sigma \\ \sigma & \sigma_2^2 \end{bmatrix}$

2.5.2 Minimum-error-rate classification

Recall:

- Minimum-error-rate means that we ignore the "cost" of each decision.
- In other words, we only select the classes based on probabilities.

Pattern of Discriminant Function

- Discriminant Function: $g_i(x) = \ln P(\omega_i|x), \forall i \in [1, c] \cap \mathbb{N}^+$
  - $g_i(x) = \ln[P(\omega_i|x)]$
  - $\implies g_i(x) = \ln[P(X|\omega_i) \times P(\omega_i)]$
  - $\implies g_i(x) = \ln[P(X|\omega_i)] + \ln[P(\omega_i)]$
  - $\implies g_i(x) = \ln[\frac{1}{|\Sigma|^{\frac{1}{2}} \times (2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)}] + \ln[P(\omega_i)]$
  - $\implies g_i(x) =$
    - $-\frac{d}{2}\ln(2\pi)$
    - $-\frac{1}{2}|\Sigma_i|$
    - $-\frac{1}{2}(X-\mu_i)^T\Sigma_i^{-1}(X-\mu_i)$
    - $+\ln[P(\omega_i)]$
- Here, $-\frac{d}{2}\ln(2\pi)$ is a constant, which can be ignored. The discriminant function is then updated as:
  - $g_i(x) =$
    - $-\frac{1}{2}\ln|\Sigma_i|$
    - $-\frac{1}{2}(X-\mu_i)^\top\Sigma_i^{-1}(X-\mu_i)$
    - $+\ln[P(\omega_i)]$

Case I: $\Sigma_i = \sigma^2 I$

- That is, $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_{|\omega|} = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$
  - All the classes have a common covariance matrix $\sigma^2 I$.
  - A diagonal matrix suggests that the distribution of data in is **isotropic** (各向同性的), with respect to any specific class.
    - That is, the variance or spread is the same in all directions.

        - In other words, there is no directional preference in the spread of the distribution.
- Therefore, we have:
    - $|\Sigma_i| = \sigma^{2d}$
    - $\Sigma_i^{-1} = \frac{1}{\sigma^2}I$
- And the discriminant function $g_i(x)$ is:
    - $g_i(x) =$
        - $-\frac{1}{2}|\Sigma_i|$
        - $-\frac{1}{2}(X-\mu_i)^T\Sigma_i^{-1}(X-\mu_i)$
        - $+\ln[P(\omega_i)]$
- Here, as $|\Sigma_i| = \sigma^{2d}$ is a constant, it is ignored. Therefore,
    - $g_i(x) =$
        - $-\frac{1}{2}(X-\mu_i)^\top\Sigma_i^{-1}(X-\mu_i)$
        - $+\ln[P(\omega_i)]$
    - $= -\frac{1}{2}(X-\mu_i)^\top \ \times \ [\frac{1}{\sigma^2}I] \ \times \ (X-\mu_i) \ + \ \ln[P(\omega_i)],$
    - $= -\frac{(X-\mu_i)^\top(X-\mu_i)}{2\sigma^2} + \ln[P(\omega_i)],$
    - $= -\frac{(X^\top-\mu_i^\top)(X-\mu_i)}{2\sigma^2} + \ln[P(\omega_i)],$
    - $= -\frac{X^\top X - X^\top\mu_i - \mu_i^\top X + \mu_i^\top\mu_i}{2\sigma^2} + \ln[P(\omega_i)],$
    - $= -\frac{X^\top X - 2\mu_i^\top X + \mu_i^\top\mu_i}{2\sigma^2} + \ln[P(\omega_i)]$, known that $a^\top b = b^\top a$
        - $= -\frac{||X-\mu_i||^2}{2\sigma^2} + \ln[P(\omega_i)]$, where $||\cdot||$ is the **Euclidean Distance**.
- Here, we ignore $X^\top X$ because it is the same for any class. (Remember X is just the random variable that needs us to classify.)
    - $g_i(x) = -\frac{-2\mu_i^\top X + \mu_i^\top\mu_i}{2\sigma^2} + \ln[P(\omega_i)]$, with $X^\top X$ ignored.
    - $= \frac{\mu_i^\top X}{\sigma^2} - \frac{\mu_i^\top\mu_i}{2\sigma^2} + \ln[P(\omega_i)]$
    - $= (\frac{\mu_i}{\sigma^2})^\top X + (-\frac{\mu_i^\top\mu_i}{2\sigma^2} + \ln[P(\omega_i)])$
    - $= w_i^T X + w_{i0}$, where
        - $w_i = \frac{\mu_i}{\sigma^2} = \begin{bmatrix} \frac{\mu_{i1}}{\sigma^2} \\ \frac{\mu_{i2}}{\sigma^2} \\ \vdots \\ \frac{\mu_{id}}{\sigma^2} \end{bmatrix}$ is the weight vector, and
        - $w_{i0} = (-\frac{\mu_i^\top\mu_i}{2\sigma^2} + \ln[P(\omega_i)])$ is the threshold/bias scalar.

