بسم الله الرحمن الرحیم

بازشناسی الگو

درس ۵

طبقه‌بندی مبتنی بر نظریه‌ی تصمیم بیز

روش‌های تخمین پارامتری تابع چگالی احتمال مجهول

Classification Based on Bayes Decision Theory

Parametric Estimation Methods for Unknown Probability Density Functions

کاظم فولادی قلعه

دانشکده مهندسی، پردیس فارابی

دانشگاه تهران

http://courses.fouladi.ir/pr

تخمین توابع چگالی احتمال مجهول

ESTIMATION OF UNKNOWN PROBABILITY DENSITY FUNCTIONS

روش طبقه‌بندی بیزی نیازمند **مدل** است.

مدل: توابع چگالی احتمال پیشین و پسین طبقه‌ها

تاکنون فرض شده بود که توابع pdf معلوم هستند، اما معمولاً این‌گونه نیست:
در بسیاری از مسائل باید pdf را از روی داده‌های آموزشی موجود، تخمین زد.

روی‌کردهای تخمین توابع چگالی احتمال مجهول	
پارامتری *Parametric*	**ناپارامتری** *Nonparametric*
pdf را می‌دانیم، اما پارامترهای آن مجهول است.	pdf را نمی‌دانیم، اما برخی آماره‌های آن (مثلاً μ یا σ^2) معلوم است.
روش‌ها: o Maximum Likelihood (ML) o Maximum a Posteriori Probability (MAP) o Maximum Entropy (ME) o Bayesian Inference o Mixture Models	روش‌ها: o Parzen Windows o *k* Nearest Neighbor

تخمین توابع چگالی احتمال مجهول

روی‌کرد پارامتری

ESTIMATION OF UNKNOWN PROBABILITY DENSITY FUNCTIONS

روش طبقه‌بندی بیزی نیازمند **مدل** است.

مدل: توابع چگالی احتمال پیشین و پسین طبقه‌ها

تاکنون فرض شده بود که توابع pdf معلوم هستند، اما معمولاً این‌گونه نیست:
در بسیاری از مسائل باید pdf را از روی داده‌های آموزشی موجود، تخمین زد.

روی‌کردهای تخمین توابع چگالی احتمال مجهول

پارامتری *Parametric*	ناپارامتری *Nonparametric*
pdf را می‌دانیم، اما پارامترهای آن مجهول است.	pdf را نمی‌دانیم، اما برخی آماره‌های آن (مثلاً μ یا σ^2) معلوم است.
روش‌ها: o Maximum Likelihood (ML) o Maximum a Posteriori Probability (MAP) o Maximum Entropy (ME) o Bayesian Inference o Mixture Models	روش‌ها: o Parzen Windows o *k* Nearest Neighbor

بازشناسی الگو

طبقه‌بندی مبتنی بر نظریه‌ی تصمیم بیز:
روش‌های تخمین پارامتری
تابع چگالی احتمال مجهول

۱

روش تخمین ماکزیمم درست‌نمایی

تخمین پارامتری توابع چگالی احتمال مجهول

روش ماکزیمم درست‌نمایی

MAXIMUM LIKELIHOOD (ML)

روش ML، پارامترها را به‌عنوان کمیت‌های **ثابت** اما **مجهول** می‌بیند.

$\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ مجموعه‌ی داده‌های آموزشی:
نمونه‌های مستقل با توزیع یکسان (iid)
$p(\mathbf{x}|\boldsymbol{\theta})$ بیرون کشیده شده از تابع چگالی احتمال

هدف: تخمین پارامتر مجهول $\boldsymbol{\theta}$ با استفاده از نمونه‌های آموزشی $\mathcal{D}$

تابع درست‌نمایی
Likelihood Function

تابع درست‌نمایی $\boldsymbol{\theta}$ نسبت به داده‌های $\mathcal{D}$

$$L(\boldsymbol{\theta}|\mathcal{D})$$

$$L(\boldsymbol{\theta}|\mathcal{D}) = p(\mathcal{D}|\boldsymbol{\theta}) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})$$

با فرض **استقلال آماری**
(بررسی وجود استقلال آماری دشوار است: اگر برای وابستگی دو متغیر دلیلی نداشتیم، آن دو را مستقل فرض می‌کنیم.)

