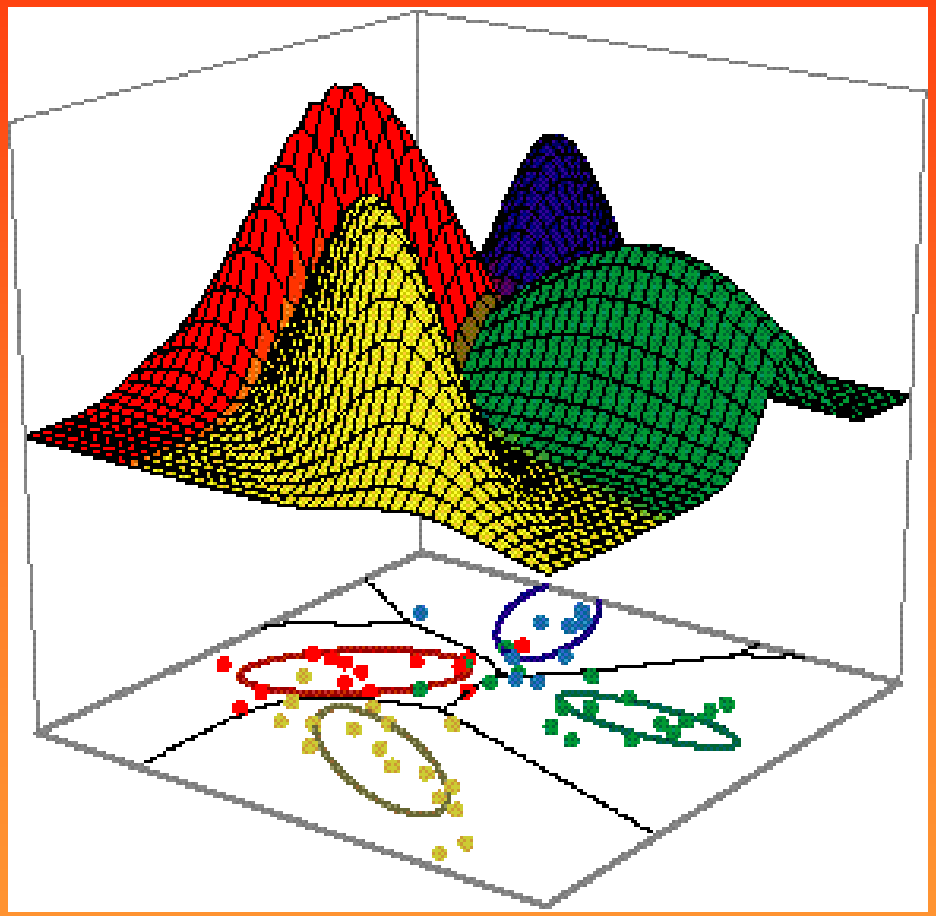


Pattern Classification

All materials in these slides were taken from
Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000
with the permission of the authors and the publisher

Chapter 3: Estymator największej wiarygodności i estymator Bayesa (part 1)

- Wstęp
- Estymator największej wiarygodności
  - Przykład dla przypadku szczególnego
  - Przypadek Gaussowski: nieznane μ i σ
  - Obciążenie estymatora
- Dodatek: Sformułowanie zadania MW

- Wstęp
  - Data availability in a Bayesian framework
    - Możemy opracować optymalny klasyfikator Bayesa, znając:
      - $P(\omega_i)$ (prawdopodobieństwa a'priori)
      - $P(x \mid \omega_i)$ (warunkowe rozkłady w klasach)

      Niestety, jedynie w wyjątkowych sytuacjach dysponujemy pełną informacją probabilistyczną
  - Opracowywanie klasyfikatora w oparciu o ciąg uczący
    - Łatwo oszacować prawdopodobieństwa a'priori
    - Ciąg uczący jest zwykle zbyt mały do estymacji rozkładów warunkowych (duży wymiar przestrzeni cech)

- Informacja aprioryczna o problemie
- Normalny rozkład $P(x \mid \omega_i)$

$$P(x \mid \omega_i) \sim N(\mu_i, \Sigma_i)$$

  - Rozkład charakteryzowany przez dwa parametry
- Techniki estymacji
  - Estymator Maksymalnej Wiarygodności (MW)
  - Estymator Bayesa
  - Wyniki końcowe są podobne, ale podejścia są różne

- Zestaw parametrów w zadaniu MW jest ustalony, ale wartości parametrów należy wyznaczyć

- Najlepsze oszacowania wartości parametrów uzyskujemy maksymalizując prawdopodobieństwo otrzymania danych, jakie zostały zaobserwowane

- W metodach Bayesowskich parametry rozpatrywane są jako zmienne losowe o znanych rozkładach

- W obu podejściach używamy $P(\omega_i \mid x)$ do konstrukcji reguł decyzyjnych klasyfikatora

Estymator Maksymalnej Wiarygodności

- Ma dobre własności zbieżności dla rosnącej długości ciągu uczącego
- Łatwiejszy w wyznaczeniu od estymatorów innych metod

- Ogólna idea działania estymatora

- Załóżmy, że jest c klas obiektów

  $P(x \mid \omega_j) \sim N(\mu_j, \Sigma_j)$

  $P(x \mid \omega_j) \equiv P(x \mid \omega_j, \theta_j)$ gdzie:

$$\theta = (\mu_j, \Sigma_j) = (\mu_j^1, \mu_j^2, ..., \sigma_j^{11}, \sigma_j^{22}, \mathrm{cov}(x_j^m, x_j^n)...)$$

