设x = [x1 X2xn]为待分类的特征向量，它已知所有类的有分类条件的高斯pdf。EE6222最小加权距离分类器由以下ω类的判别函数描述：

2. Let $\boldsymbol{x} = [x_1 \quad x_2 \quad \cdots \quad x_n]^T$ be the feature vector to be classified, which has known class-conditional Gaussian PDFs for all classes. A minimum weighted distance classifier is described by the following discriminant function of class ω_i:

$$g_i(\boldsymbol{x}) = -\sum_{j=1}^{n}(x_j - \mu_{ij})^2/\sigma_{ij}^2 + c_i$$

where μ_{ij} and σ_{ij}^2 are the class-conditional mean and variance of x_j for class ω_i. The most frequently applied linear classifier can be formulated by the following discriminant function of class ω_i:

其中μi和o为类ω下xy的类条件均值和方差。最常用的线性分类器可以用以下ω类的判别函数来表示：

$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i{}^T\boldsymbol{x} + b_i.$$

(a) Derive the conditions of $\boldsymbol{x}$ under which the minimum weighted distance classifier minimizes the probability of the misclassification or the error rate.

导出最小加权距离分类器的条件，最小化错误分类的概率或错误率。 (10 Marks)

(b) Derive the conditions of $\boldsymbol{x}$ and determine the weighting vector $\boldsymbol{w}_i$, with which the linear classifier minimizes the probability of the misclassification or the error rate. (**Hint:** You may directly use some intermediate results you derived in part (a).) (8 Marks)

导出x的条件，确定加权向量wi，使线性分类器的误分类概率或错误率最小化。(提示:您可以直接使用部分(a)中导出的一些中间结果。)

(c) Compare the complexity of the above two classifiers in the training phase and the classification phase.

比较上述两种分类器在训练阶段和分类阶段的复杂度。 (7 Marks)

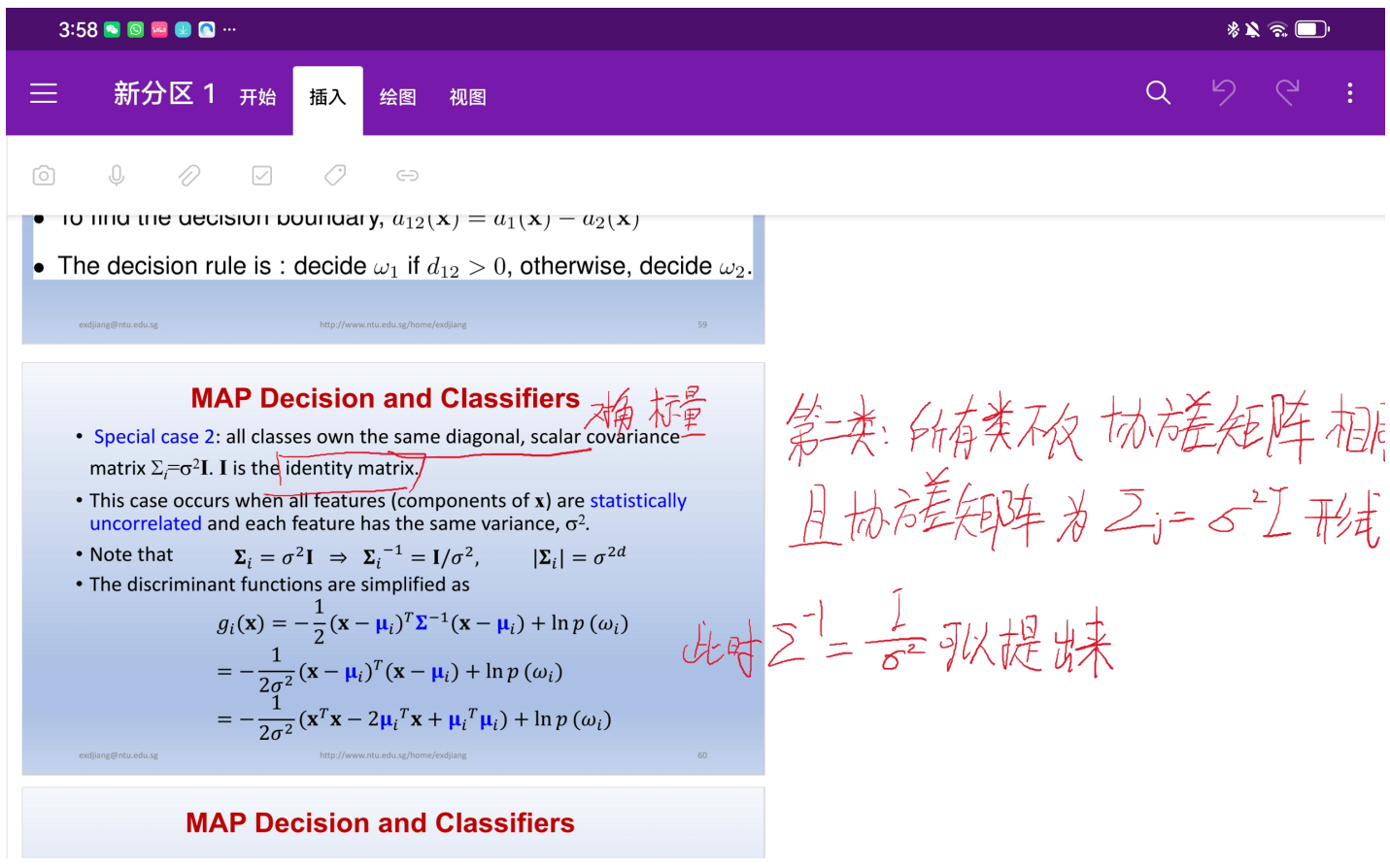
Hint: The general expression of Gaussian PDF is given by:

高斯PDF的一般表达式为：

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

它题目前两问的意思都是要找x满足的条件

题述了x满足高斯分布，然后PPT上有那个gi（x）高斯分布下用来比较大小的判别式，以及这个判别式在两种情况下的简化形式



题目给的第一个判别式满足第二类特殊情况，所以我觉得它就是想让大家说明一下这个情况下X满足的条件

应该是所有类的协方差矩阵相同且都为对角阵，对角线上的元素为分母上那个西塔方

助教-update

(a) $P(e_i|x) = \int_{-\infty}^{+\infty} P(e_i|x)P(x)dx$

$= \int_{-\infty}^{+\infty} [1 - P(w_k|x)]P(x)dx$

$= 1 - \int_{-\infty}^{+\infty} P(w_k)P(x|w_k)dx$

$\ln P(w_k)P(x|w_k) = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) + \ln p(w_i) - \frac{1}{2}\ln|\Sigma_i| - \frac{d}{2}\ln 2\pi$

assuming $c_i = \ln P(w_i) - \frac{1}{2}\ln|\Sigma_i| - \frac{d}{2}\ln 2\pi$ ~~and if x is discrete~~

It can be simplified into

$$g_i(x) = -\sum_{j=1}^{n}(x_j - \mu_{ij})^2/\sigma_j + c_i$$

$w_k = \arg\min_{w_i} P(e_i|x) \Rightarrow \arg\max_{w_i}[P(w_i)P(x|w_i)] \Rightarrow \arg\max_{w_i} \ln P(w_i)P(x|w_i)$

So the decision rule is

$w_k = \arg\max_{w_i} g_i(x)$

The class ~~w_i~~ is the class i with max $g_i(x)$

(b) ~~$g_i(x) = g_j(x) \Rightarrow$~~ $g_i(x) = -\frac{1}{2}x^T\Sigma^{-1}x + \mu_i^T\Sigma^{-1}x - \frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \ln p(w_i)$

Drop terms that are independent to class label i

$g_i(x) = \mu_i^T\Sigma^{-1}x - \frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \ln p(w_i)$