تخمین پارامتری توابع چگالی احتمال مجهول

روش ماکزیمم درست‌نمایی

MAXIMUM LIKELIHOOD (ML)

تخمین درست‌نمایی MLE برای $\boldsymbol{\theta}$ مقداری از $\widehat{\boldsymbol{\theta}}$ است تابع درست‌نمایی را ماکزیمم می‌کند:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathcal{D})$$

معمولاً کار کردن با لگاریتم تابع درست‌نمایی ساده‌تر است، و به همان تخمین برای پارامتر می‌رسد.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}|\mathcal{D}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p(\mathbf{x}_i|\boldsymbol{\theta})$$

تابع لگاریتم درست‌نمایی
Log-Likelihood Function

لگاریتم تابع درست‌نمایی $\boldsymbol{\theta}$ نسبت به داده‌های $\mathcal{D}$

$$\log L(\boldsymbol{\theta}|\mathcal{D})$$

تخمین پارامتری توابع چگالی احتمال مجهول

روش ماکزیمم درست‌نمایی: محاسبه

MAXIMUM LIKELIHOOD (ML)

اگر تعداد پارامترها p باشد، آن‌گاه

$$\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_p})^T$$

و عملگر گرادیان به صورت زیر تعریف می‌شود:

$$\nabla_{\boldsymbol{\theta}} \equiv \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\theta_1}} \\ \vdots \\ \frac{\partial}{\partial \boldsymbol{\theta_p}} \end{bmatrix}$$

در این صورت، تخمین ماکزیمم درست‌نمایی $\boldsymbol{\theta}$ باید شرط لازم زیر را برآورده کند:

$$\nabla_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}|\mathcal{D}) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}_i|\boldsymbol{\theta}) = 0$$

(مشتق مساوی صفر)

انتخاب ماکزیمم سراسری از میان اکستریم‌های محلی حاصل از مشتق لازم است. همچنین، شرایط مرزی باید بررسی شود.

تخمین پارامتری توابع چگالی احتمال مجهول

روش ماکزیمم درست‌نمایی: خصوصیات (۱): ناارییی مجانبی

MAXIMUM LIKELIHOOD (ML)

تخمین ML به‌طور مجانبی **ناریب** (unbiased) است.

اگر θ_0 مقدار واقعی پارامتر مجهول θ باشد، آن‌گاه

$$\lim_{N\to\infty} E\{\hat{\theta}_{\mathrm{ML}}\} = \theta_0$$

یعنی: برآورد ML در حالت میانگین به مقدار واقعی پارامتر همگرا می‌شود.

$\hat{\theta}_{\mathrm{ML}}$ به خودی خود یک متغیر تصادفی است
که برای مجموعه داده‌های مختلف $\mathcal{D}$ برآوردهای مختلفی دارد.

تخمین پارامتری توابع چگالی احتمال مجهول

روش ماکزیمم درست‌نمایی: خصوصیات (۲): سازگاری مجانبی

MAXIMUM LIKELIHOOD (ML)

تخمین ML به‌طور مجانبی **سازگار** (consistent) است.

اگر θ_0 مقدار واقعی پارامتر مجهول θ باشد، آن‌گاه

$$\lim_{N\to\infty} E\{||\hat{\theta}_{\mathrm{ML}} - \theta_0||^2\} = 0$$

یعنی: برآورد ML در حالت میانگین مربعات همگرا می‌شود $\Leftarrow$
برای Nهای بزرگ، واریانس برآوردهای ML به سمت صفر میل می‌کند.

$\hat{\theta}_{\mathrm{ML}}$ به خودی خود یک متغیر تصادفی است
که برای مجموعه داده‌های مختلف $\mathcal{D}$ برآوردهای مختلفی دارد.