- Wykorzystanie ciągu uczącego do estymacji

  $\theta = (\theta_1, \theta_2, \ldots, \theta_c)$, gdzie θ_i (i = 1, 2, …, c) odpowiadają klasom

- Niech D zawiera n próbek, $x_1, x_2, \ldots, x_n$

$$P(D \mid \theta) = \prod_{k=1}^{n} P(x_k \mid \theta) = F(\theta)$$

$P(D \mid \theta)$ – funkcja wiarygodności

- Estymator MW parametru θ maksymalizuje $P(D \mid \theta)$

  “Jest taką wartością parametru θ, która najlepiej odpowiada zaobserwowanym pomiarom, zebranym w ciągu uczącym”

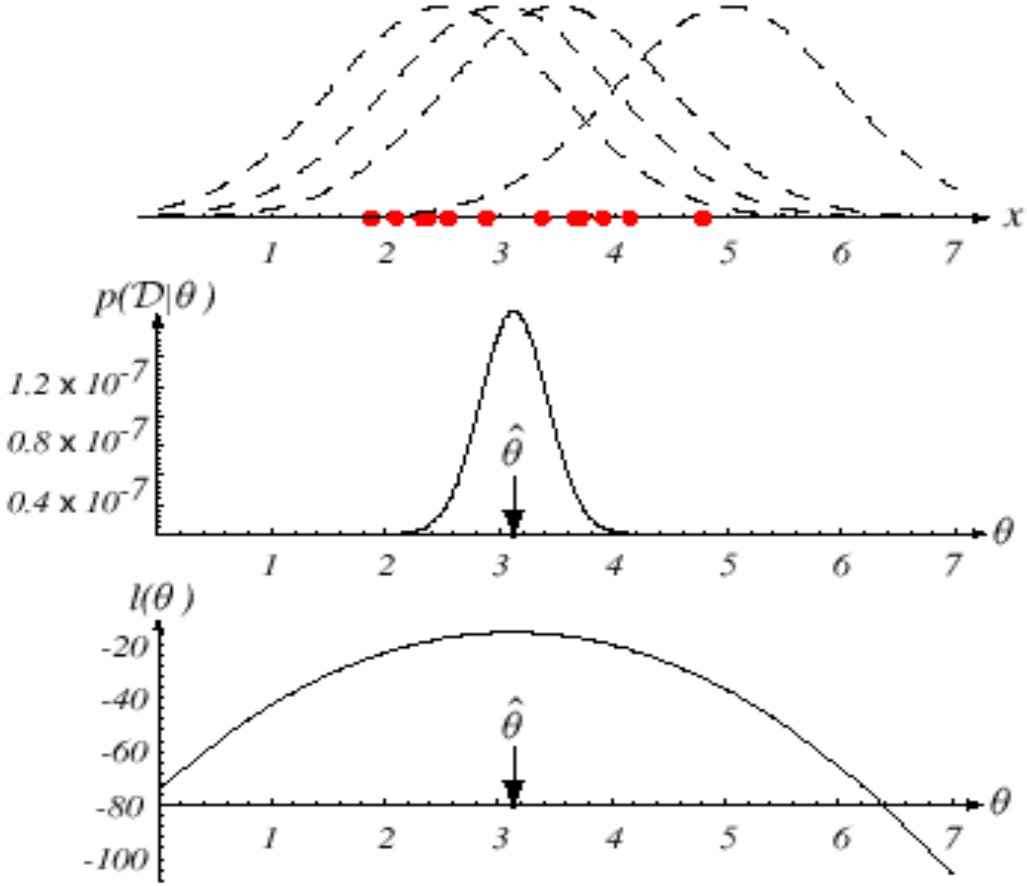


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.

- Estymator optymalny
  - Niech $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^T$ oraz niech ∇_θ oznacza gradient

$$\nabla_\theta = \left[\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, \ldots, \frac{\partial}{\partial \theta_p} \right]^T$$

  - Oznaczmy $l(\theta)$ jako:

$$l(\theta) = \ln P(D \mid \theta)$$

  - Sformułowanie problemu:
    wyznaczyć wartość θ, która maksymalizuje:

$$\hat{\theta} = \arg\max_{\theta} l(\theta)$$

Warunek konieczny optimum:

$$(\nabla_\theta l = \sum_{k=1}^{n} \nabla_\theta \ln P(x_k \mid \theta))$$

$$\nabla_\theta l = 0$$

- Szczególny przypadek: nieznane μ
  - $P(x_i \mid \mu) \sim N(\mu, \Sigma)$
    (Wartości cech obiektów są zmiennymi losowymi o wielowymiarowym rozkładzie normalnym)

$$\ln P(x_k \mid \mu) = -\frac{1}{2}\ln\left[(2\pi)^d |\Sigma|\right] - \frac{1}{2}(x_k - \mu)^T \Sigma^{-1}(x_k - \mu)$$

i $$\nabla_{\theta=\mu} \ln P(x_k \mid \mu) = \Sigma^{-1}(x_k - \mu) = 0$$

$\theta = \mu$ a zatem:

- Estymator MW parametru μ musi spełniać warunek:

$$\sum_{k=1}^{n} \Sigma^{-1}(x_k - \hat{\mu}) = 0$$

- Po wymnożeniu przez Σ oraz po prostych przekształceniach otrzymujemy:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

Jest to średnia arytmetyczna próbek z ciągu uczącego

Wniosek:

Jeżeli $P(x_k \mid \omega_j)$ (j = 1, 2, …, c) jest rozkładem Gaussa w *d*-wymiarowej przestrzeni cech to można estymować

$\theta = (\theta_1, \theta_2, \ldots, \theta_c)^T$ i dokonać optymalnej klasyfikacji (w sensie bayesowskim).

