They are identical if $\hat{P}_j$ = constant and $|\hat{\underline{\underline{C}}}_j|$ = constant

Important: N_j (no of training examples for class ω_j) large enough for

- $\hat{\underline{\underline{C}}}_j$ invertible : $N_j \gg d$ else $\hat{\underline{\underline{C}}}_j$ is rank deficient.
  Hence NECESSARY CONDITION
- good estimate for $\hat{\underline{\mu}}_j$, $\hat{\underline{\underline{C}}}_j$: $N_j \gg d$

4.3.3 Naive Bayes plug-in

A simplified version of Bayes plugin

- Naive assumption : all features $x_1 \dots x_d$ are independent
  i.e. $p(\underline{x} | \omega_j) = \prod_{i=1}^{d} p(x_i | \omega_j, \underline{\nu}_{ij}) \quad \forall \omega_j$

- Bayes plugin : $p(x_i | \omega_j ; \underline{\nu}_{ij})$ known except for $\underline{\nu}_{ij}$

$\Rightarrow$ Estimate $\underline{\nu}_{ij}$ from x_i of N_j training samples.

Special case : Naive Gaussian classifier

- Gaussian likelihood : $\underline{x} | \omega_j \sim N(\underline{\mu}_j, \underline{\underline{C}}_j)$
- "naive" : $\underline{\underline{C}}_j = \begin{bmatrix} \sigma_{1,j}^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{d,j}^2 \end{bmatrix}$

- ML parameter estimator:

  $\hat{\underline{\mu}}_j$: as in E 4.4

  σ_{ij}^2 : diagonal elements of $\hat{\underline{\underline{C}}}_j$ in E 4.4

Benefit:

- Less parameters to estimate
- Less training samples as we have less parameters to estimate

  → lower complexity

# parameters	Gaussian	naive Gaussian
$\underline{\mu}_j$	d	d
$\underline{\underline{C}}_j$	$\frac{d(d+1)}{2}$	d
Σ	$\frac{d(d+3)}{2}$	$2d$
eg $d = 100$	5150	200 per class

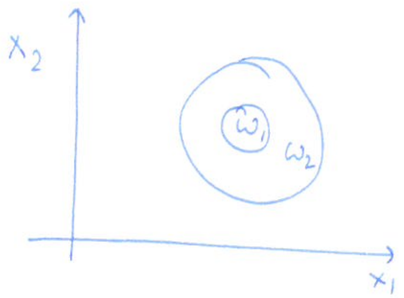
Drawbacks

- rarely independent features in real-life, naive approx

Slide 4.31

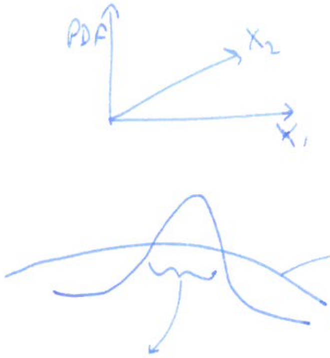
Explaination to data set 2:

true pdf



"2 donuts"

Estimated PDF



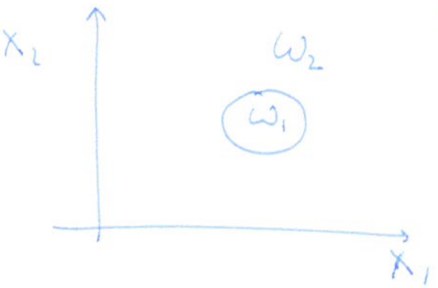
$\omega_2 \rightarrow$ higher cov as far from each other

$\omega_1 \rightarrow$ lower cov as pts located close to each other

here we predict ω_1 as
$P(\omega_1) > P(\omega_2)$

outside we have ω_2 as
$P(\omega_2) > P(\omega_1)$

Hence we get

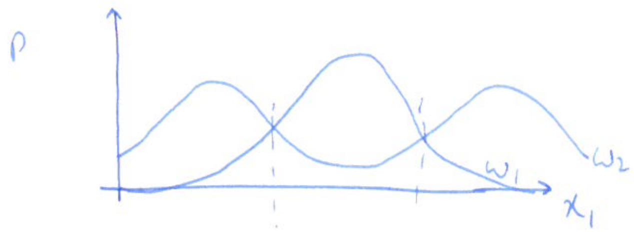


bad estimate of $P(\underline{x}/\omega_1)$, $P(\underline{x}/\omega_2)$, but we have still satisfied decision boundary $\rightarrow$ we have luck.