تخمین پارامتری توابع چگالی احتمال مجهول

روش ماکزیمم درست‌نمایی: خصوصیات (۳): همگرایی به توزیع گاوسی

MAXIMUM LIKELIHOOD (ML)

تابع چگالی تخمین ML برای $N \to \infty$ به سمت توزیع گاوسی با میانگین θ_0 میل می‌کند.

زیرا:

۱) قضیه‌ی حد مرکزی

۲) ML وابسته به مجموع N متغیر تصادفی است.

تخمین پارامتری توابع چگالی احتمال مجهول

روش ماکزیمم درست‌نمایی: خصوصیات (۴): تغییرناپذیری

MAXIMUM LIKELIHOOD (ML)

اگر $\widehat{\boldsymbol{\theta}}$ تخمین ML پارامتر $\boldsymbol{\theta}$ باشد،
آن‌گاه برای هر تابع f دلخواه،
تخمین ML پارامتر $f(\boldsymbol{\theta})$ برابر با $f(\widehat{\boldsymbol{\theta}})$ است.

تخمین پارامتری توابع چگالی احتمال مجهول

روش ماکزیمم درست‌نمایی: خصوصیات

MAXIMUM LIKELIHOOD (ML)

تخمین ML فقط برای زمانی که $N \to \infty$ خصوصیات زیر را دارد:

ناارییبی
Unbiased

می‌نیمم واریانس ممکن
Minimum Variance

همگرایی به گاوسی
Convergence to Gaussign pdf

تخمین پارامتری توابع چگالی احتمال مجهول

روش ماکزیمم درست‌نمایی: مثال

MAXIMUM LIKELIHOOD (ML)

N داده‌ی تولید شده با یک pdf گاوسی یک‌بعدی داریم

$x_1, x_2, \dots, x_N$

که میانگین آن μ و واریانس آن مجهول است.

تخمین ماکزیمم درست‌نمایی واریانس مجهول را به دست آورید.

تخمین پارامتری توابع چگالی احتمال مجهول

روش ماکزیمم درست‌نمایی: مثال

MAXIMUM LIKELIHOOD (ML)

N داده‌ی تولید شده با یک pdf گاوسی l–بعدی داریم

$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$$

که بردار میانگین آن $\boldsymbol{\mu}$ مجهول و ماتریس کوواریانس آن $\boldsymbol{\Sigma}$ معلوم است.

تخمین ماکزیمم درست‌نمایی بردار میانگین مجهول را به دست آورید.

تخمین پارامتری توابع چگالی احتمال مجهول

روش ماکزیمم درست‌نمایی: توزیع گاوسی

MAXIMUM LIKELIHOOD (ML)

Suppose that $p(\mathbf{x}|\boldsymbol{\theta}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- When $\boldsymbol{\Sigma}$ is known but $\boldsymbol{\mu}$ is unknown:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

- When both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

تخمین پارامتری توابع چگالی احتمال مجهول

روش ماکزیمم درست‌نمایی: توزیع برنولی

MAXIMUM LIKELIHOOD (ML)

- Suppose that $P(x|\theta) = \text{Bernoulli}(\theta) = \theta^x(1-\theta)^{1-x}$ where $x = 0, 1$ and $0 \leq \theta \leq 1$.
- The MLE of θ can be computed as

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

تخمین پارامتری توابع چگالی احتمال مجهول

بایاس تخمین‌گر

BIAS OF ESTIMATOR

- *Bias* of an estimator $\hat{\boldsymbol{\theta}}$ is the difference between the expected value of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$.
- The MLE of $\boldsymbol{\mu}$ is an unbiased estimator for $\boldsymbol{\mu}$ because $E[\hat{\boldsymbol{\mu}}] = \boldsymbol{\mu}$.
- The MLE of $\boldsymbol{\Sigma}$ is not an unbiased estimator for $\boldsymbol{\Sigma}$ because $E[\hat{\boldsymbol{\Sigma}}] = \frac{n-1}{n}\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}$.
- The *sample covariance*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

is an unbiased estimator for $\boldsymbol{\Sigma}$.