- We have got a **Linear Discriminant Function**.
- Having the discriminant functions defined, we get the decision surface by:
  - $g_i(X) - g_j(X) = 0$
  - $\implies w_i X + w_{i0} - (w_j X + w_{j0}) = 0$
  - $\implies \frac{\mu_i}{\sigma^2} X + w_{i0} - (\frac{\mu_j}{\sigma^2} X + w_{jo}) = 0$
  - $\implies (\frac{\mu_i - \mu_j}{\sigma^2}) X + (w_{i0} - w_{j0}) = 0$
  - $\implies (\mu_i - \mu_j) X + \sigma^2 (w_{i0} - w_{j0}) = 0$

Case II: $\Sigma_i = \Sigma$

- That is, $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_{|\omega|} = \Sigma$
  - All the classes have a common covariance matrix Σ.
  - More general than Case I.
- And the discriminant function $g_i(x)$ is:
  - $g_i(x) =$
    - $-\frac{1}{2} |\Sigma_i|$
    - $-\frac{1}{2} (X - \mu_i)^\top \Sigma_i^{-1} (X - \mu_i)$
    - $+ \ln[P(\omega_i)]$
- Here, as $|\Sigma_i| = |\Sigma|$ is a constant, it is ignored. Therefore,
  - $g_i(x) = -\frac{1}{2} (X - \mu_i)^\top \Sigma^{-1} (X - \mu_i) + \ln P(\omega_i)$
    - where $(X - \mu_i)^\top \Sigma^{-1} (X - \mu_i)$ is the **Squared Mahalanobis Distance**.
    - When $\Sigma = I$, it reduces to **Euclidean Distance**.
  - $= -\frac{1}{2} (X - \mu_i)^\top (\Sigma^{-1} X - \Sigma^{-1} \mu_i) + \ln P(\omega_i)$
  - $= -\frac{1}{2} (X^\top - \mu_i^\top)(\Sigma^{-1} X - \Sigma^{-1} \mu_i) + \ln P(\omega_i)$
  - $= -\frac{1}{2} (X^\top \Sigma^{-1} X - X^\top \Sigma^{-1} \mu_i - \mu_i^\top \Sigma^{-1} X + \mu_i^\top \Sigma^{-1} \mu_i) + \ln P(\omega_i)$
  - $= -\frac{1}{2} (X^\top \Sigma^{-1} X - 2\mu_i^\top \Sigma^{-1} X + \mu_i^\top \Sigma^{-1} \mu_i) + \ln P(\omega_i)$
- Here, $X^\top \Sigma^{-1} X$ is the same for all class, thus can be ignored.
  - $g_i(x) = (\mu_i^\top \Sigma^{-1}) X + (-\frac{\mu_i^\top \Sigma^{-1} \mu_i}{2} + \ln P(\omega_i))$

Case III: $\Sigma_i = arbitrary$

In most cases, for each class ω_i, Σ_i, the covariance/spread of data in this class is arbitrary.