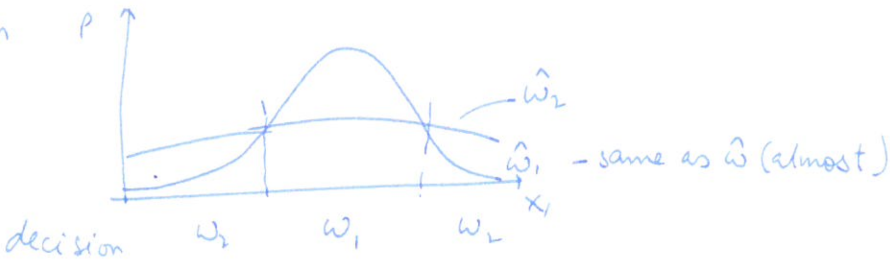
Explaination to dataset 3:

true PDF : $x_1, x_2 \rightarrow$ independent

$x_2 \rightarrow$ uniform in $[0, 1]$



estimated distribution



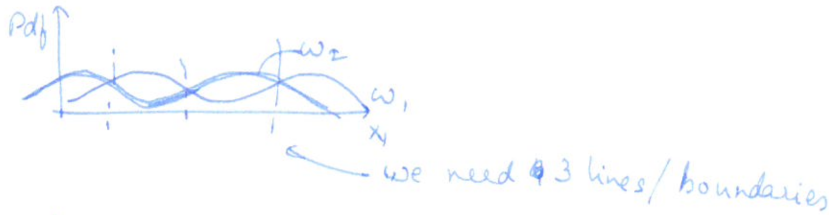
good estimate for $p(\underline{x}|\omega_1)$

bad estimate for $p(\underline{x}|\omega_2)$ } we have luck

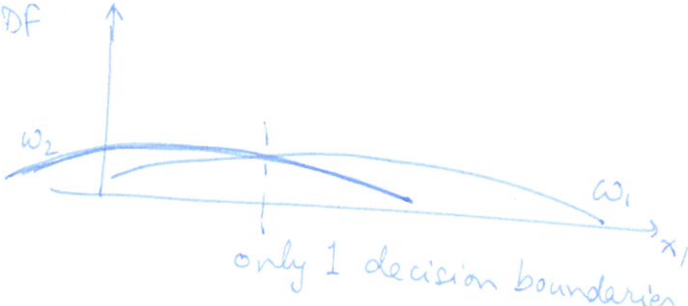
Satisfied decision boundary

Explaination to dataset 4:

true PDF



estimated PDF



bad estimate of $p(\underline{x}|\omega_1)$, $p(\underline{x}|\omega_2)$

bad decision boundary ; hence high error rate

4.3.4 Gaussian Mixture Models

Upto now: One Gaussian likelihood $N(\underline{\mu}, \underline{\underline{C}})$ per class

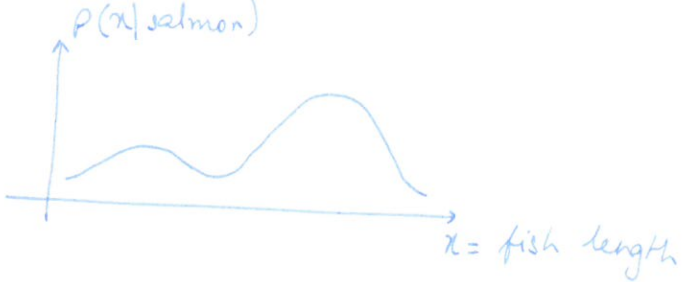
$\underline{\nu} = \{\underline{\mu}, \underline{\underline{C}}\}$ contains $\frac{d(d+3)}{2}$ parameters

Limitations → only one mode / component / maximum
→ unimodal pdf

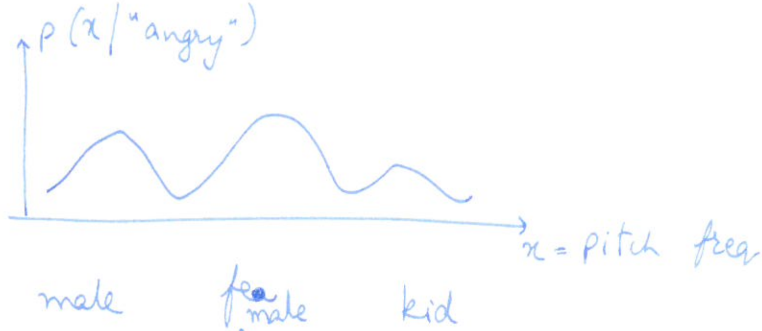
Wish → a multimodal pdf

E 4.7: Multimodal PDF

(a) fish:
multiple species of fish:



(b) Emotion recognition:
from speech



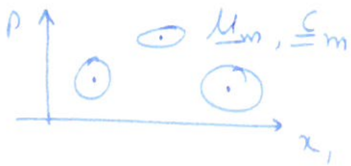
Idea: Gaussian Mixture Model (GMM), a mixture of Gaussian PDFs per class

GMM for one class:

- M Gaussian modes / sub-classes:

$P_m(\underline{x}; \underline{\nu}_m) \sim N(\underline{\mu}_m, \underline{\underline{C}}_m)$

$\underline{\nu}_m = \{\underline{\mu}_m, \underline{\underline{C}}_m\} \quad 1 \leq m \leq M$



- Random $\underbrace{\text{mode variable}}$ $Z \in \{1 \dots M\}$
  (sub class label)

  $Z = m$: $\underline{x}$ generated by mode m $\hat{=}$ subclass label m.

- law of total probability :

$$\text{PDF} \quad p(\underline{x}; \underline{\nu}) = \sum_{m=1}^{M} \underbrace{p(\underline{x} \mid z = m)}_{P_m(\underline{x}; \underline{\nu}_m)} \underbrace{P(z = m)}_{\alpha_m}$$

$$= \sum_{m=1}^{M} \alpha_m P_m(\underline{x}; \underline{\nu}_m)$$

- $\alpha_m = P(z = m) \geq 0 \quad \sum_{m=1}^{M} \alpha_m = 1$: prior of Z, mode weight, $M-1$ parameters

- $\underline{\nu} = \{\underline{\nu}_1 \dots \underline{\nu}_M ; \alpha_1 \dots \alpha_{M-1}\}$ contains $M \underbrace{\frac{d(d+3)}{2}}_{\text{for each } \underline{\nu}_i} + M - 1$ parameters

  if $\underline{\underline{C}}_m$ is non-diagonal

  If naive then $2Md + M - 1$ if $\underline{\underline{C}}_m$ is diagonal.

i.e. 2 levels of class label :

- class label ω : $\quad \omega_1 \quad \omega_2$
  to be classified with labelled samples

- mode m : $\quad 1 \dots M_1 \quad 1 \dots M_2$
  not relevant for classification, no labels
  but relevant for GMM parameter estimation

Q How to estimate $\underline{\nu}$?

A Direct ML parameter estimation

$$\max_{\underline{\nu}} p(\underline{x}; \underline{\nu})$$

- $M = 1$: closed form solution see 4.4
- $M > 1$: no closed form soln., hard to solve.

Slide 4.34

B Assumption: N iids measurements of $\underline{x}_n \in \mathbb{R}^d$

and $z_n \in \{1, \ldots M\}$ $1 \leq n \leq N$

z_n : additional mode measurements

Let $\underline{X} = \begin{bmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_N \end{bmatrix}$ $\underline{Z} = \begin{bmatrix} z_1 \\ \vdots \\ z_N \end{bmatrix}$

continuous valued — discrete valued

joint PDF/PMF model for $(\underline{X}, \underline{Z})$:

$$p(\underline{X}, \underline{Z}; \underline{\nu}) = \prod_{n=1}^{N} p(\underline{x}_n, z_n; \underline{\nu})$$

$$p(\underline{x}_n, \underline{z}_n; \underline{\nu}) = \underbrace{p(\underline{x}_n | z_n)}_{p_{z_n}(\underline{x}_n ; \nu)} \underbrace{p(z_n)}_{\alpha_{z_n}}$$

$$= \alpha_{z_n} p_{z_n}(\underline{x}_n; \underline{\nu})$$

$$\max_{\underline{\nu}} p(\underline{X}, \underline{Z}; \underline{\nu})$$

unrealistic due to missing $\underline{z}$

c) Consider $\underline{Z}$ as unknown: ("missing variable)

$$\max_{\underline{\nu}} \; \underset{\checkmark}{P} \left(\underset{\checkmark}{\underline{X}}, \; \underset{?}{\underline{z}} \, ; \underset{?}{\underline{\nu}} \right)$$

How? Expectation Maximization Algorithm (EM Algorithm)