CT 720 Tópicos em Aprendizagem de Máquina e Classificação de Padrões

3-Estimação de Parâmetros: Máxima Verosimilhança e Bayes

Conteúdo

1. Introdução
2. Estimador de máxima verosimilhança
3. Estimador de Bayes
4. Teoria geral estimadores Bayesianos
5. Problemas de dimensionalidade
6. Modelos de Markov
7. Resumo

1-Introdução

- Teoria Bayesiana de decisão
  - assume $P(\omega_i)$ e $p(\mathbf{x}/\omega_i)$ completamente conhecidos
  - na prática estes valores não são conhecidos
  - projeto necessita de dados de treinamento
  - problema de estimar função $\rightarrow$ estimar parâmetros

- Este capítulo
  - apresenta métodos principais de estimação
  - problema da dimensão e complexidade
  - classificação estática e dinâmica

- Estimador de máxima verosimilhança
  - parâmetros de $p(\mathbf{x}/\omega_i)$ são valores fixos, mas desconhecidos
  - melhor estimativa: maximiza a probabilidade de obter as observações

- Estimador de Bayes
  - parâmetros são variáveis aleatórias com distribuições *a priori* dadas
  - observações convertem estas distribuições em *a posteriori*

- Aprendizagem
  - amostras **x** obtidas selecionando estado ω_i com probabilidade $P(\omega_i)$
  - amostras independentemente selecionadas de acordo com $p(\mathbf{x}/\omega_i)$
  - supervisionada: classe (estado) ω_i de cada amostra é conhecida
  - não supervisionada

2-Estimador de máxima verosimilhança

- Características
  - boa convergência quando número de amostras de treinamento aumenta
  - mais simples que métodos alternativos (Bayes, EM, etc.)

- Princípio geral

  $\mathcal{D}_1$,...., $\mathcal{D}_c$: c conjuntos de dados

  $\mathcal{D}_j$: conjunto de amostras independentemente de $p(\mathbf{x}/\omega_j)$ (*i.i.d.*)

  $p(\mathbf{x}/\omega_j)$: forma paramétrica é, por hipótese, conhecida

  $\boldsymbol{\theta}_j$: vetor de parâmetros que caracteriza $p(\mathbf{x}/\omega_j)$ de forma única

  $p(\mathbf{x}/\omega_j) = p(\mathbf{x}/\omega_j,\boldsymbol{\theta}_j)$

  problema: $\boldsymbol{\theta}_j$ *?*

- Hipótese

  $\mathcal{D}_i$: não tem informação sobre $\boldsymbol{\theta}_j$ se $i \neq j$

  parâmetros são funcionalmente independentes

  permite tratar cada classe separadamente

- Problema de estimação

  Estimar o vetor de parâmetros $\boldsymbol{\theta}_j$ a partir das amostras em $\mathcal{D}$

  amostras estas geradas independentemente a partir de $p(\mathbf{x}/\boldsymbol{\theta})$

  $\mathcal{D} = \{\mathbf{x}_1, ...,\mathbf{x}_n\}$

  $$p(\mathcal{D}\,|\,\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k\,|\,\boldsymbol{\theta})$$ *likelihood* de $\boldsymbol{\theta}$ com relação a $\mathcal{D}$ (1)

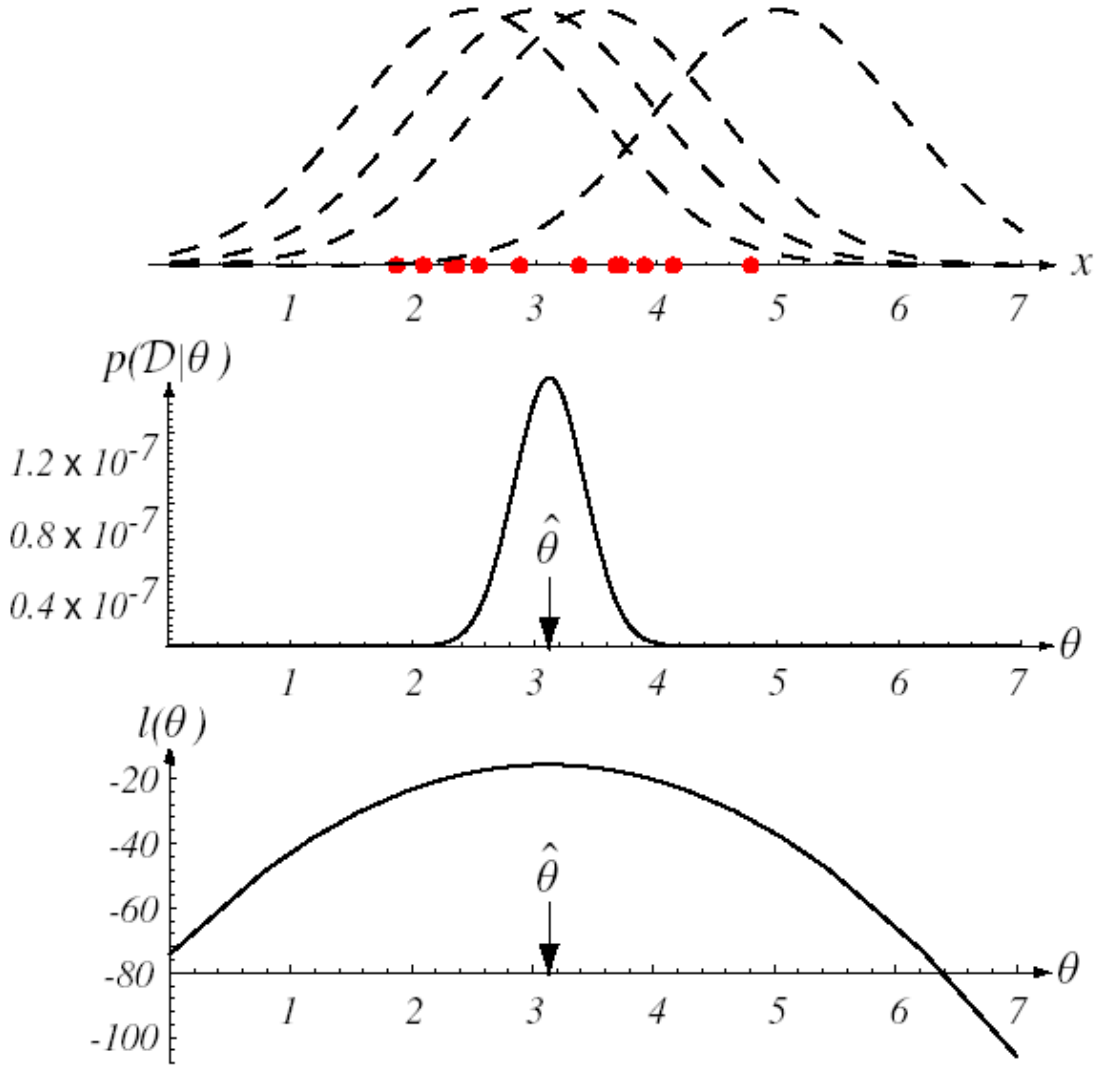
- Estimador de máxima verosimilhança (MV) $\hat{\boldsymbol{\theta}}$
  - maximiza $p(\mathcal{D}/\boldsymbol{\theta})$
  - valor de $\boldsymbol{\theta}$ mais aderente aos dados de treinamento
  - em geral usa-se *log-likelihood*

$$l(\boldsymbol{\theta}) = \ln p(\mathcal{D} \mid \boldsymbol{\theta})$$

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

$$l(\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

$$\nabla_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta}) = 0 \qquad \text{condição necessária} \quad (7)$$



- Caso Gaussiano: $\boldsymbol{\mu}$ desconhecido

$$\ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) = -\frac{1}{2}\ln[(2\pi)^d \mid \boldsymbol{\Sigma} \mid -\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

$$\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) = 0 \qquad \text{(multiplicando por } \Sigma^{-1} \text{ e rearranjando)}$$

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

- Caso Gaussiano: μ e $\sum$ desconhecidos

$$\theta_1 = \mu \text{ e } \theta_2 = \sigma^2$$

$$\ln p(x_k \mid \boldsymbol{\theta}) = -\frac{1}{2}\ln[2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

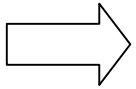
$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0$$

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

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

$$\sigma_2^2 = \frac{1}{n}\sum_{k=1}^{n} (x_k - \hat{\mu})^2$$



$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n} (\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$$

- Tendenciosidade (bias)

$\hat{\boldsymbol{\theta}}$ é um estimador não tendencioso de $\boldsymbol{\theta}$ se se somente se $E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$

$$E\left[\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$ tendencioso

$$\mathbf{C} = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$$ não tendencioso

3-Estimador de Bayes

- Densidades condicionais de classe
  - $P(\omega_i|\,\mathbf{x})$ é essencial em classificação Bayesiana
  - como obter $P(\omega_i|\,\mathbf{x})$ se $P(\mathbf{x}|\omega_i)$ e $P(\omega_i)$ dão desconhecidos ?
  - usar conhecimento e.g. forma funcional e faixas dos parâmetros
  - amostras para treinamento: conjunto $\mathcal{D}$
  - $P(\omega_i|\,\mathbf{x}, \mathcal{D})$?