- Estymator MW:
  - Rozkład normalny: *nieznane* μ *oraz* σ

$\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$

$$l = \ln P(x_k \mid \theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_\theta l = \begin{pmatrix} \frac{\partial}{\partial\theta_1}(\ln P(x_k \mid \theta)) \\ \frac{\partial}{\partial\theta_2}(\ln P(x_k \mid \theta)) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} \frac{1}{\theta_2}(x_k - \theta_1) = 0 \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} = 0 \end{cases}$$

Sumowanie:

$$\begin{cases} \sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \theta_1) = 0 & (1) \\ -\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 & (2) \end{cases}$$

Rozwiązanie układu daje wynik:

$$\hat{\theta}_1 = \mu = \sum_{k=1}^{n} \frac{x_k}{n} \quad ; \quad \hat{\theta}_2 = \sigma^2 = \frac{\sum_{k=1}^{n} (x_k - \mu)^2}{n}$$

- Obciążenie estymatora
  - Estymator MW parametru σ^2 jest obciążony

$$E\left[\frac{1}{n}\Sigma(x_i-\bar{x})^2\right]=\frac{n-1}{n}.\sigma^2\neq\sigma^2$$

  - Nieobciążony estymator macierzy Σ ma postać:

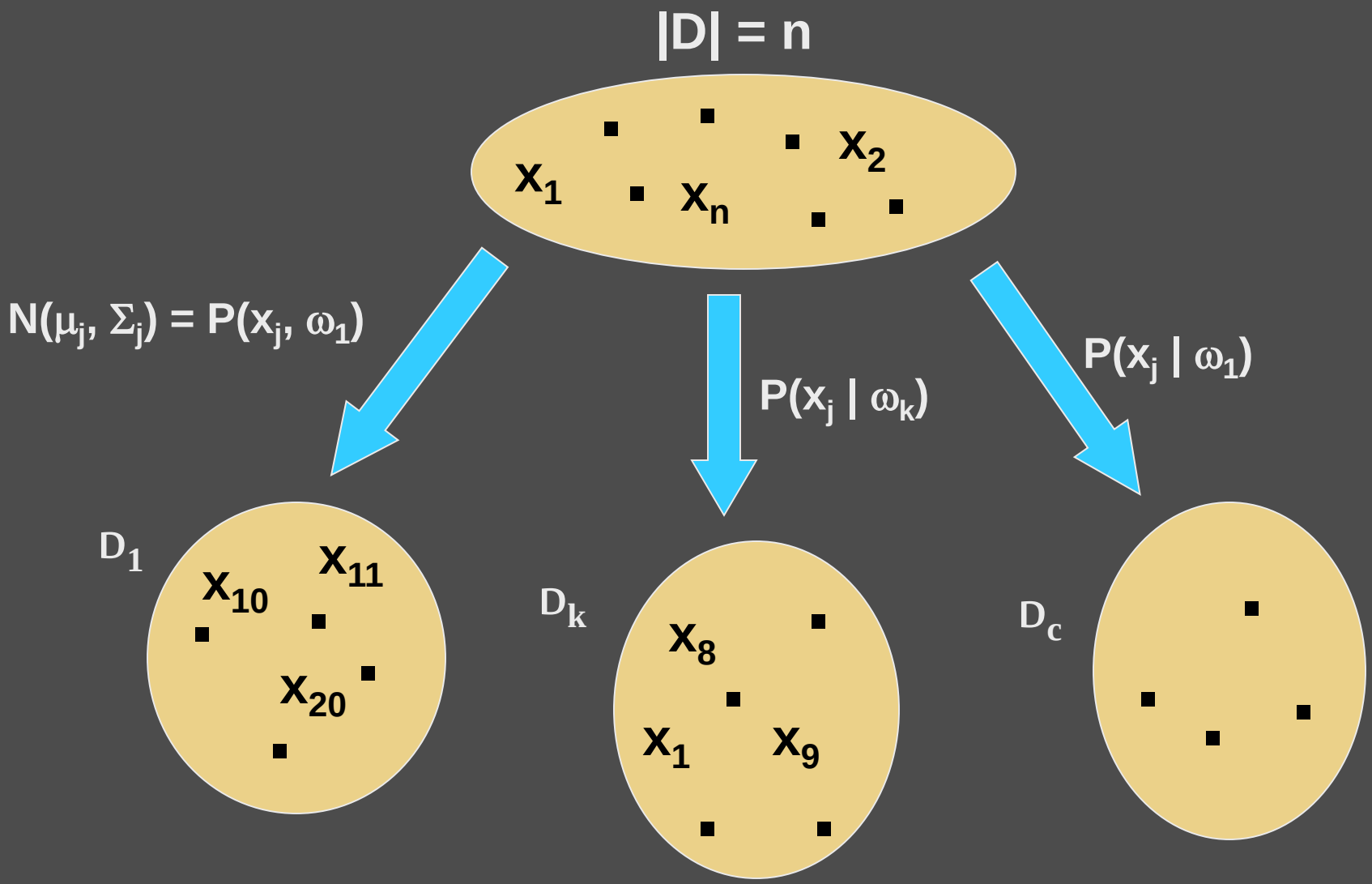
$$\underbrace{C=\frac{1}{n-1}\sum_{k=1}^{k=n}(x_k-\mu)(x_k-\hat{\mu})^T}_{\text{macierz kowariancji z próby}}$$