❖ Maximum Likelihood (ML)

- Let $\underline{x}_1, \underline{x}_2, ..., \underline{x}_N$ known and independent
- Let $p(\underline{x})$ known within an unknown vector parameter $\underline{\theta}$: $p(\underline{x}) \equiv p(\underline{x}; \underline{\theta})$
- $X = \{\underline{x}_1, \underline{x}_2, ..., \underline{x}_N\}$
- $p(X; \underline{\theta}) \equiv p(\underline{x}_1, \underline{x}_2, ..., \underline{x}_N; \underline{\theta})$

$$= \prod_{k=1}^{N} p(\underline{x}_k; \underline{\theta})$$

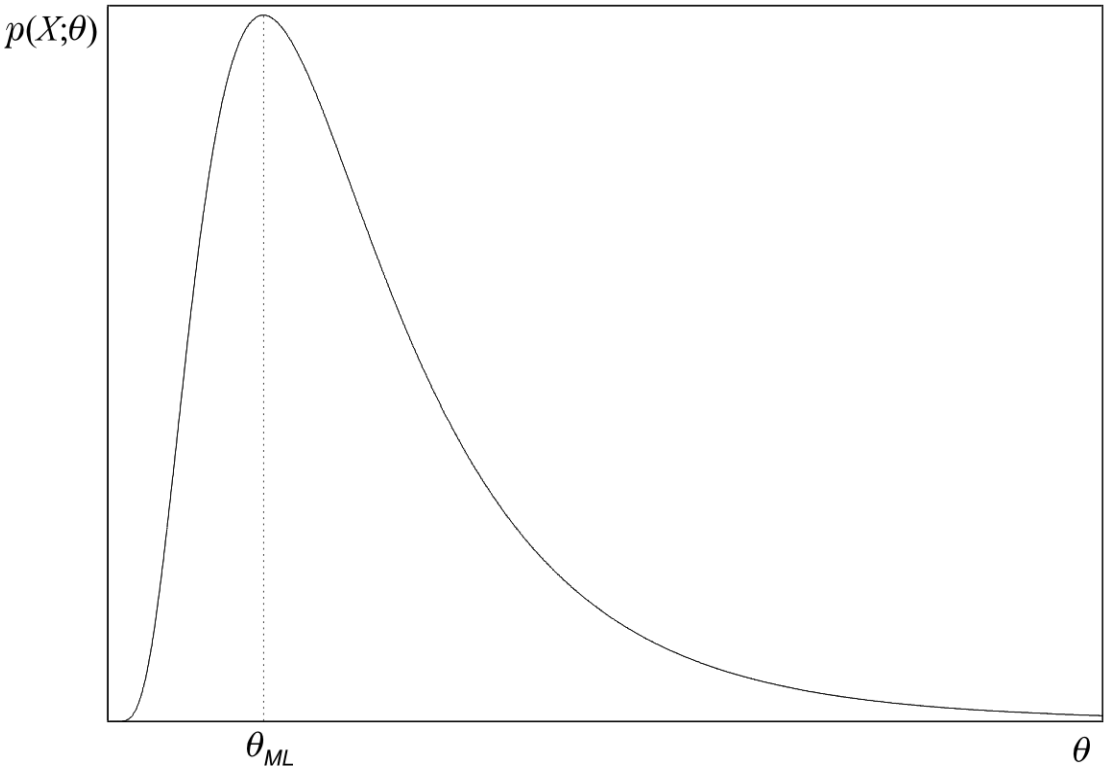
which is known as the Likelihood of $\underline{\theta}$ w.r. to X

The method :

- $$\hat{\underline{\theta}}_{\mathrm{ML}} : \arg\max_{\underline{\theta}} \prod_{k=1}^{N} p(\underline{x}_k ; \underline{\theta})$$

- $$L(\underline{\theta}) \equiv \ln p(X ; \underline{\theta}) = \sum_{k=1}^{N} \ln p(\underline{x}_k ; \underline{\theta})$$