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

- Hipóteses
  - probabilidades *a priori* conhecidas/calculadas: $P(\omega_i|\mathcal{D}) = P(\omega_i)$
  - $\mathcal{D}_1$,...., $\mathcal{D}_c$ conjuntos de dados de treinamento

    $\mathcal{D}_i$: não influencia $p(\mathbf{x}|\omega_j, \mathcal{D})$ se $i \neq j$
  - classes são tratadas separadamente: *c* problemas independentes

    $\mathcal{D}_i$ para estimar $p(\mathbf{x}|\omega_j, \mathcal{D})$

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

- Aprendizagem Bayesiana

  Usa um conjunto $\mathcal{D}$ de amostras observadas independentemente de acordo com uma distribuição de probabilidade fixa, mas desconhecida $p(\mathbf{x})$ para estimar $p(\mathbf{x}|\ \mathcal{D})$

- Distribuição de parâmetros

  – densidade de probabilidade $p(\mathbf{x})$ desconhecida
  – assume-se forma paramétrica de $p(\mathbf{x})$ conhecida: $p(\mathbf{x}|\boldsymbol{\theta})$
  – problema: determinar vetor de parâmetros $\boldsymbol{\theta}$
  – conhecimento a priori sobre $\boldsymbol{\theta}$: $p(\boldsymbol{\theta})$
  – observações de $\mathcal{D}$ converte $p(\boldsymbol{\theta})$ em $p(\boldsymbol{\theta}|\mathcal{D})$

$$p(\mathbf{x}\,|\,\mathcal{D}) = \int p(\mathbf{x},\boldsymbol{\theta}\,|\,\mathcal{D})d\boldsymbol{\theta}$$

$$p(\mathbf{x},\boldsymbol{\theta}\,|\,\mathcal{D}) = p(\mathbf{x}\,|\,\boldsymbol{\theta},\mathcal{D})\,p(\boldsymbol{\theta}\,|\,\mathcal{D})$$

$$p(\mathbf{x}\,|\,\boldsymbol{\theta},\mathcal{D}) = p(\mathbf{x},\boldsymbol{\theta}) \qquad \text{seleção de } \mathbf{x} \text{ são independentes}$$

$$p(\mathbf{x}\,|\,\mathcal{D}) = \int p(\mathbf{x}\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta}\,|\,\mathcal{D})d\boldsymbol{\theta} \qquad \text{(integrar numericamente)} \qquad (25)$$

- Exemplo: caso Gaussiano

$p(\boldsymbol{\theta}|\mathcal{D}) = ?, \qquad p(\mathbf{x}|\mathcal{D}) = ?$

$p(\mathbf{x}|\mu) \sim N(\mu, \boldsymbol{\Sigma})$

1) Caso univariado: $p(\mu|\mathcal{D})$, μ é o único parâmetro desconhecido

$$p(x \mid \mu) = N(\mu, \sigma^2) \qquad (26)$$

$$p(\mu) = N(\mu_o, \sigma_o^2)$$

$\mathcal{D} = \{x_1, ...,x_n\}$

$$p(\mu \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mu)\, p(\mu)}{\int p(\mathcal{D} \mid \mu)\, p(\mu) d\mu} = \alpha \prod_{k=1}^{n} p(x_k \mid \mu)\, p(\mu)$$

$$p(\mu \mid \mathcal{D}) = \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k - \mu}{\sigma} \right)^2 \right]}^{p(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2}\sigma_o} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_o}{\sigma_o} \right)^2 \right]}^{p(\mu)}$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n} \left(\frac{\mu - x_k}{\sigma} \right)^2 + \left(\frac{\mu - \mu_o}{\sigma_o} \right)^2 \right) \right]$$

$$= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma^2} \right)\mu^2 - 2\left(\frac{1}{\sigma^2} \sum_{k=1}^{n} x_k + \frac{\mu_o}{\sigma_o^2} \right)\mu \right] \right] \qquad (29)$$

$$p(\mu \mid \mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n} \right)^2 \right] \sim N(\mu_n, \sigma_n) \qquad (30)$$

igualando (29) e (30)

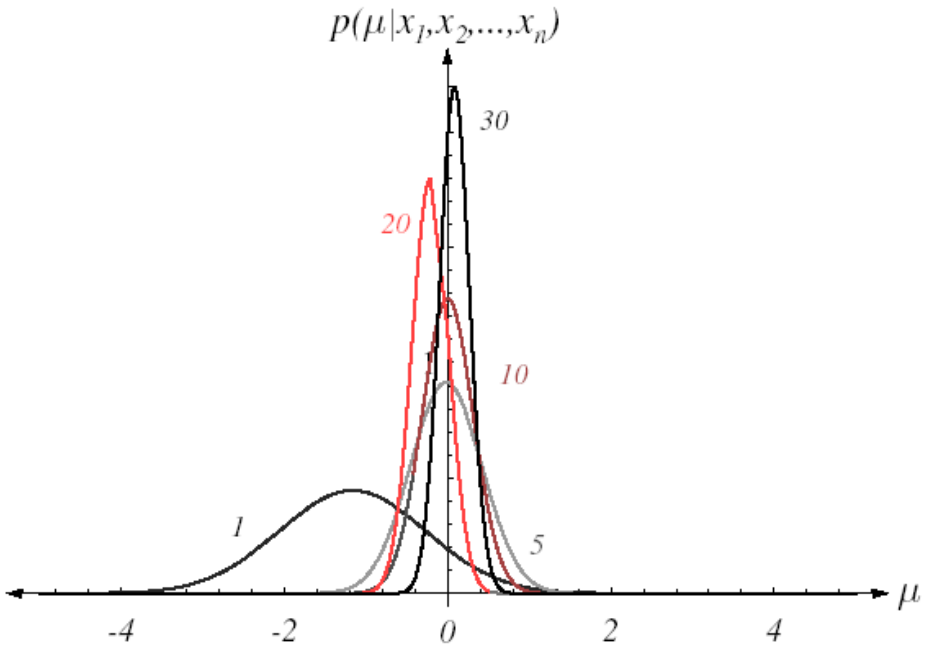
$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_o^2}$$

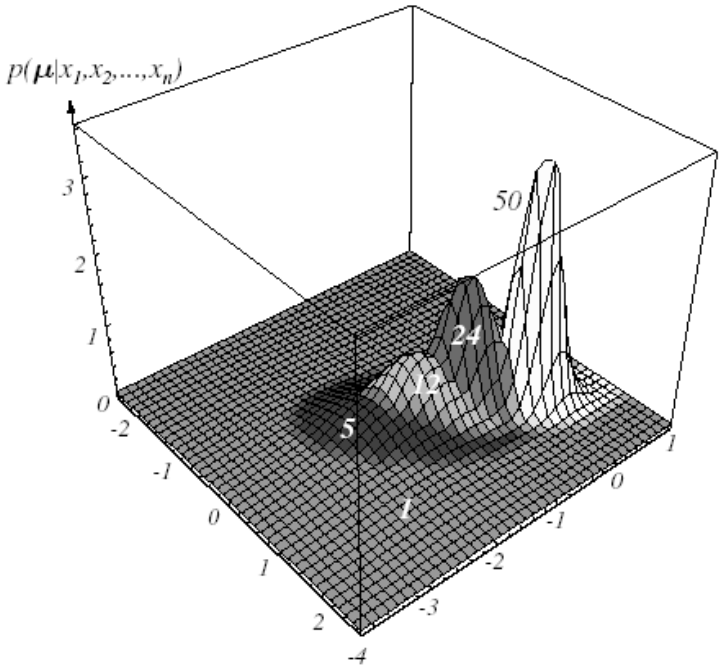
$$\frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_o}{\sigma_o^2} \qquad \hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$\mu_n = \left(\frac{n\sigma_o^2}{n\sigma_o^2 + \sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_o^2 + \sigma^2}\mu_o$$

$$\sigma_n^2 = \frac{\sigma_o^2\sigma^2}{n\sigma_o^2 + \sigma^2}$$

- Aprendizagem Bayesiana





2) Caso univariado: $p(x|\mathcal{D})$

$$p(x \mid \mathcal{D}) = \int p(x \mid \theta)\, p(\theta \mid \mathcal{D}) d\theta \tag{25}$$

$$p(x \mid \mu) = N(\mu, \sigma^2) \tag{26}$$

$$p(\mu \mid \mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n} \right)^2 \right] \sim N(\mu_n, \sigma_n) \tag{30}$$

$$p(x \mid \mathcal{D}) = \int p(x \mid \mu)\, p(\mu \mid \mathcal{D}) d\mu$$

$$= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right] \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_n}{\sigma_n} \right)^2 \right] d\mu$$

$$= \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2} \frac{(x-\mu_n)^2}{\sigma^2 + \sigma_n^2} \right] f(\sigma, \sigma_n)$$

$$f(\sigma, \sigma_n) = \int \exp\left[-\frac{1}{2} \frac{\sigma^2 + \sigma_n^2}{\sigma^2 \sigma_n^2} \left(\mu - \frac{\sigma_n^2 x + \sigma^2 \mu_n}{\sigma^2 + \sigma_n^2} \right)^2 \right] d\mu$$

$$p(x \mid \mathcal{D}) \sim N(\mu_n, \sigma^2 + \sigma_n^2) = p(x \mid \omega_j, \mathcal{D}_j)$$