$= w_i^T x + w_{i0}$

$w_i = \mu_i^T\Sigma^{-1}$

~~In This~~

$w_k = \arg\min_{w_i} P(e_i|x)$

The class is the class with ~~A~~ max $g_i(x)$

(c) ~~Rei~~ Design / train phase

training: Linear : Compute μ & Σ $o(nd^2)$

inverse of Σ $o(d^3)$

$o(nd^2 + d^3)$

minimum weight distance classifier & : $wx+b$ $o(d^3)$

Online phase: linear : $wx+b$ $o(d)$ minimum: $O(d^2)$

助教

(a) $P(e_i|x) = \int_{-\infty}^{+\infty} P(e_k|x)P(x)dx$

$= \int_{-\infty}^{+\infty} [1 - P(w_k|x)]P(x)dx$

$= 1 - \int_{-\infty}^{+\infty} P(w_k)p(x|w_k)dx$

$\ln P(w_k)p(x|w_k) = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) + \ln p(w_i) - \frac{1}{2}\ln|\Sigma_i| - \frac{d}{2}\ln 2\pi$

assuming $c_i = \ln P(w_i) - \frac{1}{2}\ln|\Sigma_i| - \frac{d}{2}\ln 2\pi$ ~~and if x is discre~~

It can be simplified into

$$g_i(x) = -\sum_{j=1}^{n}(x_j-\mu_{ij})^2/\sigma_j + c_i$$

$w_k = \arg\min_{w_i} P(e_i|x) \Rightarrow \arg\max_{w_i}[P(w_i)P(x|w_i)] \Rightarrow \arg\max_{w_i} \ln P(w_i)P(x|w_i)$

So the decision rule is

$w_k = \arg\max_{w_i} g_i(x)$

The class ~~k~~ is the class i with max $g_i(x)$

(b) ~~$g_i(x) = g_j(x)$~~ $\Rightarrow g_i(x) = -\frac{1}{2}x^T\Sigma^{-1}x + \mu_i^T\Sigma^{-1}x - \frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \ln p(w_i)$

Drop terms that are independent to class label i

$g_i(x) = \mu_i^T\Sigma^{-1}x - \frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \ln p(w_i)$

$= w_i^T x + w_{i0}$

$w_i = \mu_i^T\Sigma^{-1}$

~~In This~~

$w_k = \arg\min_{w_i} P(e_i|x)$

The class is the class with max $g_i(x)$

群里

Let $\boldsymbol{x} = [x_1 \quad x_2 \quad \cdots \quad x_n]^T$ be the feature vector to be classified, which has known class-conditional Gaussian PDFs for all classes. A minimum weighted distance classifier is described by the following discriminant function of class ω_i:

$$g_i(\boldsymbol{x}) = -\sum_{j=1}^{n}(x_j - \mu_{ij})^2/\sigma_{ij}^2 + c_i$$

where μ_{ij} and σ_{ij}^2 are the class-conditional mean and variance of x_j for class ω_i. The most frequently applied linear classifier can be formulated by the following discriminant function of class ω_i:

$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i{}^T\boldsymbol{x} + b_i.$$

(a) Derive the conditions of $\boldsymbol{x}$ under which the minimum weighted distance classifier minimizes the probability of the misclassification or the error rate.

(10 Marks)

(b) Derive the conditions of $\boldsymbol{x}$ and determine the weighting vector $\boldsymbol{w}_i$, with which the linear classifier minimizes the probability of the misclassification or the error rate. (**Hint:** You may directly use some intermediate results you derived in part (a).)

(8 Marks)

(c) Compare the complexity of the above two classifiers in the training phase and the classification phase.

(7 Marks)

Hint: The general expression of Gaussian PDF is given by:

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

(b)

$\Sigma_i = \Sigma_j = \Sigma$ 不同 w_i，有相同的 Σ

$W_i = \mu_i^T \Sigma^{-1}$

$$W_i = [\mu_{i1}\ \mu_{i2} \cdots \mu_{in}] \begin{bmatrix} \sigma_{i1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{i2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{in}^2 \end{bmatrix}$$

$$= [\mu_{i1}\sigma_{i1}^2\ \ \mu_{i2}\sigma_{i2}^2 \cdots \mu_{in}\sigma_{in}^2]$$