- Dodatek: Sformułowanie zadania estymacji MW

  - Niech D = $\{x_1, x_2, \ldots, x_n\}$

    $$P(x_1,\ldots, x_n \mid \theta) = \prod_{k=1}^{n} P(x_k \mid \theta); |D| = n$$

    Należy wyznaczyć wartość $\hat{\theta}$, przy której ciąg uczący jest najbardziej reprezentatywny


|D| = n
x1
xn
x2
N(μj, Σj) = P(xj, ω1)
P(xj | ωk)
P(xj | ω1)
D1
x10
x11
x20
Dk
x8
x1
x9
Dc

$\theta = (\theta_1, \theta_2, \ldots, \theta_c)$

Problem: wyznaczyć $\hat{\theta}$ takie, że:

$$\max_{\theta} P(D \mid \theta) = \max P(x_1, \ldots, x_n \mid \theta)$$

$$= \max \prod_{k=1}^{n} P(x_k \mid \theta)$$

Chapter 3: Estymator największej wiarygodności i estymator Bayesa (part 1)

- Metoda Bayesa (MB)
  - Parametryczny estymator Bayesa: rozkład normalny
  - Parametryczny estymator Bayesa: przypadek ogólny
- Problemy z wielowymiarowością
- Złożoność obliczeniowa
- Analiza komponentów i analiza dyskryminacyjna
- Ukryte łańcuchy Markowa

- Estymator Bayesa (uczenie bayesowskie w rozpoznawaniu obrazów)
  - In MLE θ was supposed fix
  - In BE θ is a random variable
  - Wyznaczanie prawdopodobieństw a'posteriori $P(\omega_i \mid x)$
  - Wyznaczyć $P(\omega_i \mid x, D)$
    Wzór Bayesa dla próbki D:

$$P(\omega_i \mid x, \mathrm{D}) = \frac{P(x \mid \omega_i, \mathrm{D}) P(\omega_i \mid \mathrm{D})}{\sum_{j=1}^{c} P(x \mid \omega_j, \mathrm{D}) P(\omega_j \mid \mathrm{D})}$$

$$P(x,\mathsf{D}\,|\,\omega_i) = P(x\,|\,\mathsf{D},\omega_i)P(\mathsf{D}\,|\,\omega_i)$$

$$P(x\,|\,\mathsf{D}) = \sum_j P(x,\omega_j\,|\,\mathsf{D})$$

$$P(\omega_i) = P(\omega_i\,|\,\mathsf{D})$$

Ostatecznie :

$$P(\omega_i\,|\,x,\mathsf{D}) = \frac{P(x\,|\,\omega_i,\mathsf{D}_i)P(\omega_i)}{\sum_{j=1}^{c} P(x\,|\,\omega_j,\mathsf{D})P(\omega_j)}$$

- Parametryczny estymator Bayesa: rozkład normalny

Cel: Estymacja θ z wykorzystaniem rozkładu a'posteriori $P(\theta \mid D)$

- Przypadek jednowymiarowy: $P(\mu \mid D)$

  μ jest jedynym nieznanym parametrem

$$P(x \mid \mu) \sim N(\mu, \sigma^2)$$

$$P(\mu) \sim N(\mu_0, \sigma_0^2)$$

$$P(\mu \mid \mathrm{D}) = \frac{P(\mathrm{D} \mid \mu)P(\mu)}{\int P(\mathrm{D} \mid \mu)P(\mu)d\mu} \qquad (1)$$

$$= \alpha \prod_{k=1}^{k} P(x_k \mid \mu)P(\mu)$$

- Estymacja rozkładu

$$P(\mu \mid \mathrm{D}) \sim N(\mu_n, \sigma_n^2) \qquad (2)$$

Porównanie (1) i (2):

$$\mu_n = \left(\frac{n\sigma_0^2}{n_0\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0$$