3) Caso Gaussiano multivariado

$p(\boldsymbol{\theta}|\mathcal{D}) = ?$ $\quad p(\mathbf{x}|\mathcal{D}) = ?$

$p(\mathbf{x}|\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\quad p(\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}_o, \boldsymbol{\Sigma}_o)$

$\mathcal{D} = \{\mathbf{x}_1, ...,\mathbf{x}_n\}$, $\mathbf{x}_1, ...,\mathbf{x}_n$ amostras independentes

Após observar as n amostras de $\mathcal{D}$ e usando a fórmula de Bayes:

$$p(\boldsymbol{\mu} \mid \mathcal{D}) = \alpha \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\mu})\, p(\boldsymbol{\mu}) \qquad (39)$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}^t (n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_o^{-1})\boldsymbol{\mu} - 2\boldsymbol{\mu}^t \left(\boldsymbol{\Sigma}^{-1} \sum_{k=1}^{n} \mathbf{x}_k + \boldsymbol{\Sigma}_o^{-1}\boldsymbol{\mu}_o \right) \right) \right]$$

$p(\boldsymbol{\mu}|\mathcal{D})$ tem a forma

$$p(\boldsymbol{\mu} \mid \mathcal{D}) = \alpha'' \exp\left[-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_n)^t \boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_n)\right] \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \qquad (40)$$

igualando (39) e (40)

$$\Sigma_n^{-1} = n\Sigma^{-1} + \Sigma_{\mathrm{o}}^{-1} \qquad \hat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

$$\Sigma_n^{-1}\boldsymbol{\mu}_n = n\Sigma^{-1}\hat{\boldsymbol{\mu}}_n + \Sigma_{\mathrm{o}}^{-1}\boldsymbol{\mu}_{\mathrm{o}}$$

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_{\mathrm{o}}\left(\boldsymbol{\Sigma}_{\mathrm{o}} + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\hat{\boldsymbol{\mu}}_n + \frac{1}{n}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_{\mathrm{o}} + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu}_{\mathrm{o}}$$

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_{\mathrm{o}}\left(\boldsymbol{\Sigma}_{\mathrm{o}} + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\frac{1}{n}\boldsymbol{\Sigma}$$

$$p(\mathbf{x} \mid \mathcal{D}) = \int p(\mathbf{x} \mid \boldsymbol{\mu})\, p(\boldsymbol{\mu} \mid \mathcal{D}) d\boldsymbol{\mu} = N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_n)$$

4-Teoria geral estimadores Bayesianos

- Hipóteses
  - forma da densidade $p(\mathbf{x}|\theta)$ é conhecida
  - valor de θ não é conhecido exatamente
  - conhecimento inicial sobre θ contido densidade *a priori* $p(\theta)$
  - restante do conhecimento sobre θ contido em conjunto $\mathcal{D}$
  - $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, cada $\mathbf{x}_i$ obtido independente de acordo com $p(\mathbf{x})$
  - $p(\mathbf{x})$ desconhecida

- Problema básico

determinar distribuição *a posteriori* $p(\boldsymbol{\theta}|\mathcal{D})$ pois com ela calculamos

$$p(x \mid \mathcal{D}) = \int p(x \mid \theta)\, p(\theta \mid \mathcal{D}) d\theta \tag{49}$$

solução:

1) fórmula de Bayes

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{50}$$

3) hipótese de independência

$$p(\mathcal{D} \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta}) \tag{51}$$

- Análise

1) $p(\mathcal{D}|\boldsymbol{\theta})$ tem um pico em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$

$p(\boldsymbol{\theta}) \neq 0$ para $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ e não varia significativamente na vizinhança

(50) $\rightarrow$ $p(\boldsymbol{\theta}|\mathcal{D})$ tem um pico neste ponto

(49) $\rightarrow$ $p(x\,|\,\mathcal{D}) \approx p(\mathbf{x}\,|\,\hat{\boldsymbol{\theta}})$

mesmo resultado que o de MV se $\hat{\boldsymbol{\theta}}$ fosse o verdadeiro

2) se pico de $p(\mathcal{D}|\boldsymbol{\theta})$ é muito acentuado, a influência da informação *a priori* sobre incerteza de $\boldsymbol{\theta}$ pode ser desprezada

3) solução Bayesiana usa toda informação disponível

- Aprendizagem Bayesiana incremental

1) $\mathcal{D}^n = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$

2) de (51), se $n > 1$

$$p(\mathcal{D}^n \mid \boldsymbol{\theta}) = p(\mathbf{x}_n \mid \boldsymbol{\theta})\, p(\mathcal{D}^{n\text{-}1} \mid \boldsymbol{\theta}) \qquad (52)$$

3) densidade *a posteriori*

$$p(\boldsymbol{\theta} \mid \mathcal{D}^n) = \frac{p(\mathbf{x}^n \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D}^{n-1})}{\int p(\mathcal{D}^n \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D}^{n-1}) d\boldsymbol{\theta}} \qquad (53)$$

$$p(\boldsymbol{\theta} \mid \mathcal{D}^0) = p(\boldsymbol{\theta})$$

- Exemplo: caso unidimensional, distribuição uniforme

$$p(x \mid \boldsymbol{\theta}) \sim U(0,\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \textit{caso contrário} \end{cases}$$

$\mathcal{D} = \{4, 7, 2, 8\}$

$$p(\theta \mid \mathcal{D}^0) = p(\theta) = U(0,10)$$

1) $x_1 = 4$ e usando (53)

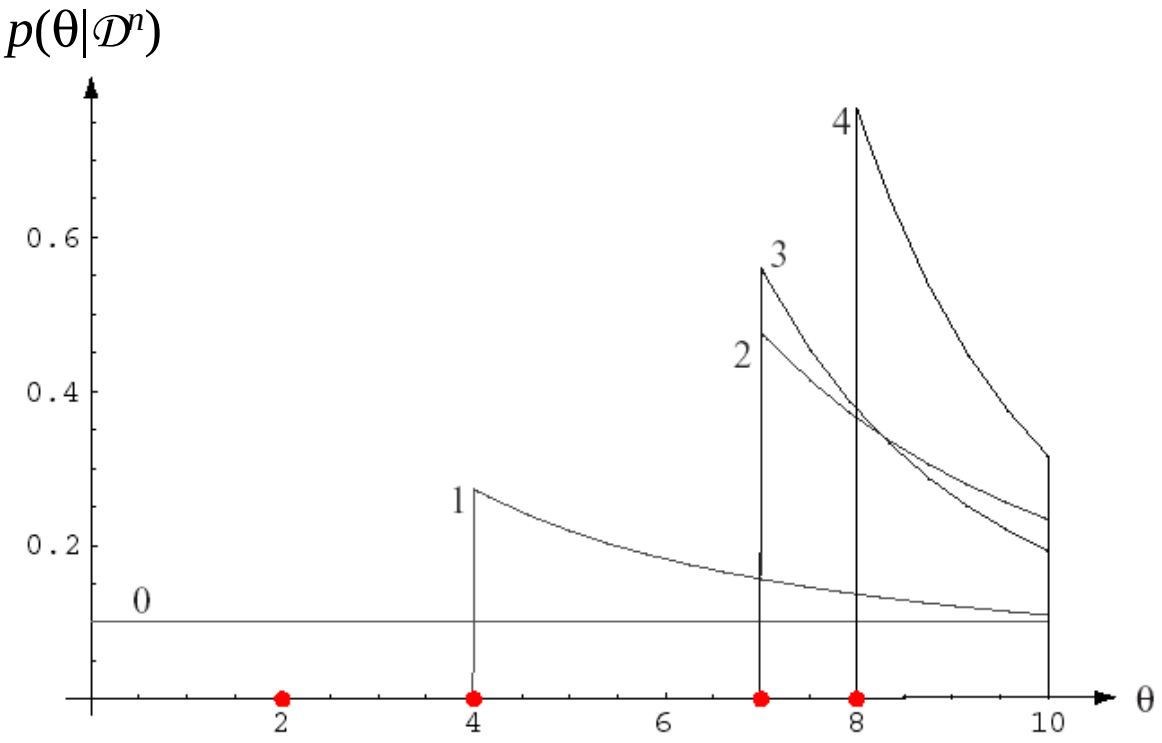
$$p(\theta \mid \mathcal{D}^1) \;\alpha\; p(x \mid \theta)\, p(\theta \mid \mathcal{D}^0) = \begin{cases} 1/\theta & 4 \le \theta \le 10 \\ 0 & \text{c.c.} \end{cases}$$

2) $x_2 = 7$ e usando (53)