- $$\hat{\underline{\theta}}_{ML} : \frac{\partial L(\underline{\theta})}{\partial(\underline{\theta})} = \sum_{k=1}^{N} \frac{1}{p(\underline{x}_k ; \underline{\theta})} \frac{\partial p(\underline{x}_k ; \underline{\theta})}{\partial(\underline{\theta})} = \underline{0}$$


$p(X;\theta)$
θ_{ML}
θ

If, indeed, there is a $\underline{\theta}_0$ such that

$p(\underline{x}) = p(\underline{x}; \underline{\theta}_0)$, then

$$\lim_{N\to\infty} E[\underline{\theta}_{ML}] = \underline{\theta}_0$$

$$\lim_{N\to\infty} E\left\|\underline{\hat{\theta}}_{ML} - \underline{\theta}_0\right\|^2 = 0$$

Asymptotically unbiased and consistent

❖ Example: $p(\underline{x}): N(\underline{\mu}, \Sigma): \underline{\mu}$ unknown, $\underline{x}_1, \underline{x}_2, ..., \underline{x}_N \quad p(\underline{x}_k) \equiv p(\underline{x}_k; \underline{\mu})$

$$L(\underline{\mu}) = \ln \prod_{k=1}^{N} p(\underline{x}_k; \underline{\mu}) = C - \frac{1}{2} \sum_{k=1}^{N} (\underline{x}_k - \underline{\mu})^T \Sigma^{-1} (\underline{x}_k - \underline{\mu})$$

$$p(\underline{x}_k; \underline{\mu}) = \frac{1}{(2\pi)^{\frac{l}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\underline{x}_k - \underline{\mu})^T \Sigma^{-1} (\underline{x}_k - \underline{\mu})\right)$$

$$\frac{\partial L(\underline{\mu})}{\partial (\underline{\mu})} \equiv \begin{bmatrix} \frac{\partial L}{\partial \mu_1} \\ . \\ . \\ . \\ \frac{\partial L}{\partial \mu_l} \end{bmatrix} = \sum_{k=1}^{N} \Sigma^{-1} (\underline{x}_k - \underline{\mu}) = \underline{0} \Rightarrow \underline{\mu}_{ML} = \frac{1}{N} \sum_{k=1}^{N} \underline{x}_k$$

Remember: if $A = A^T \Rightarrow \dfrac{\partial (\underline{\alpha}^T A \underline{\alpha})}{\partial \underline{\alpha}} = 2A\underline{\alpha}$

طبقه‌بندی مبتنی بر نظریه‌ی تصمیم بیز:
روش‌های تخمین پارامتری
تابع چگالی احتمال مجهول

۲

روش تخمین ماکزیمم احتمال پسین

تخمین پارامتری توابع چگالی احتمال مجهول

روش ماکزیمم احتمال پسین

MAXIMUM A POSTERIORI PROBABILITY (MAP)

در روش MAP، به پارامتر به‌عنوان یک متغیر تصادفی نگاه می‌کنیم که دارای یک توزیع پیشین معلوم است.
مشاهده‌ی نمونه‌های جدید، این چگالی پیشین را به یک چگالی پسین تبدیل می‌کند.

❖ Maximum Aposteriori Probability Estimation (MAP)

- In ML method, $\underline{\theta}$ was considered as a parameter
- Here we shall look at $\underline{\theta}$ as a random vector described by a pdf $p(\underline{\theta})$, assumed to be known
- Given

$$X = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$$

Compute the maximum of

$$p(\underline{\theta}|X)$$

- From Bayes theorem $p(\underline{\theta})p(X|\underline{\theta}) = p(X)p(\underline{\theta}|X)$ or