$$\text{and} \quad \sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}$$

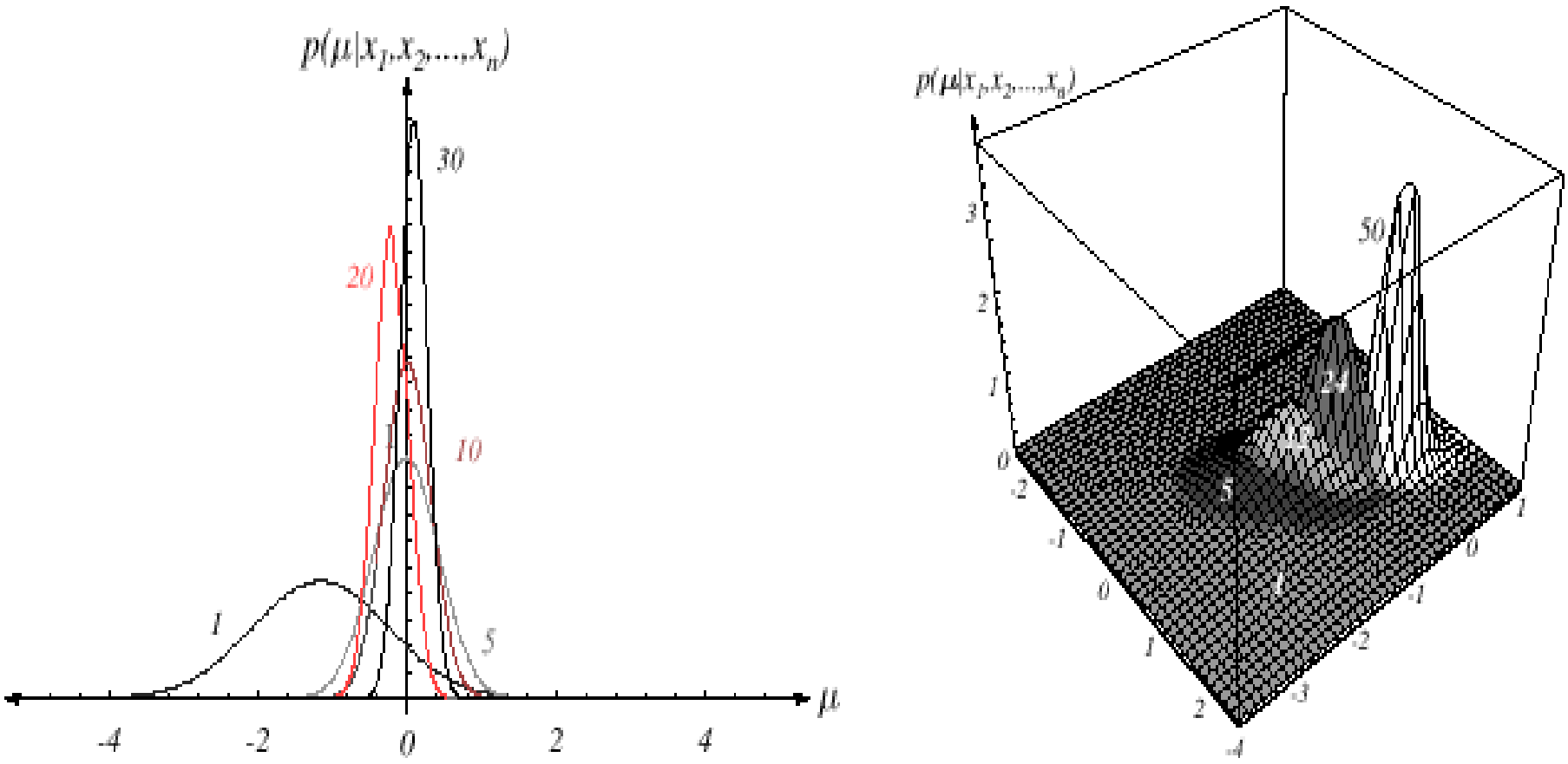


FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Przypadek jednowymiarowy P(x | D)
  - P(μ | D) jest już wyznaczone
  - P(x | D) pozostało do wyznaczenia!

$$P(x|\mathrm{D}) = \int P(x|\mu)P(\mu|\mathrm{D})d\mu$$ jest rozkładem normalnym

A więc: $$P(x|\mathrm{D}) \sim N(\mu_n, \sigma^2 + \sigma_n^2)$$

(szukany rozkład w klasach P(x | D_j, ω_j))

Mając P(x | D_j, ω_j) wraz z P(ω_j) i wzorem Bayesa, otrzymujemy regułę decyzyjną dla klasyfikatora Bayesa:

$$\max_{\omega_j}\left[P(\omega_j \mid x, \mathrm{D}\right] \equiv \max_{\omega_j}\left[P(x \mid \omega_j, \mathrm{D}_j)P(\omega_j)\right]$$

- Parametryczny estymator Bayesa: przypadek ogólny
  - Wyznaczenie P(x | D) może być przeprowadzone w każdej sytuacji, w której nieznany rozkład daje się parametryzować. Szczegółowe założenia są następujące:
    - Postać P(x | θ) jest znana z dokładnością do wartości θ, którą należy wyznaczyć
    - Wiedza o θ jest zawarta w znanym rozkładzie a'priori P(θ)
    - Pozostała część wiedzy o θ jest zawarta w zbiorze D n zmiennych losowych $x_1, x_2, \ldots, x_n$

Podstawowe zadanie:

"Wyznacz rozkład a'posteriori P(θ | D)" a następnie "wyznacz P(x | D)"

Ze wzoru Bayesa mamy:

$$P(\theta \mid \mathrm{D}) = \frac{P(\mathrm{D} \mid \theta) P(\theta)}{\int P(\mathrm{D} \mid \theta) P(\theta) d\theta},$$

Z warunku niezależności otrzymujemy:

$$P(\mathrm{D} \mid \theta) = \prod_{k=1}^{k=n} P(x_k \mid \theta)$$

- Problemy z wielowymiarowością
  - Zadania z 50 lub 100 binarnymi cechami
    - Dokładność klasyfikacji zależy od wymiaru i ilości danych uczących
    - Przypadek wielowymiarowy z rozkładem normalnym i dwiema klasami o tej samej kowariancji

$$P(\text{blad}) = \frac{1}{\sqrt{2\pi}} \int_{r/2}^{\infty} e^{\frac{-u^2}{2}} du$$

$$\text{gdzie}: \quad r^2 = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$$

$$\lim_{r \to \infty} P(\text{blad}) = 0$$

- Jeżeli cechy są niezależne, to:

$$\Sigma = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$$

$$r^2 = \sum_{i=1}^{i=d} \left(\frac{\mu_{i1} - \mu_{i2}}{\sigma_i} \right)^2$$

- Najbardziej użyteczne cechy to takie, których różnica między średnią jest duża w porównaniu do odchylenia standardowego

- Często obserwuje się w praktyce, że – począwszy od pewnego momentu – uwzględnianie kolejnych cech prowadzi do pogorszenia jakości klasyfikacji!.