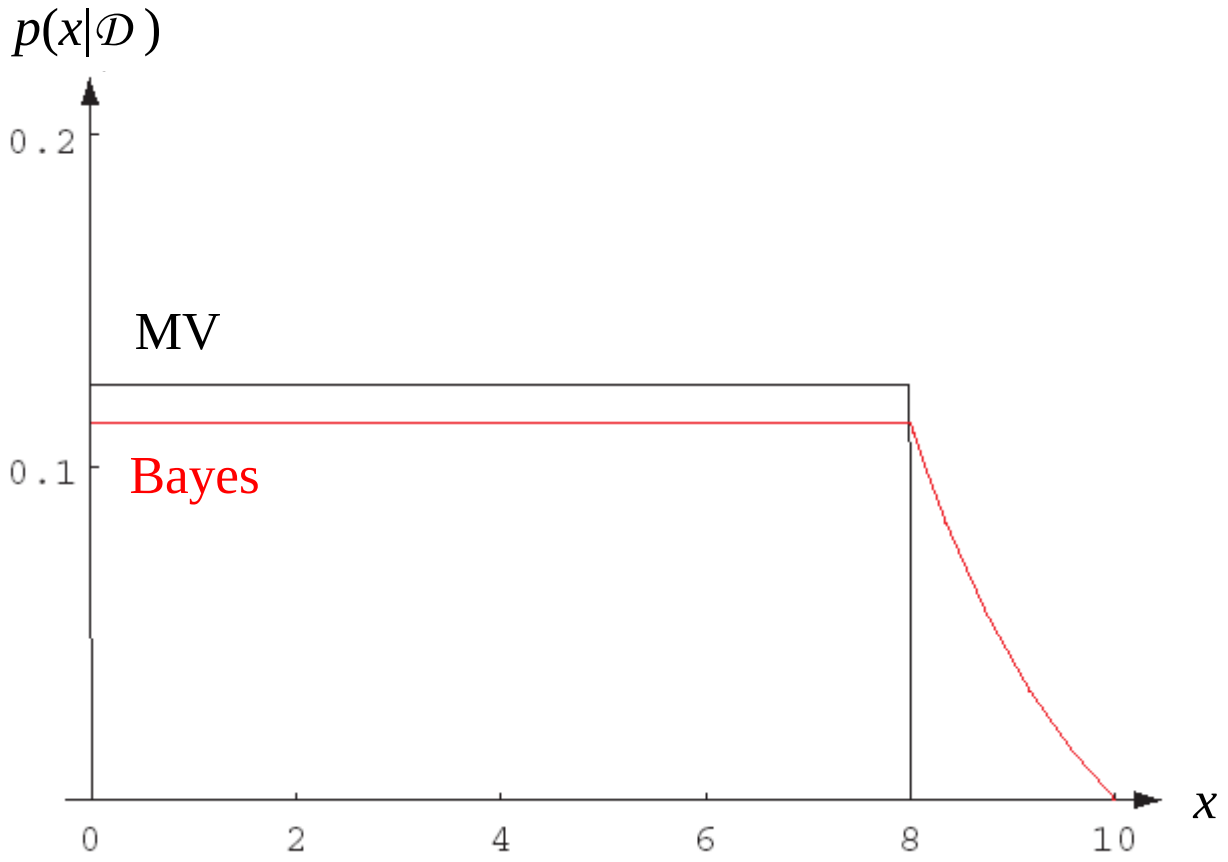
$$p(\theta \mid \mathcal{D}^2) \;\alpha\; p(x \mid \theta)\, p(\theta \mid \mathcal{D}^1) = \begin{cases} 1/\theta^2 & 7 \le \theta \le 10 \\ 0 & \text{c.c.} \end{cases}$$

n) $x_n = 8$ ($n = 4$) e usando (53)

$$p(\theta \mid \mathcal{D}^n) \;\alpha\; p(x \mid \theta)\, p(\theta \mid \mathcal{D}^{n-1}) = \begin{cases} 1/\theta^n & \max\limits_{x}[\mathcal{D}^n] \le \theta \le 10 \\ 0 & \text{c.c.} \end{cases}$$



$$p(\theta \mid \mathcal{D}^n) \;\alpha\; p(x \mid \theta)\, p(\theta \mid \mathcal{D}^{n-1}) = \begin{cases} 1/\theta^n & \max_x[\mathcal{D}^n] \le \theta \le 10 \\ 0 & \text{c.c.} \end{cases}$$



$$p(x \mid \mathcal{D}) = \int p(x \mid \theta)\, p(\theta \mid \mathcal{D}) d\theta$$

5-Problemas de dimensionalidade

- Questões
  - como a precisão de classificação depende da:
    - dimensão do espaço de atributos
    - quantidade de amostras de treinamento
  - complexidade computacional do classificador
  - *overfitting*

- Precisão, dimensão e quantidade dados treinamento

  – resultados teóricos para atributos independentes

  – exemplo com dois atributos: $p(\mathbf{x}|\omega_j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}),\ j = 1, 2$

  $P(\omega_1) = P(\omega_2)$

  erro classificação Bayes

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

$$r^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

– $P(e)$ diminui quando r aumenta; $P(e) \rightarrow 0$ quando $r \rightarrow \infty$

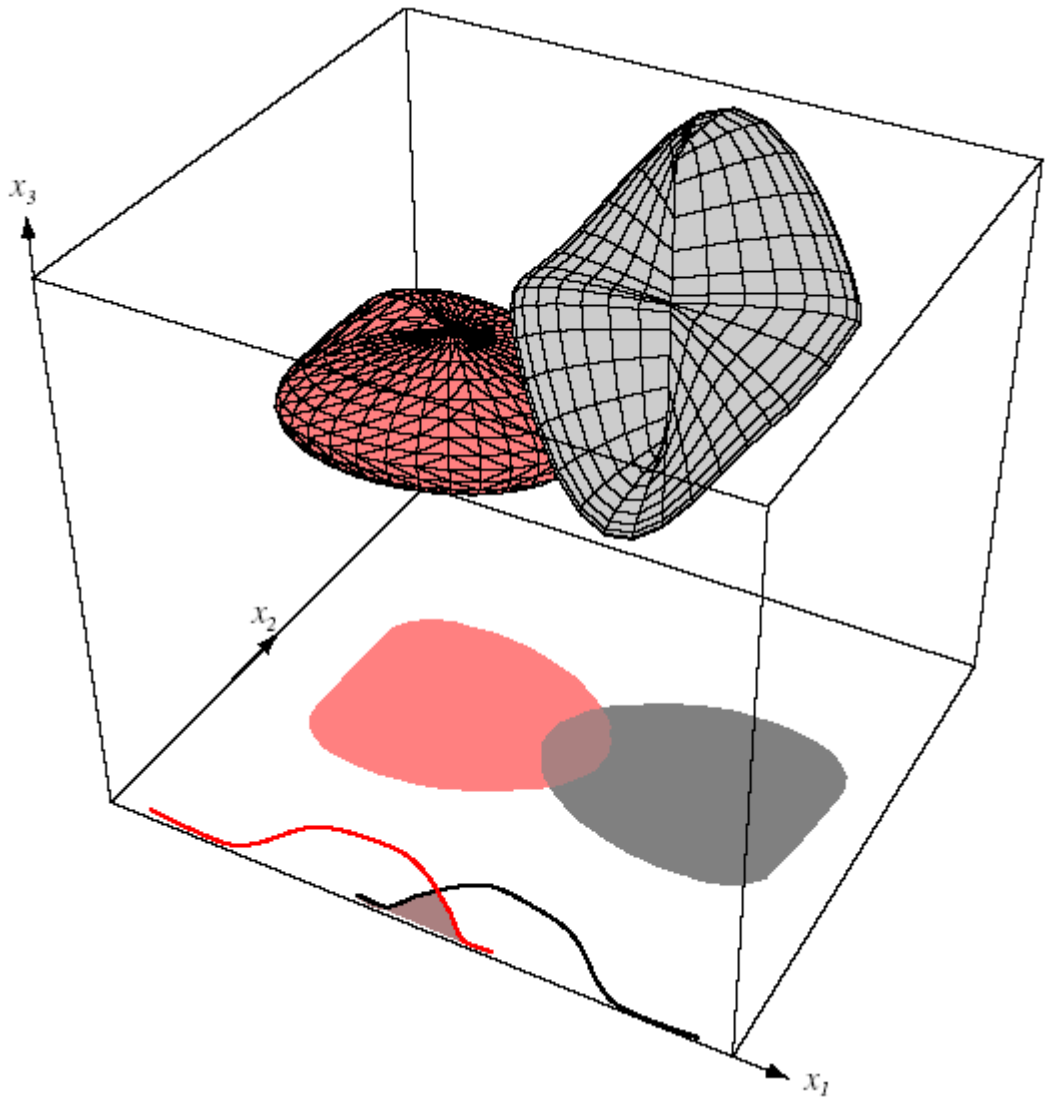
– caso condicionalmente independente $\mathbf{\Sigma} = \text{diag}(\sigma_1{}^2, ..., \sigma_d{}^2)$

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

– atributos mais relevantes: aqueles em que a diferença das médias é grande comparada com o desvio padrão

– atributo é útil se suas médias para os classificadores diferem

– como reduzir erro? adicionar novos atributos independentes


x_3
x_2
x_1

– observa-se na prática que acrescentar atributos além de um certo limite deteriora o desempenho do classificador.