- $g_i(x) = -\frac{1}{2} (X - \mu_i)^\top \Sigma_i^{-1} (X - \mu_i) - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)$

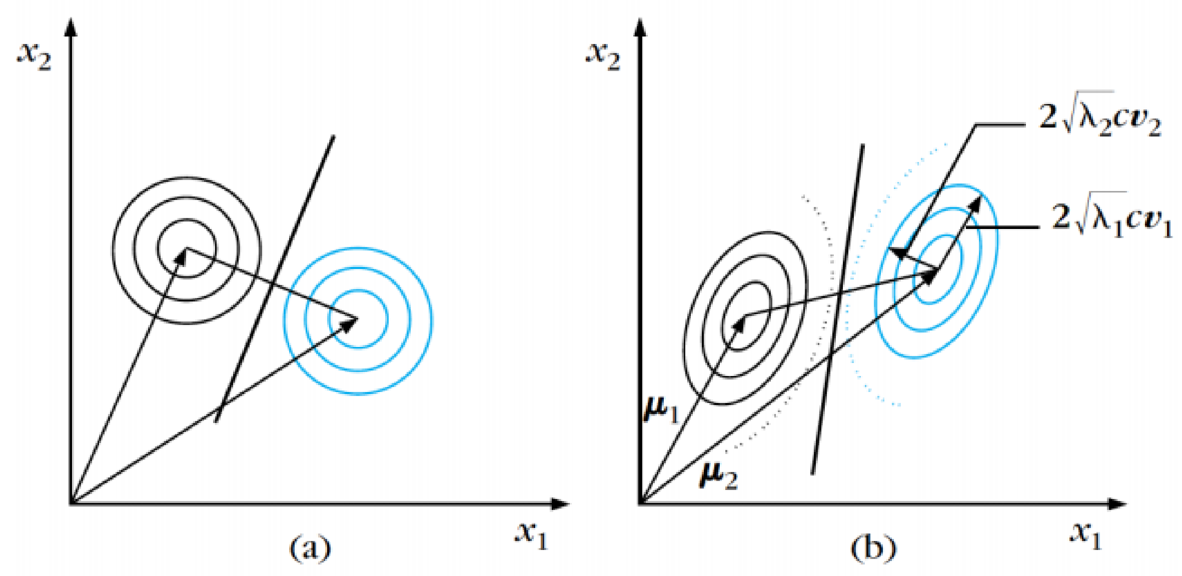
- $= -\frac{1}{2}(X-\mu_i)^\top(\Sigma_i^{-1}X - \Sigma_i^{-1}\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$
- $= -\frac{1}{2}(X^\top-\mu_i^\top)(\Sigma_i^{-1}X - \Sigma_i^{-1}\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$
- $= -\frac{1}{2}(X^\top\Sigma_i^{-1}X - X^\top\Sigma_i^{-1}\mu_i - \mu_i^\top\Sigma_i^{-1}X + \mu_i^\top\Sigma_i^{-1}\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$
- $= -\frac{1}{2}(X^\top\Sigma_i^{-1}X - 2\mu_i^\top\Sigma_i^{-1}X + \mu_i^\top\Sigma_i^{-1}\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$
- $= X^\top(-\frac{1}{2}\Sigma_i^{-1})X + (\mu_i^\top\Sigma_i^{-1})X + (-\frac{\mu_i^\top\Sigma^{-1}\mu_i}{2} - \frac{\ln|\Sigma_i|}{2} + \ln P(\omega_i))$

  Thus,
- $g_i(X) = X^\top W_i X + w_i^\top X + w_{i0}$, where
  - $W_i = -\frac{1}{2}\Sigma_i^{-1}$ is the Quadratic matrix.
  - $w_i = \mu_i^\top\Sigma_i^{-1}$ is the Weight Vector
  - $w_{i0} = -\frac{\mu_i^\top\Sigma^{-1}\mu_i}{2} - \frac{\ln|\Sigma_i|}{2} + \ln P(\omega_i)$ is the Threshold/Bias.