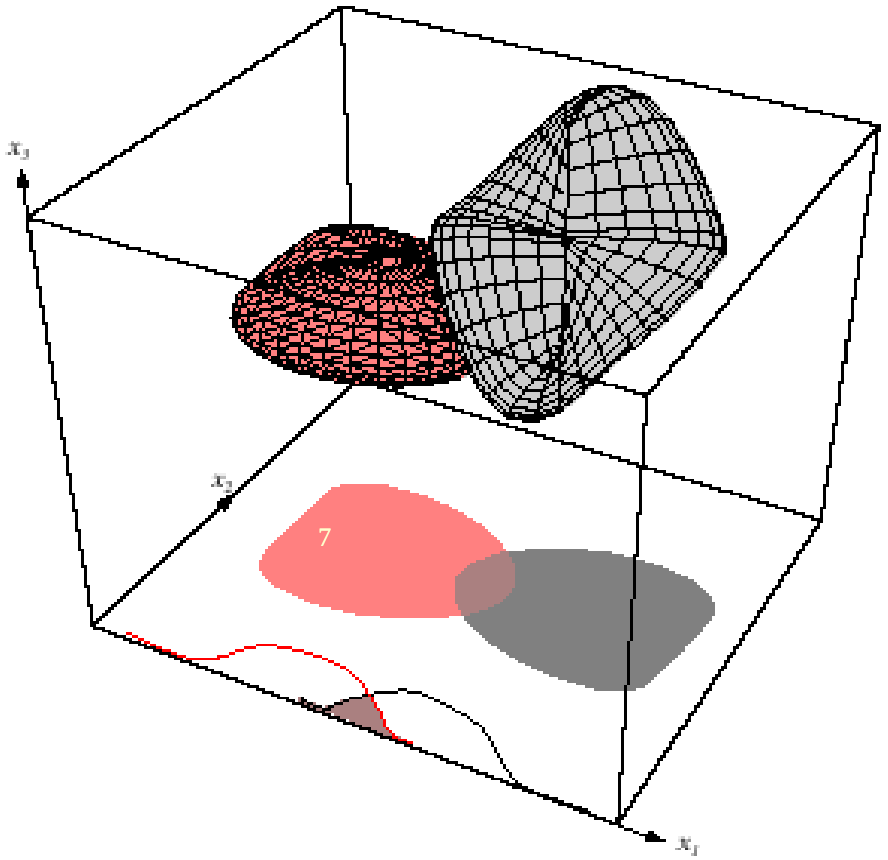


FIGURE 3.3. Two three-dimensional distributions have nonoverlapping densities, and thus in three dimensions the Bayes error vanishes. When projected to a subspace—here, the two-dimensional $x_1 - x_2$ subspace or a one-dimensional x_1 subspace—there can be greater overlap of the projected distributions, and hence greater Bayes error. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Złożoność obliczeniowa

  - Notacja “wielkie o”

    f(x) = O(h(x))

    Jeżeli: $\exists (c_0, x_0) \in \Re^2;\ |f(x)| \le c_0 |h(x)|$

    (górna granica f(x) rośnie nie szybciej niż h(x) wystarczająco x o wystarczająco dużym wymiarze!)

    $f(x) = 2+3x+4x^2$

    $g(x) = x^2$

    $f(x) = O(x^2)$

$f(x) = O(x^2)$; $f(x) = O(x^3)$; $f(x) = O(x^4)$

- Notacja “wielkie theta”

  $f(x) = \theta(h(x))$

  Jeżeli:

$$\exists (x_0, c_1, c_2) \in \Re^3; \forall x > x_0$$
$$0 \le c_1 g(x) \le f(x) \le c_2 g(x)$$

$f(x) = \theta(x^2)$ ale $f(x) \neq \theta(x^3)$

- Złożoność estymatora MW
  - Rozkłady normalne a'priori, d-wymiarowy klasyfikator, na każdą z c klas przypada n próbek ciągu uczącego
  - Dla każdej klasy wyznaczamy funkcję dyskryminującą:

$$g(x) = -\frac{1}{2}\underbrace{(x-\hat{\mu})^T}_{O(d.n)}\overbrace{\hat{\Sigma}^{-1}}^{O(n.d^2)}(x-\hat{\mu}) - \overbrace{\frac{d}{2}\ln 2\pi}^{O(1)} - \underbrace{\frac{1}{2}\ln\left|\hat{\Sigma}\right|}_{O(d^2.n)} + \underbrace{\ln P(\omega)}_{O(n)}$$

Złożoność ogółem= $O(d^2.n)$

Złożoność przy c klasach = $O(cd^2.n) \cong O(d^2.n)$

- Koszt obliczeniowy jest znaczący kiedy d i n są duże

- Analiza komponentów i analiza dyskryminacyjna
  - Łączenie cech w celu redukcji wymiaru przestrzeni cech
  - Najprościej: kombinacje liniowe cech
  - Projekcja danych z przestrzeni wysokowymiarowych do przestrzeni o mniejszej liczbie wymiarów
  - Dwa klasyczne podejścia do znajdywania „optymalnej” transformacji liniowej
    - PCA (Principal Component Analysis) “Projekcja dająca najlepszą reprezentację danych w sensie średniokwadratowym”
    - MDA (Multiple Discriminant Analysis) “Projekcja najlepiej separująca dane w sensie średniokwadratowym”

- Ukryte łańcuchy Markowa:
  - Łańcuchy Markowa
  - Cel: wyznaczyć sekwencję decyzji
    - Procesy dynamiczne, na stany w czasie t wpływają stany w czasie t-1
    - Zastosowania: rozpoznawanie i tagowanie mowy, rozpoznawanie gestów, sekwencjonowanie DNA,
    - Proces bez pamięci:

      $\omega^T = \{\omega(1), \omega(2), \omega(3), \ldots, \omega(T)\}$ – sekwencja stanów, np.

      $$\omega^6 = \{\omega 1, \omega 4, \omega 2, \omega 2, \omega 1, \omega 4\}$$
    - Proces może „odwiedzić” ten sam stan w różnych krokach, przy czym pewne stany w ogóle nie muszą być odwiedzane