– razões principais são as seguintes:

1. hipóteses erradas sobre o modelo
   (e.g. Gaussiano, condicionamento)

2. número amostras treinamento pequeno

- Complexidade computacional

  – parâmetros distribuição normal para o MV

$$g(\mathbf{x}) = -\frac{1}{2}\overbrace{(\mathbf{x}-\hat{\boldsymbol{\mu}})^t}^{O(dn)}\overbrace{\hat{\boldsymbol{\Sigma}}^{-1}}^{O(nd^2)}(\mathbf{x}-\hat{\boldsymbol{\mu}}) - \overbrace{\frac{d}{2}\ln 2\pi}^{O(1)} - \overbrace{\frac{1}{2}\ln|\hat{\boldsymbol{\Sigma}}|}^{O(d^2n)} + \overbrace{\ln P(\omega)}^{O(n)}$$

  – classificação: $O(d^2)$

  – aprendizagem Bayesiana: mais complexo devido à integração

- *Overftting*

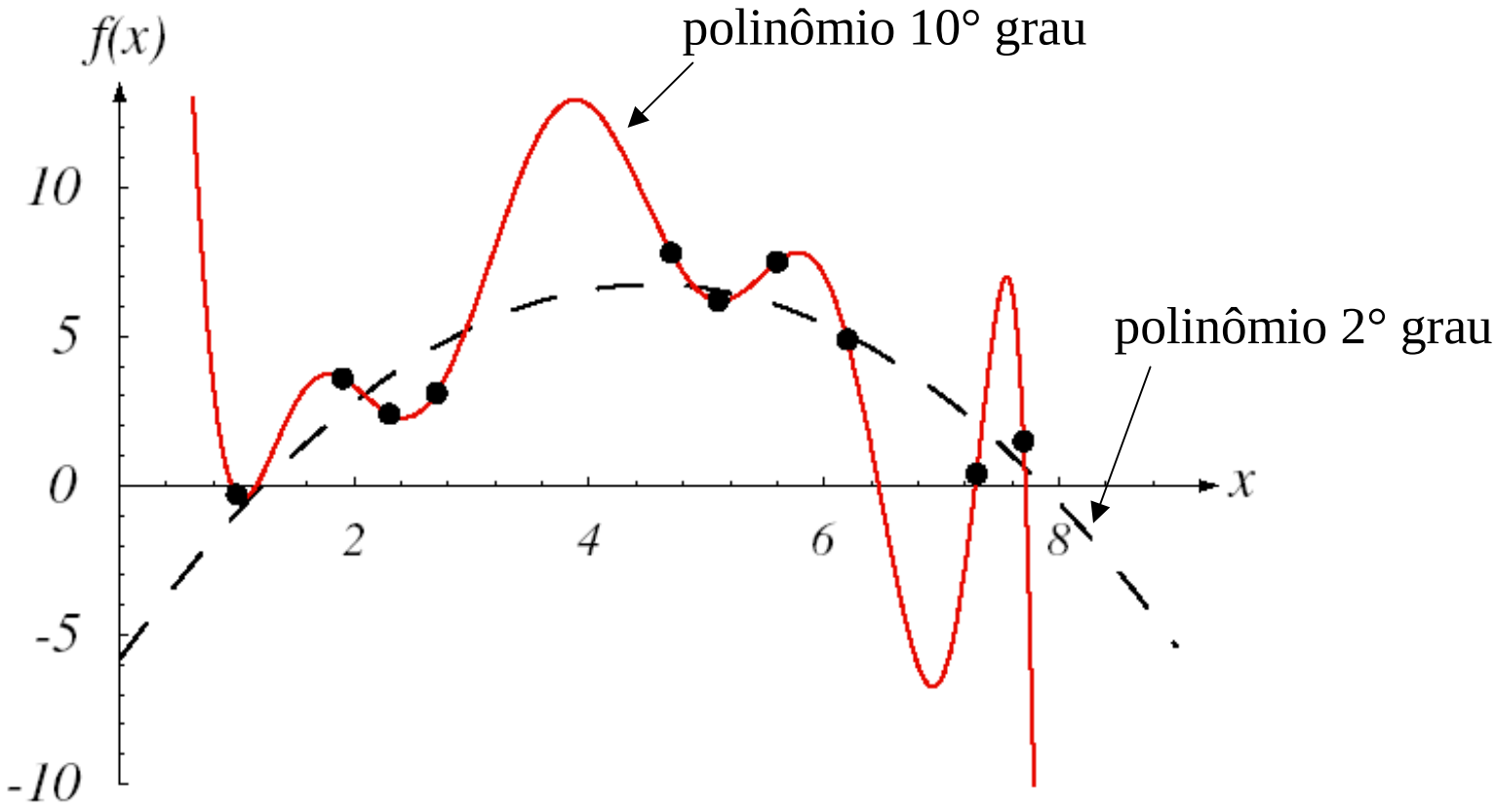
  – erro dados de treinamento $\times$ generalização

  – em geral, interpolação ou extrapolação só pode ser feita de forma confiável se a solução é sobredeterminada, isto é, o número de pontos é maior do que o número de parâmetros a serem determinados.

  – heurísticas: e.g. *shrinkage*

$$\boldsymbol{\Sigma}_i(\alpha) = \frac{(1-\alpha)n_i\boldsymbol{\Sigma}_i + \alpha n\boldsymbol{\Sigma}}{(1-\alpha)n_i + \alpha n}$$

$$\boldsymbol{\Sigma}(\beta) = (1-\beta)\boldsymbol{\Sigma} + \beta\mathbf{I} \qquad 0 < \alpha, \beta < 1$$



$$f(x) = ax^2 + bx + c + \varepsilon \qquad p(\varepsilon) \sim N(0, \sigma^2)$$

6-Modelos de Markov

- Modelos de Markov de 1ª ordem

  – $\omega(t)$: estado em t

  $\boldsymbol{\omega}^T = \{\omega(1), \omega(2), ..., \omega(T)\}$ sequência de tamanho T

  exemplo: $\boldsymbol{\omega}^6 = \{\omega_1, \omega_4, \omega_2, \omega_2, \omega_1, \omega_2\}$

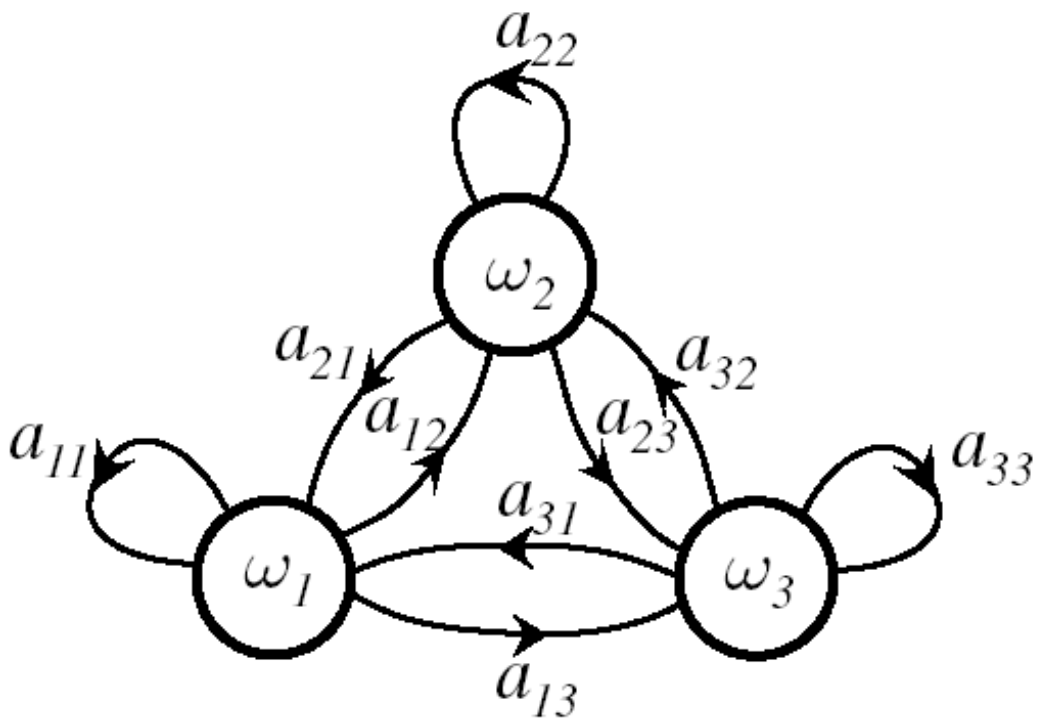
  – $P(\omega_j(t+1)|\omega_i(t)) = a_{ij}$ probabilidade de transição de estado

  – modelo $\boldsymbol{\theta}$: conjunto de todos os valores a_{ij}

  probabilidade modelo gerar $\boldsymbol{\omega}^T$: produto das probabilidades

  exemplo: $P(\boldsymbol{\omega}^6|\boldsymbol{\theta}) = a_{14}a_{42}a_{22}a_{21}a_{14}$

- Modelo Markov de 1ª ordem



Modelo de Markov de 1ª ordem discreto: estado em $t + 1$ depende somente do estado em t e das probabilidades de transição.