(c)

complexity of training

linear > distance

complexity of classification

linear < distance

(a) $$P(w|x) = \frac{P(x|w)\,P(w)}{P(x)}$$

$$\ln P(w|x) = \ln p(x|w) + \ln p(w) - \ln P(x)$$

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) + \ln p(w_i) - \frac{1}{2}\ln|\Sigma_i^{-1}|$$

$$g_i(x) = -(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) + 2\ln p(w_i) - \ln|\Sigma_i^{-1}|$$

当 $\Sigma_i^{-1} = \begin{bmatrix} \sigma_{i1}^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_{i2}^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_{i3}^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_{in}^2 \end{bmatrix}$ 满足最小 $g_i(x)$

∴ 要满足 X 的特征向量相互独立

满足 $\forall i \neq j \quad P_{ij} = 0$

(a) 推导最小加权距离分类器最小化误分类概率的条件

GPT

为了最小化误分类概率，分类器必须实现贝叶斯决策规则，即选择使后验概率 $P(\omega_i|x)$ 最大的类别。对于高斯分布的类条件概率密度函数，贝叶斯判别函数由对数似然函数导出：

$$g_i(x) = \ln P(x|\omega_i) + \ln P(\omega_i)。$$

假设特征是独立的，且所有类别的类条件概率密度函数都是具有已知均值 μ_{ij} 和方差 σ_{ij}^2 的高斯分布，那么联合概率密度函数 $P(x|\omega_i)$ 可以表示为各个特征概率密度函数的乘积：

$$P(x|\omega_i) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left(-\frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2}\right)。$$

取对数：

$$\ln P(x|\omega_i) = -\sum_{j=1}^{n} \left(\frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2} + \frac{1}{2}\ln(2\pi\sigma_{ij}^2)\right)。$$

加上先验概率 $\ln P(\omega_i)$，判别函数为：

$$g_i(x) = -\sum_{j=1}^{n} \frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2} - \frac{1}{2}\sum_{j=1}^{n} \ln(2\pi\sigma_{ij}^2) + \ln P(\omega_i)。$$

忽略对所有类别都相同的常数项（因为它们不影响决策），我们可以简化为：

$$g_i(x) = -\sum_{j=1}^{n} \frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2} + c_i,$$

其中 $c_i = -\frac{1}{2}\sum_{j=1}^{n} \ln \sigma_{ij}^2 + \ln P(\omega_i)$。

将其与最小加权距离分类器的判别函数比较：

$$g_i(x) = -\sum_{j=1}^{n} \frac{(x_j - \mu_{ij})^2}{\sigma_{ij}^2} + c_i,$$

可以注意到，在求和项中存在一个 $\frac{1}{2}$ 的差异。由于将判别函数乘以一个正的常数不会影响最终的决策，因此分类器在以下条件下有效地实现了贝叶斯决策规则：

- **特征统计独立性：** 给定类别 ω_i，特征 x_j 是统计独立的。
- **高斯类条件概率密度函数：** 类条件概率密度函数是具有已知均值 μ_{ij} 和方差 σ_{ij}^2 的高斯分布。
- **协方差矩阵为对角矩阵：** 协方差矩阵 Σ_i 是对角矩阵，意味着类别内特征之间没有相关性。
- **包含先验概率：** 常数项 c_i 适当地包含了先验概率 $P(\omega_i)$。

在这些条件下，最小加权距离分类器通过有效地计算负的马氏距离并在类别之间进行比较，从而最小化了误分类的概率。

(b) 推导条件并确定线性分类器的权重向量 w_i，使其最小化误分类概率

为了使线性分类器最小化误分类概率，它必须在判别函数为线性的条件下符合贝叶斯决策规则。当类条件概率密度函数具有**相等的协方差矩阵** $\Sigma_i = \Sigma$ 时，判别函数将变为线性。

从一般的高斯判别函数开始：

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

展开二次型：

$$\begin{aligned} g_i(x) &= -\frac{1}{2}\left(x^\top \Sigma^{-1} x - 2x^\top \Sigma^{-1}\mu_i + \mu_i^\top \Sigma^{-1}\mu_i\right) + \ln P(\omega_i) \\ &= -\frac{1}{2}x^\top \Sigma^{-1} x + x^\top \Sigma^{-1}\mu_i - \frac{1}{2}\mu_i^\top \Sigma^{-1}\mu_i + \ln P(\omega_i)。 \end{aligned}$$