- Model Markowa pierwszego rzędu
  - Prawdopodobieństwa przejść do kolejnych stanów:

$$P(\omega_j(t+1) \mid \omega_i(t)) = a_{ij}$$

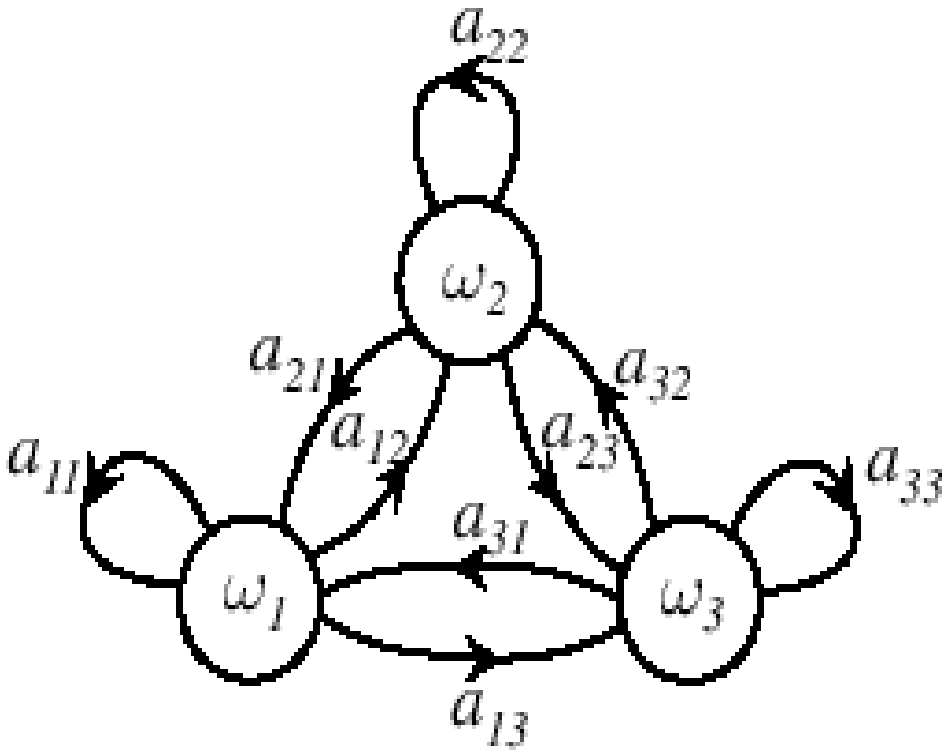


FIGURE 3.8. The discrete states, ω_i, in a basic Markov model are represented by nodes, and the transition probabilities, a_{ij}, are represented by links. In a first-order discrete-time Markov model, at any step t the full system is in a particular state $\omega(t)$. The state at step $t+1$ is a random function that depends solely on the state at step t and the transition probabilities. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

$\theta = (a_{ij}, \omega^T)$

$P(\omega^T \mid \theta) = a_{14} \cdot a_{42} \cdot a_{22} \cdot a_{21} \cdot a_{14} \cdot P(\omega(1) = \omega_i)$

Przykład: rozpoznawanie mowy

“wypowiadanie kolejnych słów”

“obraz” słowa to reprezentacja przy użyciu fonemów: /p/ /a/ /tt/ /er/ /n/ // (// = stan ciszy)

Przejścia: /p/ do /a/, /a/ do /tt/, /tt/ do /er/, /er/ do /n/ i /n/ do stanu ciszy

Chapter 3:
Estymator największej wiarygodności i estymator Bayesa (part 3)

- Ukryte modele Markowa

- Ukryty model Markowa
  - Związek stanów procesu ze stanami ukrytymi
    $\sum b_{jk} = 1$ *dla wszystkich j, dla których* $b_{jk} = P(V_k(t) \mid \omega_j(t))$.
  - 3 zadania związane z tym modelem
    - The evaluation problem
    - Zadanie dekodowania
    - Zadanie uczenia

- The evaluation problem

Prawdopodobieństwo, że w trakcie działania procesu zaobserwowana zostanie sekwencja stanów V^T:

$$P(V^T) = \sum_{r=1}^{r_{\max}} P(V^T \mid \omega_r^T) P(\omega_r^T)$$

gdzie *r* jest numerem sekwencji T stanów ukrytych

$$\omega_r^T = \{\omega(1), \omega(2), ..., \omega(T)\}$$

$$(1) \qquad P(V^T \mid \omega_r^T) = \prod_{t=1}^{t=T} P(v(t) \mid \omega(t))$$

$$(2) \qquad \mathrm{P}(\omega_{\mathrm{r}}^{\mathrm{T}}) = \prod_{t=1}^{t=T} P(\omega(t) \mid \omega(t-1))$$

Z zależności (1) i (2) wynika:

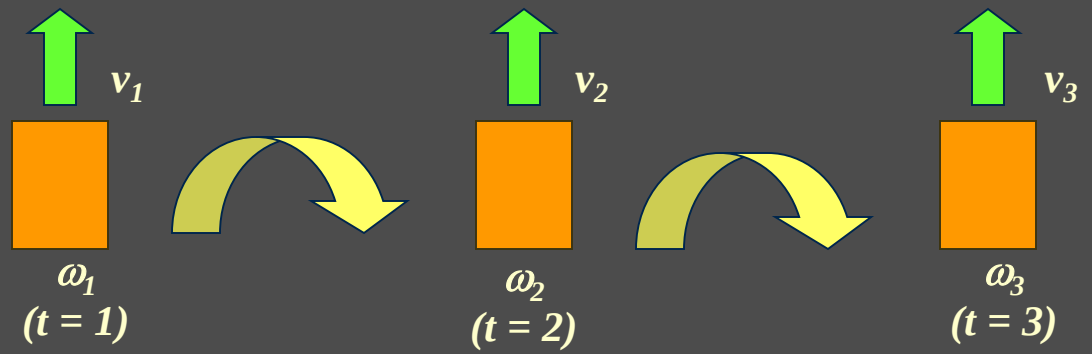
$$P(V^T) = \sum_{r=1}^{r_{\max}} \prod_{t=1}^{t=T} P(v(t) \mid \omega(t))\, P(\omega(t) \mid \omega(t-1))$$