- Hidden Markov Models de 1ª ordem

  – $\omega(t)$: estado sistema em t

  – estado emite símbolos visíveis $v(t)$

  $\mathbf{V}^T = \{v(1), v(2), ..., v(T)\}$ sequência de símbolos visíveis

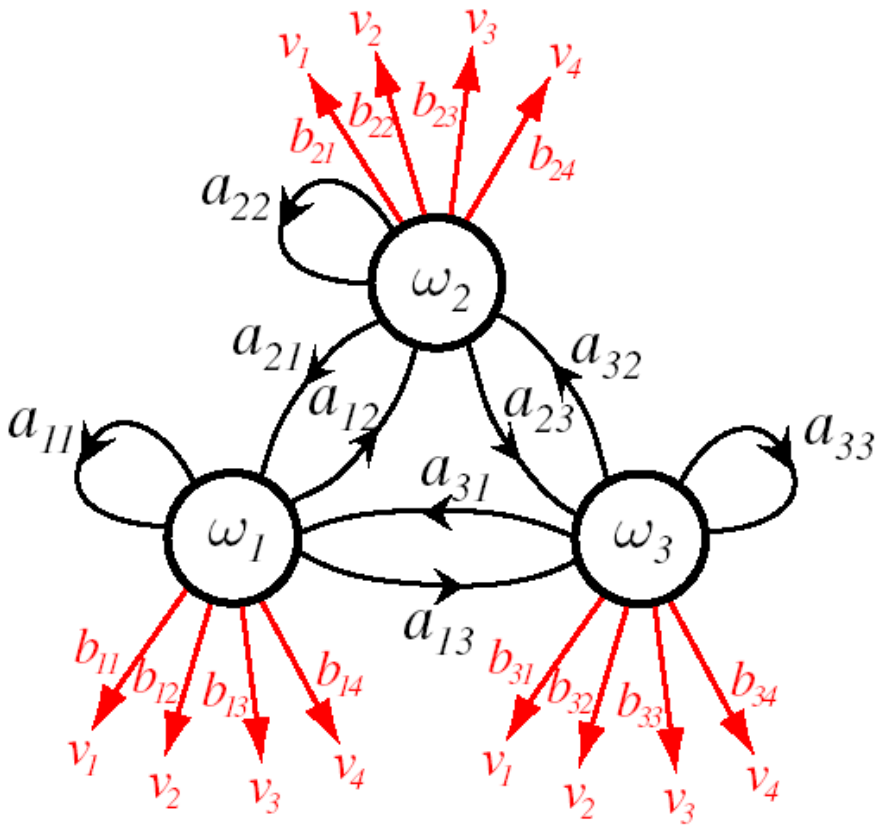
  exemplo: $\mathbf{V}^6 = \{v_5, v_1, v_1, v_5, v_2, v_3\}$

  – $P(v_k(t)|\omega_i(t)) = b_{jk}$ probabilidade de emitir símbolo v_k

  – ω não é observável; acesso somente a símbolos visíveis

  modelos escondidos de Markov de 1ª ordem

- Modelo escondido de Markov de 1ª ordem



- Características
  - grafos são máquinas de estado finito
  - grafos + probabilidades transição = modelos Markov
  - MM são estritamente causais
  - ergódigos: $a_{ij} \neq 0 \ \forall i, j$
  - absorção: estado ω_o com $a_{oo} = 1$

$$a_{ij} = P(\omega_j(t+1) \mid \omega_i(t)) \qquad \sum_j a_{ij} = 1 \quad \forall i$$

$$b_{jk} = P(v_k(t) \mid \omega_j(t)) \qquad \sum_k a_{jk} = 1 \quad \forall j$$

- Problemas importantes em HMM

1) Avaliação: temos HMM com a_{ij} e b_{jk} ; qual probabilidade que uma sequência particular $\mathbf{V}^T$ foi gerada pelo modelo?

2) Decodificação: temos HMM e $\mathbf{V}^T$; determinar a sequência mais provável de estados escondidos $\boldsymbol{\omega}^T$ que produziu $\mathbf{V}^T$.

3) Aprendizagem: dado a estrutura do modelo e um conjunto de observações de treinamento, determinar a_{ij} e b_{jk} .

1) Avaliação

dado um modelo HMM, determinar a probabilidade que este modelo gerou uma sequência particular $\mathbf{V}^T$ de estados visíveis

$$P(\mathbf{V}^T) = \sum_{r=1}^{r_{max}} P(\mathbf{V}^T \,|\boldsymbol{\omega}_r^T) P(\boldsymbol{\omega}_r^T)$$

$$\boldsymbol{\omega}_r^T = \{\omega(1), \omega(2), \ldots, \omega(T)\}$$

c estados escondidos $\rightarrow r_{max} = c^T$ termos possíveis

$$P(\boldsymbol{\omega}_r^T) = \prod_{i=1}^{T} P(\omega(t) \,|\, \omega(t-1)$$

$$P(\mathbf{V}^T \,|\, \boldsymbol{\omega}_r{}^T) = \prod_{i=1}^{T} P(v(t) \,|\, \omega(t)$$

$$P(\mathbf{V}^T) = \sum_{r=1}^{r_{max}} \prod_{t=1}^{T} P(v(t) \mid \omega(t)) P(\omega(t) \mid \omega(t-1)) \qquad (135)$$

cálculo de (135): $O(Tc^T)$ $\quad c = 10$ e $T = 20 \rightarrow 10^{21}$ operações

$P(\mathbf{V}^T)$ calculado recursivamente: envolve $v(t)$, $\omega(t)$ e $\omega(t-1)$

$$\alpha_j(t) = \begin{cases} 0 & t = 0 \quad j \neq \text{estado inicial} \\ 1 & t = 0 \quad j = \text{estado inicial} \\ [\sum_i \alpha_i(t-1) a_{ij}] b_{jk} v(t) & c.c. \end{cases}$$

$b_{jk}v(t)$: probabilidade b_{jk} associada estado visível $v(t)$

Algoritmo HMM Forward

1 **inicialize** $t \leftarrow 0$, a_{ij}, b_{jk}, sequência visível $\mathbf{V}^T$, $\alpha_j(0)$

2 **for** $t \leftarrow t + 1$

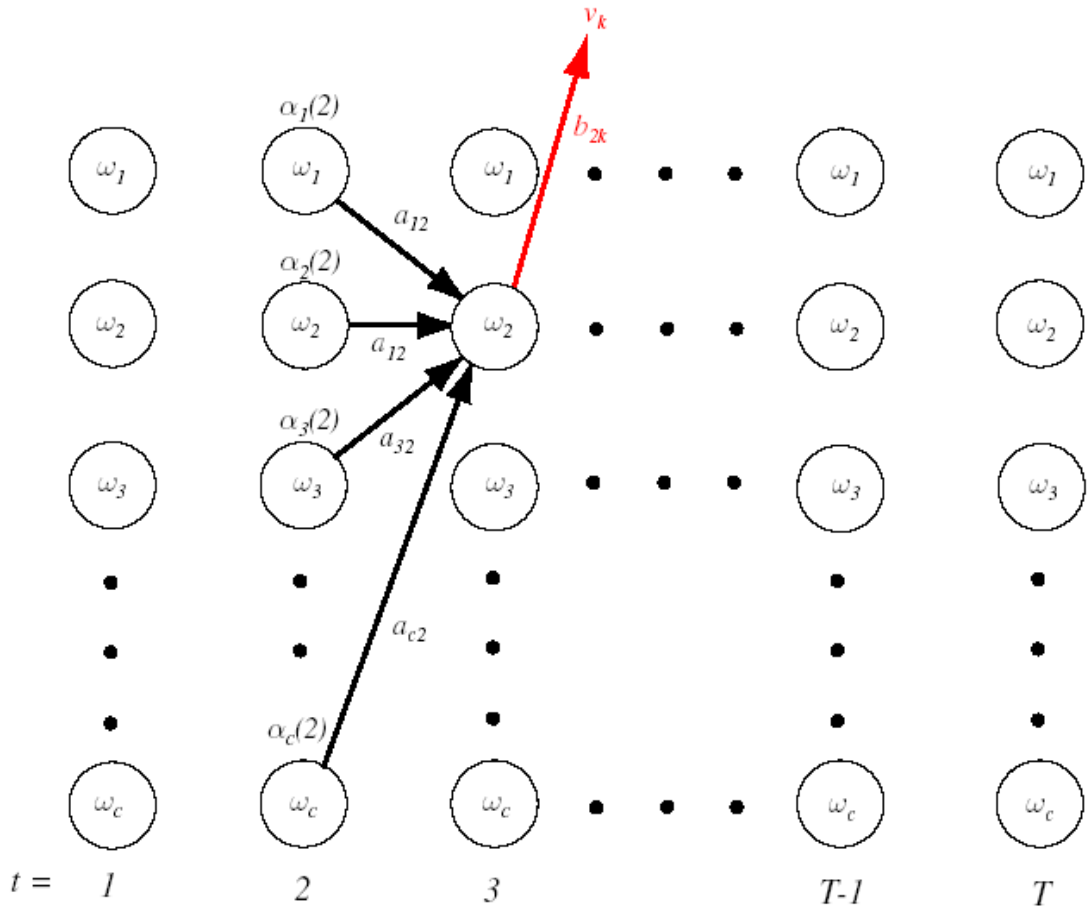
3 $\alpha_j(t) \leftarrow b_{jk}\, v(t)[\, \sum_{i = 1,..,c}(\alpha_i(t-1)\, a_{ij}\,]$