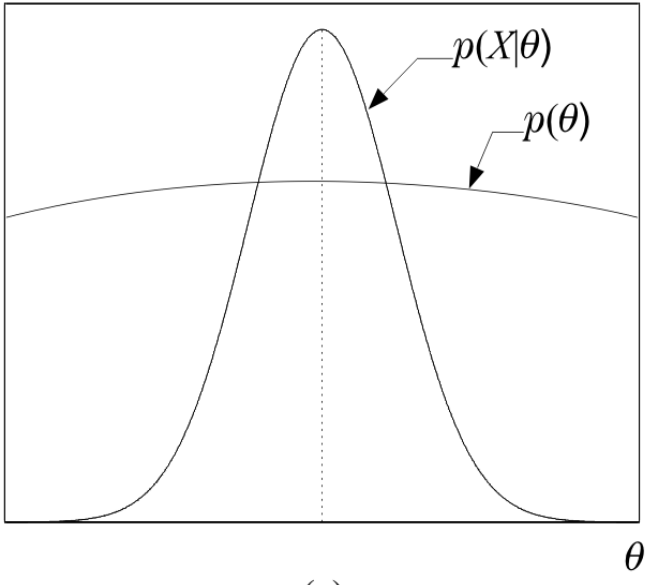
$$p(\underline{\theta}\,|\,X) = \frac{p(\underline{\theta})p(X|\underline{\theta})}{p(X)}$$

➢ The method:

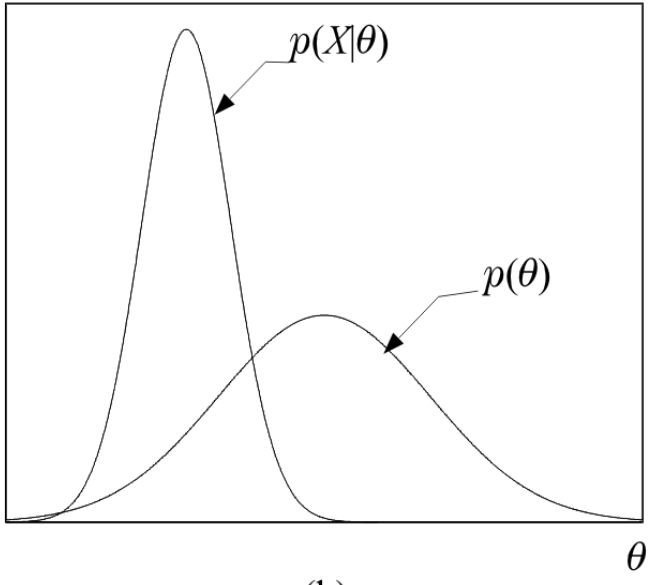
$$\underline{\hat{\theta}}_{MAP} = \arg\max_{\underline{\theta}} p(\underline{\theta}|X) \text{ or}$$

$$\underline{\hat{\theta}}_{MAP} : \frac{\partial}{\partial \underline{\theta}} (P(\underline{\theta}) p(X|\underline{\theta}))$$

If $p(\underline{\theta})$ is uniform or broad enough $\underline{\hat{\theta}}_{MAP} \cong \underline{\theta}_{ML}$



(a)



(b)

ML and MAP estimates of will be approximately the same in (a) and different in (b).

تخمین پارامتری توابع چگالی احتمال مجهول

روش ماکزیمم احتمال پسین: مثال

MAXIMUM A POSTERIORI PROBABILITY (MAP)

N داده‌ی تولید شده با یک pdf گاوسی l–بعدی داریم

$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$$

که بردار میانگین آن $\boldsymbol{\mu}$ مجهول و ماتریس کوواریانس آن $\boldsymbol{\Sigma}$ معلوم است.

اگر توزیع پیشین بردار میانگین مجهول به‌صورت زیر باشد،

$$P(\boldsymbol{\mu}) = \frac{1}{(2\pi)^{l/2}\sigma_\mu^l} \exp\left(-\frac{1}{2}\frac{||\boldsymbol{\mu}-\boldsymbol{\mu}_0||^2}{\sigma_\mu^2}\right)$$

تخمین ماکزیمم احتمال پسین بردار میانگین مجهول را به دست آورید.