Again, for special covariance matrices:

- $\Sigma_i = \sigma^2 I$:
  - Assign x to ω_i if there is a smaller **Euclidean Distance:** $d_{Euclidean} = \|X-\mu_i\|$
- $\Sigma_i = \Sigma$:

  - Assign x to ω_i if there is a smaller **Mahalanobis Distance:**
  $d_{Mahalanobis} = \sqrt{(X-\mu_i)^\top\Sigma^{-1}(X-\mu_i)}$



2.6 (Additional) Geometric Description of Covariance Matrix

2.6.1 Meta Matrices

Take the 2-D case as an example. Geometrically, the covariance matrix tells how the original Euclidean Space could be transformed into a Mahalanobis space.

The transformation info of a 2-D covariance matrix could be described as:

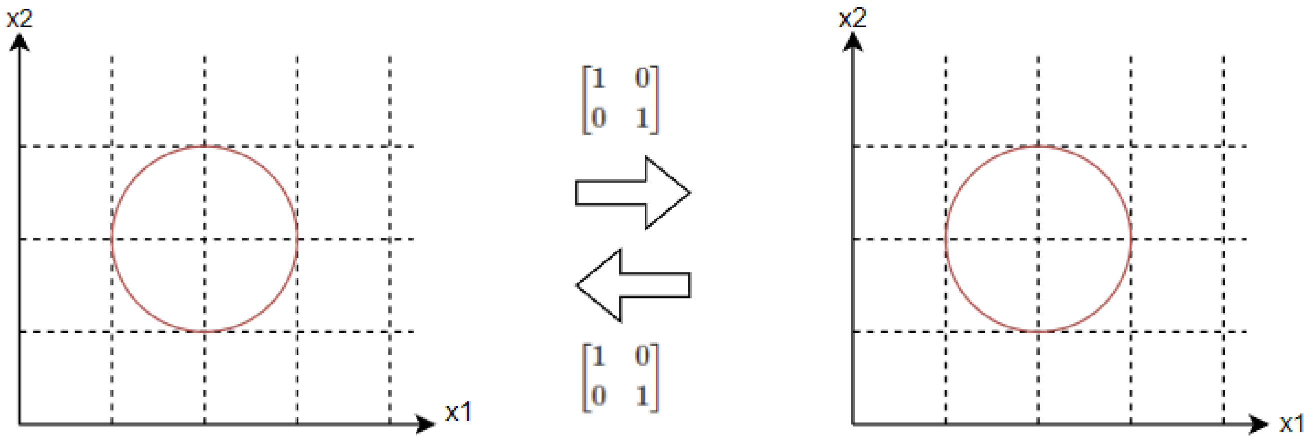
- Scale factor: Multiplication
- Skew factor: Addition

$$\begin{bmatrix} x_1 \; scale \; factor & x_1 \; skew \; factor \\ x_2 \; skew \; factor & x_2 \; scale \; factor \end{bmatrix}$$

1. Identity Matrix

In the trivial case, I^2 as the covariance matrix does no effect on the original Euclidean space. The inverse of the identity matrix is itself.