4 **until** $t = T$

5 **return** $P(\mathbf{V}^T) \leftarrow \alpha_0(T)$

$O(c^2T) \sim 2000$ operações para $c = 10$ e $T = 20$

classificação: Bayes $$P(\boldsymbol{\theta} \mid \mathbf{V}^T) = \frac{P(\mathbf{V}^T \mid \boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{V}^T)}$$



$$\alpha_2(3) = [\sum_{i=1}^{c} \alpha_i(2) a_{i2}] b_{2k}$$

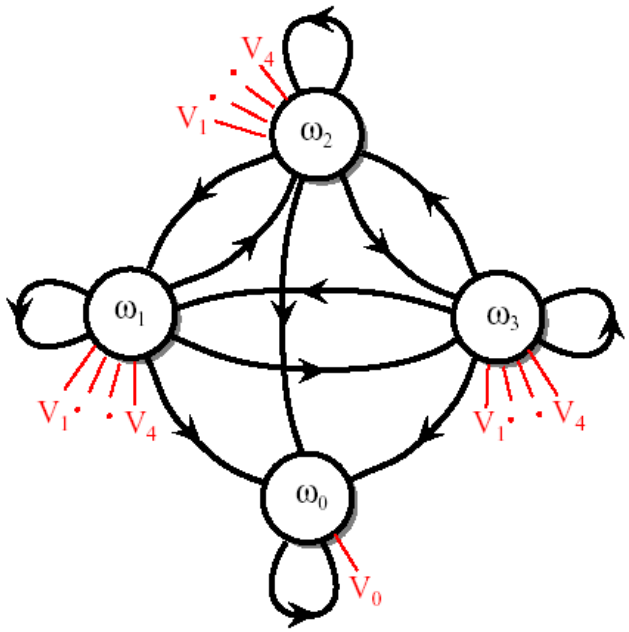
- Exemplo: avaliação

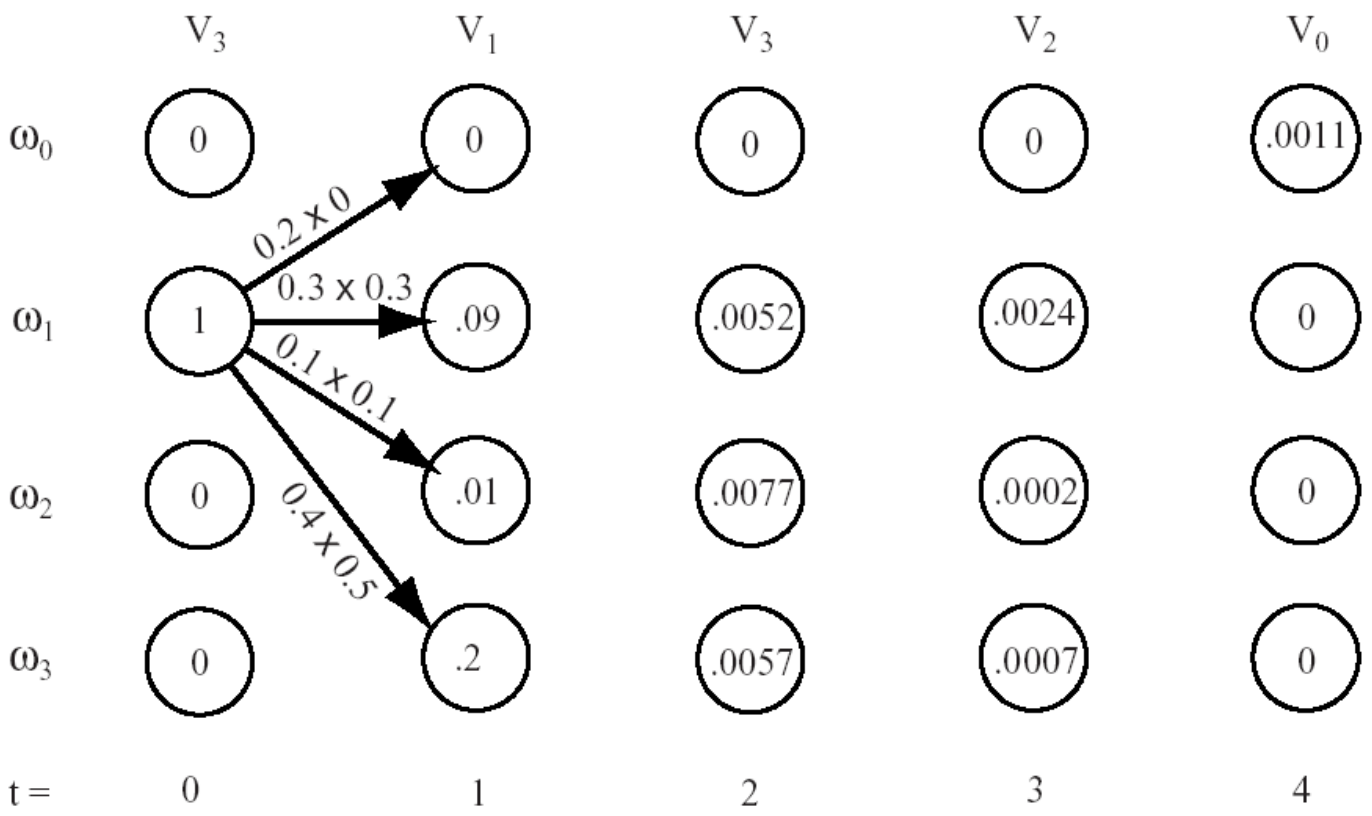
$\mathbf{V}^4 = \{v_1, v_3, v_2, v_0\}$

v_0: *absorbing state*

$$[a_{ij}] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.2 & 0.3 & 0.1 & 0.4 \\ 0.2 & 0.5 & 0.2 & 0.1 \\ 0.8 & 0.1 & 0.0 & 0.1 \end{bmatrix}$$

$$[b_{jk}] = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0.3 & 0.4 & 0.1 & 0.2 \\ 0 & 0.1 & 0.1 & 0.7 & 0.1 \\ 0 & 0.5 & 0.2 & 0.1 & 0.2 \end{bmatrix}$$





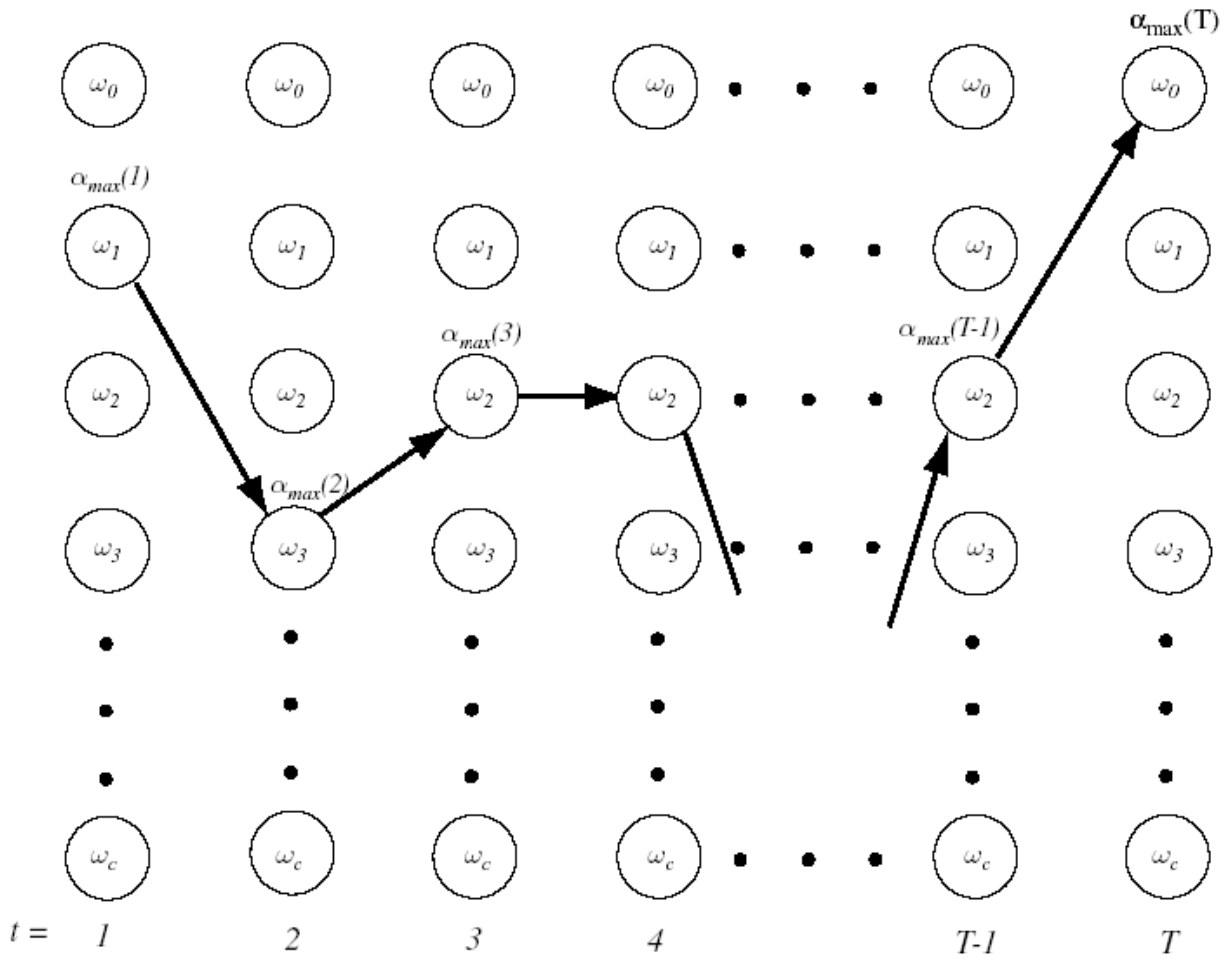
$$P(\mathbf{V}^T|\theta) = 0.0011$$

2) Decodificação

dada uma sequência $\mathbf{V}^T$ de estados visíveis, determinar a sequência mais provável de estados escondidos