بازشناسی الگو

طبقه‌بندی مبتنی بر نظریه‌ی تصمیم بیز:
روش‌های تخمین پارامتری
تابع چگالی احتمال مجهول

۳

روش تخمین استنتاج بیزی

تخمین پارامتری توابع چگالی احتمال مجهول

روش استنتاج بیزی

BAYESIAN INFERENCE

روش ML و MAP، یک تخمین خاص از بردار پارامتر مجهول $\boldsymbol{\theta}$ ارائه می‌کردند،
اما در **استنتاج بیزی**، مسیر دیگری دنبال می‌شود:
با داشتن مجموعه‌ی $\mathcal{D}$ از بردارهای آموزشی و اطلاعات پیشین $p(\boldsymbol{\theta})$ می‌خواهیم $p(\mathbf{x}|\mathcal{D})$ را محاسبه کنیم.

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

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

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})$$

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

توزیع شرطی $p(\boldsymbol{\theta}|\mathcal{D})$ نیز به‌عنوان تخمین pdf پسین معلوم است:
(زیرا دانایی به‌روزشده در مورد خواص آماری $\boldsymbol{\theta}$ پس از مشاهده‌ی مجموعه داده $\mathcal{D}$ است.)

❖ Bayesian Inference

- ML, MAP $\Rightarrow$ a single estimate for $\underline{\theta}$.

  Here a different root is followed.

  Given :

  $X = \{\underline{x}_1, \ldots, \underline{x}_N\}$, $p(\underline{x}|\underline{\theta})$ and $p(\underline{\theta})$

  The goal :

  estimate $p(\underline{x}|X)$

  How??

$$p(\underline{x}|X) = \int p(\underline{x}|\underline{\theta})\, p(\underline{\theta}|X)\, d\underline{\theta}$$

$$p(\underline{\theta}|X) = \frac{p(X|\underline{\theta})\, p(\underline{\theta})}{p(X)} = \frac{p(X|\underline{\theta})\, p(\underline{\theta})}{\int p(X|\underline{\theta})\, p(\underline{\theta})\, d\underline{\theta}}$$

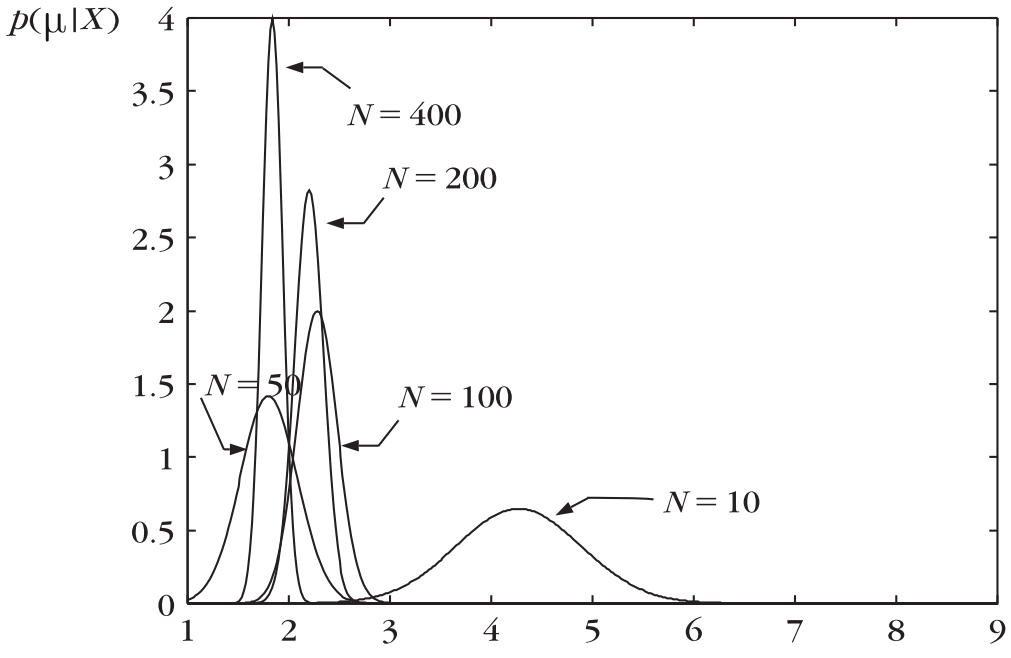
$$p(X|\underline{\theta}) = \prod_{k=1}^{N} p(\underline{x}_k|\underline{\theta})$$