$\Sigma = \Sigma^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

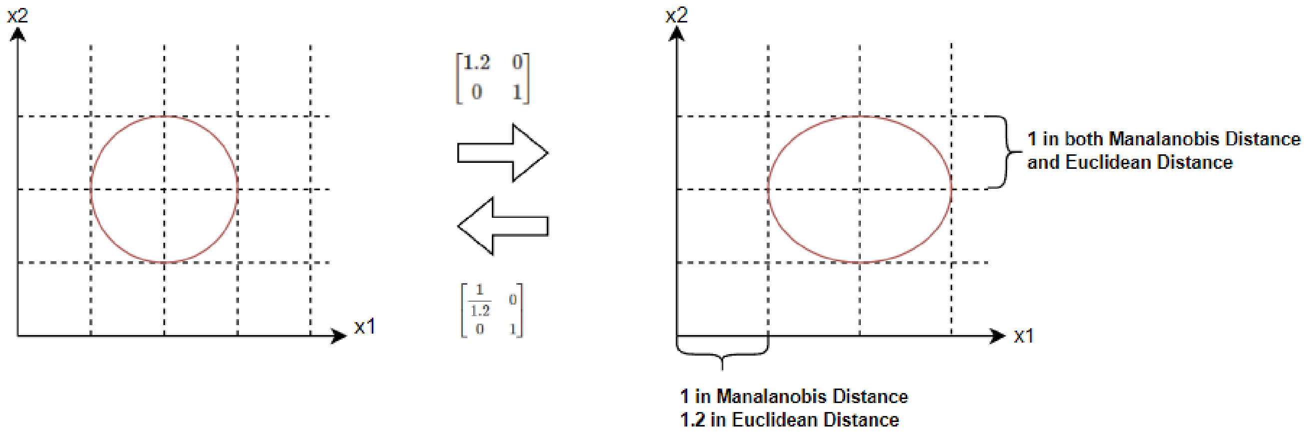


2. Scales x_1

By setting the x_1 scale factor non-1, the matrix scales the Euclidean space on the x_1 axis. The inverse of covariance matrix does the opposite, that is to scale a coordinate back.

$\Sigma = \begin{bmatrix} 1.2 & 0 \\ 0 & 1 \end{bmatrix}$ $\Sigma^{-1} = \begin{bmatrix} \frac{1}{1.2} & 0 \\ 0 & 1 \end{bmatrix}$

For $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, $\Sigma \times v = \begin{bmatrix} 1.2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 1.2v_1 \\ v_2 \end{bmatrix}$

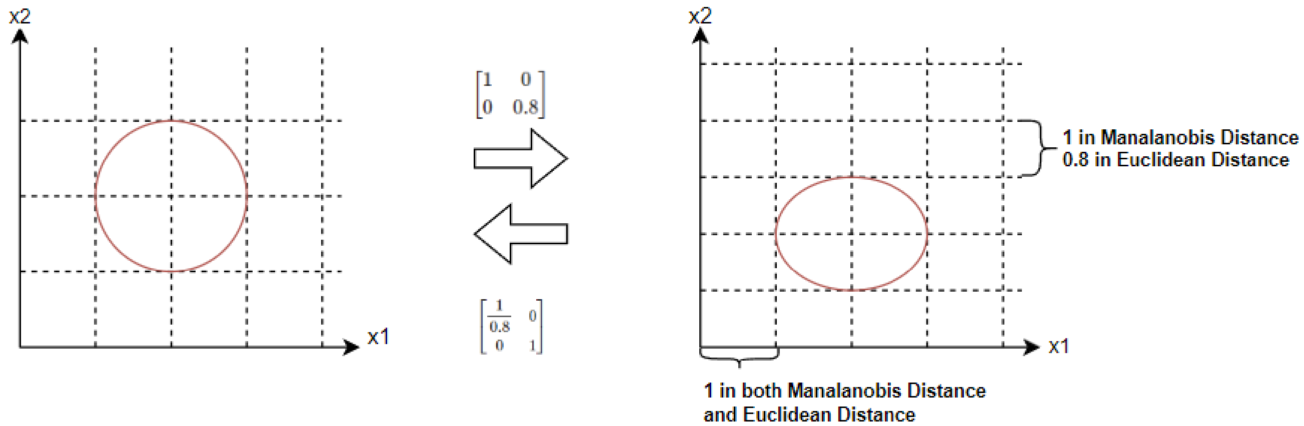


3. Scales x_2

By setting the x_2 scale factor non-1, the matrix scales the Euclidean space on the x_2 axis. The inverse of covariance matrix does the opposite, that is to scale a coordinate back.