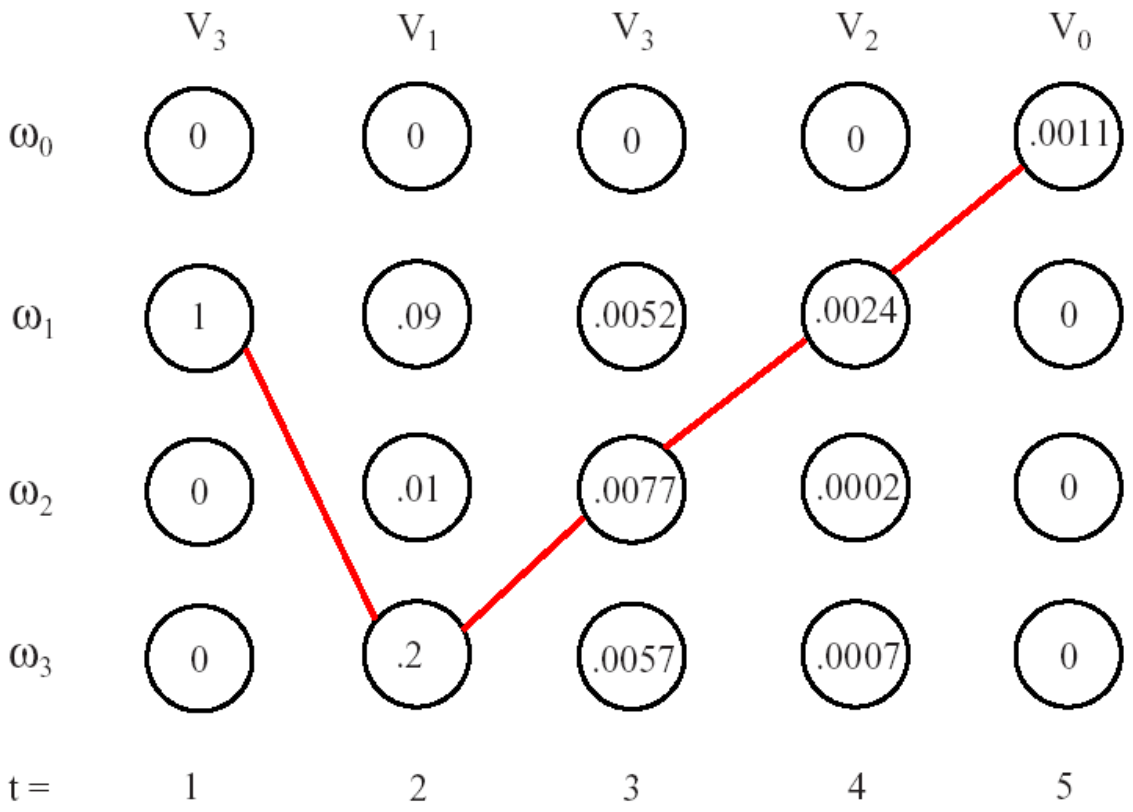
Algoritmo HMM Decoding

1 **inicialize** $t \leftarrow 0$, Path $\leftarrow \{\}$
2 **for** $t \leftarrow t + 1$
3 $j \leftarrow j + 1$
4 **for** $j \leftarrow j + 1$
5 $\alpha_j(t) \leftarrow b_{jk}\, v(t)[\, \sum_{i = 1,..,c}(\alpha_i(t-1)\, a_{ij}\,]$
6 **until** $t = T$
7 $j' = \arg\max_j \alpha_j(t)$
8 Append $\omega j'$ to Path
9 **until** $t = T$
10 **return** Path



obs: máximos locais → não garante consistência da solução global

- Exemplo: decodificação (exemplo de avaliação)



solução: $\{\omega_1, \omega_3, \omega_2, \omega_1, \omega_0\}$ (inconsistente !! $a_{32} = 0$)

3) Aprendizagem

determinar os parâmetros do modelo, a_{ij}, b_{jk}

algoritmo *forward-backward*

$\alpha_i(t)$: probabilidade modelo estar no estado $\omega_i(t)$ e gerou sequência de referência até t

$\beta_i(t)$: probabilidade modelo está no estado $\omega_i(t)$ e vai gerar sequência de referência de $t + 1$ até T

$$\beta_i(t) = \begin{cases} 0 & \omega_i(t) \neq \omega_0 \quad t = T \\ 1 & \omega_i(t) = \omega_0 \quad t = T \\ \sum_j \beta_j(t+1) a_{ij} b_{jk} v(t+1) & c.c. \end{cases} \qquad (138)$$

- Justificativa de (138)

  – supor $\alpha_i(t)$ conhecido até $T - 1$

  – probabilidade que o modelo gerar o último símbolo visível ?

  – esta probabilidade é $\beta_i(T)$

  – $\beta_i(T)$ = probabilidade transição para $\omega_i(T)$

  $\times$

  probabilidade estado emitir símbolo visível correto

  – definição: $\beta_i(T) = 0$, se $\omega_i(T) \neq \omega_0$

  $\beta_i(T) = 1$, se $\omega_i(T) = \omega_0$

  – logo $\beta_i(T - 1) = \sum_j a_{ij} b_{jk}(T)\, \beta_i(T)$

– $\gamma_{ij}(t)$ probabilidade transição entre $\omega_i(t-1)$ e $\omega_j(t)$ dado que o modelo gerou toda a sequência de treinamento $\mathbf{V}^T$ em qualquer caminho

– definimos $\gamma_{ij}(t)$

$$\gamma_{ij}(t) = \frac{\alpha_i(t-1)a_{ij}b_{jk}\beta_j(t)}{P(\mathbf{V}^T \mid \boldsymbol{\theta})}$$

– $\gamma_{ij}(t)$ probabilidade transição de $\omega_i(t-1)$ para $\omega_j(t)$ dado que o modelo gerou a sequência visível $\mathbf{V}^T$ completamente

- Estimativa das probabilidades de transição

$\sum_{t=1}^{T} \gamma_{ij}(t)$ número esperado de transições de $\omega_i(t-1)$ para $\omega_j(t)$ na sequência de treinamento

$\sum_{t=1}^{T} \sum_{k} \gamma_{ik}$ número total esperado de transições de $\omega_j(t)$ para qualquer outro estado

$$\hat{a}_{ij}(t) = \frac{\sum_{t=1}^{T} \gamma_{ij}(t)}{\sum_{t=1}^{t} \sum_{k} \gamma_{ik}(t)} \quad (140) \qquad \hat{b}_{jk}(t) = \frac{\sum_{t=1}^{T} \sum_{l} \gamma_{jl}(t)}{\sum_{t=1}^{t} \sum_{l} \gamma_{jl}(t)} \quad (141)$$

$v(t)=v_k$

Algoritmo Forward-Backward

inicialize $z \leftarrow 0$, a_{ij}, b_{jk}, sequência treinamento $\mathbf{V}^T$, ε convergência
    do $t \leftarrow t + 1$
        calcular $\hat{a}(z)$ usando $a(z-1)$ e (140)
        calcular $\hat{b}(z)$ usando $b(z-1)$ e (141)
        $a_{ij}(z) \leftarrow \hat{a}_{ij}(z-1)$
        $b_{jk}(z) \leftarrow \hat{b}_{jk}$
    until $\max[a_{ij}(z) - a_{ij}(z-1), a_{ij}(z) - a_{ij}(z-1)] < \varepsilon$
return $a_{ij} \leftarrow a_{ij}(z)$, $b_{ij} \leftarrow b_{ij}(z)$,

9-Resumo

- Forma densidades condicionais classe conhecida
- Aprendizagem
  - estimação de parâmetros MV
  - estimação densidades Bayes
- Impacto da dimensão espaço atributos e dados de treinamento
- Decisão sequencial com modelos de Markov

Observação

Este material refere-se às notas de aula do curso CT 720 Tópicos Especiais em Aprendizagem de Máquina e Classificação de Padrões da Faculdade de Engenharia Elétrica e de Computação da Unicamp e do Centro Federal de Educação Tecnológica do Estado de Minas Gerais. Não substitui o livro texto, as referências recomendadas e nem as aulas expositivas. Este material não pode ser reproduzido sem autorização prévia dos autores. Quando autorizado, seu uso é exclusivo para atividades de ensino e pesquisa em instituições sem fins lucrativos.