由于项 $-\frac{1}{2}x^\top \Sigma^{-1} x$ 对所有类别都是相同的，可以忽略。判别函数简化为：

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

确定权重向量 w_i 和偏置 b_i：

$$w_i = \Sigma^{-1}\mu_i, \qquad b_i = -\frac{1}{2}\mu_i^\top \Sigma^{-1}\mu_i + \ln P(\omega_i)。$$

因此，线性判别函数为：

$$g_i(x) = w_i^\top x + b_i。$$

使线性分类器最小化误分类概率的条件：

- **相等的协方差矩阵：** 对所有类别，$\Sigma_i = \Sigma$，确保二次项抵消，判别函数为线性。
- **高斯类条件概率密度函数：** 类条件概率密度函数是具有已知均值 μ_i 和共享协方差矩阵 Σ 的高斯分布。
- **权重向量和偏置的确定：**
  - $w_i = \Sigma^{-1}\mu_i$
  - $b_i = -\frac{1}{2}\mu_i^\top \Sigma^{-1}\mu_i + \ln P(\omega_i)$

在这些条件下，线性分类器以线性形式实现了贝叶斯决策规则，最小化了误分类的概率。

(c) 比较上述两个分类器在训练阶段和分类阶段的复杂度

训练阶段的复杂度：

- **最小加权距离分类器：**
  - **需要估计：** 对于 c 个类别和 n 个特征，计算 $c \times n$ 个均值 μ_{ij} 和方差 σ_{ij}^2。
  - **复杂度：** 估计均值和方差需要 $O(c \times n)$ 的操作。
- **线性分类器：**
  - **需要估计：**
    - 计算 c 个均值向量 μ_i，每个大小为 n（复杂度 $O(c \times n)$）。
    - 估计共享的协方差矩阵 Σ，大小为 $n \times n$（复杂度 $O(N \times n^2)$，其中 N 为训练样本总数）。
    - **矩阵求逆：** 计算 Σ^{-1} 来得到 $w_i = \Sigma^{-1}\mu_i$（复杂度 $O(n^3)$）。

分类阶段的复杂度：

- **最小加权距离分类器：**
  - **对每个样本、每个类别的计算：**
    - n 次减法：$x_j - \mu_{ij}$。
    - n 次除法：$(x_j - \mu_{ij})/\sigma_{ij}$。
    - n 次平方：$\left(\frac{x_j - \mu_{ij}}{\sigma_{ij}}\right)^2$。
    - n 次加法：对特征求和。
  - **每个类别的总复杂度：** $O(n)$ 次操作。
- **线性分类器：**
  - **对每个样本、每个类别的计算：**
    - 计算点积 $w_i^\top x$ 需要 n 次乘法和 n 次加法。
    - 加上偏置项 b_i。
  - **每个类别的总复杂度：** $O(n)$ 次操作，但由于每个特征的操作较少，通常比最小加权距离分类器的计算量更小。

总体比较：

- **训练阶段：**
  - **最小加权距离分类器** 的复杂度较低，只需要对每个特征和类别计算均值和方差。
  - **线性分类器** 的复杂度较高，因为需要估计协方差矩阵并进行矩阵求逆，尤其当特征数量 n 很大时，计算量更大。
- **分类阶段：**
  - **线性分类器** 计算效率更高，每个类别只需要点积和加法。
  - **最小加权距离分类器** 由于每个特征需要额外的操作（平方和除法），在分类时计算效率较低。

结论：

- **最小加权距离分类器** 在训练阶段复杂度较低，但在分类阶段计算量更大。
- **线性分类器** 在训练阶段复杂度较高（主要由于协方差矩阵的估计和求逆），但在分类阶段计算效率更高。
- 选择哪种分类器可能取决于应用需求，例如是优先考虑更快的训练还是更快的分类。