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 0.8 \end{bmatrix} \Sigma^{-1} = \begin{bmatrix} \frac{1}{0.8} & 0 \\ 0 & 1 \end{bmatrix}$$

For $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, $\Sigma \times v = \begin{bmatrix} 1 & 0 \\ 0 & 0.8 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} v_1 \\ 0.8v_2 \end{bmatrix}$



4. Skews x_1

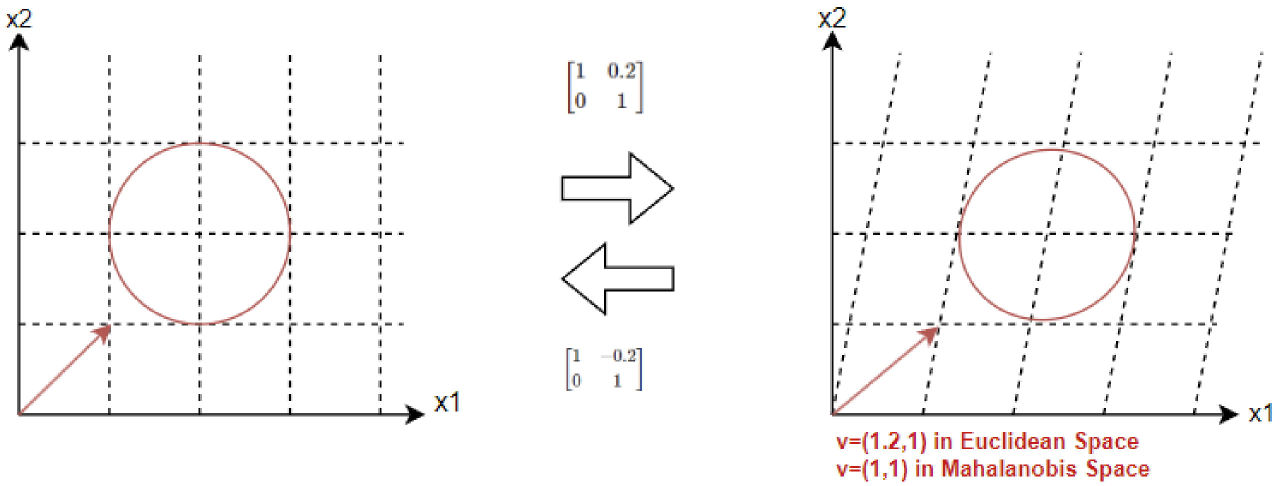
By setting the x_1 skewing factor non-zero, the matrix pans (平移) the x_1 coordinate of a vector by the multiplication of:

- The factor
- And the x_2 coordinate of that vector.

Therefore, the larger x_2 coordinate the vector has, the more it is panned.

$$\Sigma = \begin{bmatrix} 1 & 0.2 \\ 0 & 1 \end{bmatrix} \Sigma^{-1} = \begin{bmatrix} 1 & -0.2 \\ 0 & 1 \end{bmatrix}$$

For $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, $\Sigma \times v = \begin{bmatrix} 1 & 0.2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} v_1 + 0.2v_2 \\ v_2 \end{bmatrix}$