Interpretacja: Prawdopodobieństwo zaobserwowania konkretnej sekwencji T stanów procesu V^T jest sumą po wszystkich r_{max} możliwych sekwencji stanów ukrytych, iloczynów prawdopodobieństw warunkowych przejścia do poszczególnych stanów i prawdopodobieństw zaobserwowania mierzalnych stanów w sekwencji.

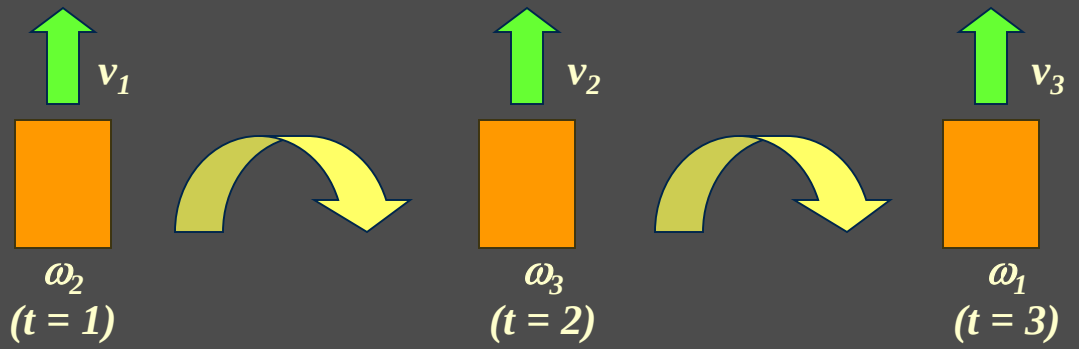
Przykład: Niech ω_1, ω_2, ω_3 oznaczają ukryte stany; v_1, v_2, v_3 stany mierzalne a $V^3 = \{v_1, v_2, v_3\}$ jest sekwencją stanów widocznych.

$P(\{v_1, v_2, v_3\}) = P(\omega_1)\, P(v_1 \mid \omega_1)\, P(\omega_2 \mid \omega_1)\, P(v_2 \mid \omega_2)\, P(\omega_3 \mid \omega_2)\, P(v_3 \mid \omega_3)$
+…+ (liczba wszystkich składników sumy to (3^3= 27)!)

Pierwszy przypadek:



Drugi przypadek:



$P(\{v_1, v_2, v_3\}) = P(\omega_2).P(v_1 \mid \omega_2).P(\omega_3 \mid \omega_2).P(v_2 \mid \omega_3).P(\omega_1 \mid \omega_3).P(v_3 \mid \omega_1) + \ldots +$

Zatem:

$$P(\{v_1, v_2, v_3\}) = \sum_{\text{możliwe sekwencje stanów ukrytych}} \prod_{t=1}^{t=3} P(v(t) \mid \omega(t)) P(\omega(t) \mid \omega(t-1))$$

- Zadanie dekodowania (optymalna sekwencja stanów)

  Dla sekwencji stanów mierzalnych V^T, the decoding problem polega na znalezieniu najbardziej prawdopodobnej sekwencji stanów ukrytych.

$\hat{\omega}(1), \hat{\omega}(2), ..., \hat{\omega}(T)$, która spełnia

$$\hat{\omega}(1), \hat{\omega}(2), ..., \hat{\omega}(T) = \arg\max_{\omega(1),\omega(2),...,\omega(T)} P\left[\omega(1), \omega(2), ..., \omega(T), v(1), v(2), ..., V(T) \mid \lambda\right]$$

gdzie: $\lambda = [\pi, A, B]$

$\pi = P(\omega(1) = \omega)$ *(prawdopodobieństwo stanu początkowego)*

$A = a_{ij} = P(\omega(t+1) = j \mid \omega(t) = i)$

$B = b_{jk} = P(v(t) = k \mid \omega(t) = j)$

Zauważmy, że brak jest tu sumowania, ponieważ poszukiwana jest jedna, najlepsza sekwencja

W poprzednim przykładzie, te obliczenia odpowiadają wyborowi najlepszej ścieżki spomiędzy:

$$\{\omega_1(t=1),\omega_2(t=2),\omega_3(t=3)\},\ \{\omega_2(t=1),\omega_3(t=2),\omega_1(t=3)\}$$
$$\{\omega_3(t=1),\omega_1(t=2),\omega_2(t=3)\},\ \{\omega_3(t=1),\omega_2(t=2),\omega_1(t=3)\}$$
$$\{\omega_2(t=1),\omega_1(t=2),\omega_3(t=3)\}$$

- Zadanie uczenia (estymacja parametrów)

Zadanie polega na zaproponowaniu metody estymacji parametrów $\lambda = [\pi, A, B]$ optymalizującej pewne kryterium. Należy znaleźć najlepszy model

$$\hat{\lambda} = [\hat{\pi}, \hat{A}, \hat{B}]$$

maksymalizujący prawdopodobieństwo zaobserwowania otrzymanej sekwencji :

$$\max_{\lambda} P(V^T \mid \lambda)$$

Do znalezienia lokalnego optimum można użyć metod iteracyjnych, np. metody Bauma-Welcha lub metody gradientowej