5. Skews x_2

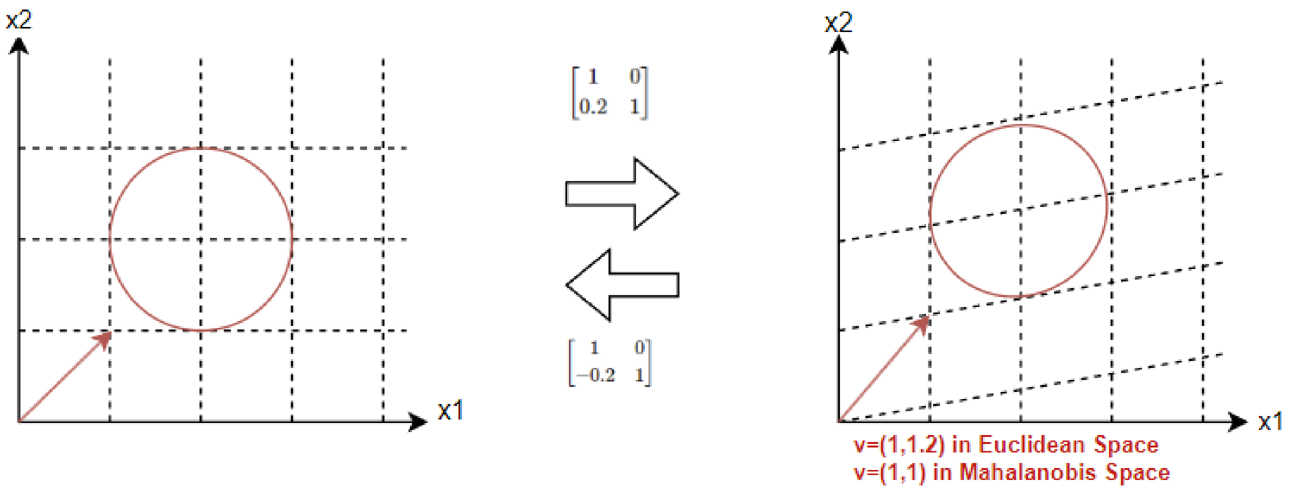
By setting the x_2 skewing factor non-zero, the matrix pans (平移) the x_2 coordinate of a vector by the multiplication of:

- The factor
- And the x_1 coordinate of that vector.

Therefore, the larger x_1 coordinate the vector has, the more it is panned.

$\Sigma = \begin{bmatrix} 1 & 0 \\ 0.2 & 1 \end{bmatrix}$ $\Sigma^{-1} = \begin{bmatrix} 1 & 0 \\ -0.2 & 1 \end{bmatrix}$

For $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, $\Sigma \times v = \begin{bmatrix} 1 & 0 \\ 0.2 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} v_1 \\ 0.2v_1 + v_2 \end{bmatrix}$



2.6.2 Combined Matrices

Several meta matrices could be combined to a single covariance matrix, performing batch operations.

$$\begin{bmatrix}1.2 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 0.8\end{bmatrix}\begin{bmatrix}1 & 0.2\\ 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0.2 & 1\end{bmatrix}$$

$$= \begin{bmatrix}1.2 & 0\\ 0 & 0.8\end{bmatrix}\begin{bmatrix}1 & 0.2\\ 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0.2 & 1\end{bmatrix}$$

$$= \begin{bmatrix}1.2 & 0.24\\ 0 & 0.8\end{bmatrix}\begin{bmatrix}1 & 0\\ 0.2 & 1\end{bmatrix}$$

$$= \begin{bmatrix}1.2 & 0.24\\ 0 & 0.8\end{bmatrix}\begin{bmatrix}1 & 0\\ 0.2 & 1\end{bmatrix}$$

$$= \begin{bmatrix}1.248 & 0.24\\ 0.16 & 0.8\end{bmatrix}$$

Calculation of Squared Mahalanobis Distance

$d_M^2 = (X-\mu_i)^\top \Sigma^{-1}(X-\mu_i) = (X-\mu_i)^\top [\Sigma^{-1}(X-\mu_i)]$

The inverse of the covariance matrix Σ^{-1} reverses the transformed space back to its original Euclidean Space to compute the Mahalanobis distance.