A bit more insight via an example

- *Let* $p(x|\mu) \rightarrow N(\mu, \sigma^2)$
- $p(\mu) \rightarrow N(\mu_0, \sigma_0^2)$
- It turns out that: $p(\mu|X) \rightarrow N(\mu_N, \sigma_N^2)$

$$\mu_N = \frac{N\sigma_0^2 \bar{x} + \sigma^2 \mu_0}{N\sigma_0^2 + \sigma^2}, \qquad \sigma_N^2 = \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2}, \qquad \bar{x} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

➢ The above is a sequence of Gaussians as $N \to \infty$



❖ Example:

$$p(\underline{x}):\ N(\underline{\mu},\Sigma),\ \ \underline{\mu}\ \ \text{unknown},\ \ \ X=\{\underline{x}_1,\ldots,\underline{x}_N\}$$

$$p(\underline{\mu})=\frac{1}{(2\pi)^{\frac{l}{2}}\sigma_\mu^l}\exp\left(-\frac{\left\|\underline{\mu}-\underline{\mu}_0\right\|^2}{2\sigma_\mu^2}\right)$$

$$\underline{\hat{\theta}}_{MAP}:\ \frac{\partial}{\partial\underline{\mu}}\ln\left(\prod_{k=1}^{N}p(\underline{x}_k|\underline{\mu})p(\underline{\mu})\right)=\underline{0}\ \ \text{or}\ \ \sum_{k=1}^{N}\frac{1}{\sigma^2}(\underline{x}_k-\underline{\mu})-\frac{1}{\sigma_\mu^2}(\underline{\hat{\mu}}-\underline{\mu}_0)=\underline{0}\Rightarrow$$

$$\underline{\hat{\mu}}_{MAP}=\frac{\underline{\mu}_0+\frac{\sigma_\mu^2}{\sigma^2}\sum_{k=1}^{N}\underline{x}_k}{1+\frac{\sigma_\mu^2}{\sigma^2}N}\ \ \text{for}\ \ \frac{\sigma_\mu^2}{\sigma^2}>>1,\ \text{or for}\ \ N\to\infty$$

$$\underline{\hat{\mu}}_{\text{MAP}}\cong\underline{\hat{\mu}}_{ML}=\frac{1}{N}\sum_{k=1}^{N}\underline{x}_k$$

بازشناسی الگو

طبقه‌بندی مبتنی بر نظریه‌ی تصمیم بیز:
روش‌های تخمین پارامتری
تابع چگالی احتمال مجهول

۴

روش تخمین ماکزیمم آنتروپی

تخمین پارامتری توابع چگالی احتمال مجهول

روش ماکزیمم آنتروپی

MAXIMUM ENTROPY (ME)

روش ME تابع چگالی احتمالی را می‌یابد که آنتروپی را ماکزیمم می‌کند.
(متناظر با توزیعی که بالاترین میزان تصادفی بودن ممکن را نشان می‌دهد)

❖ Maximum Entropy

- Entropy

$$H = -\int p(\underline{x}) \ln p(\underline{x}) d\underline{x}$$

- $\hat{p}(x)$: maximum H

  subject to the available constraints

➢ Example: x is nonzero in the interval $[x_1, x_2]$ and zero otherwise. Compute the ME pdf

- The constraint: $\int_{x_1}^{x_2} p(x)dx = 1$

- Lagrange Multipliers $H_L = H + \lambda(\int_{x_1}^{x_2} p(x)dx - 1)$

- $$\hat{p}(x) = \exp(\lambda - 1)$$

- $$\hat{p}(x) = \begin{cases} \dfrac{1}{x_2 - x_1} & x_1 \le x \le x_2 \\ 0 & \text{otherwise} \end{cases}$$

طبقه‌بندی مبتنی بر نظریه‌ی تصمیم بیز:
روش‌های تخمین پارامتری
تابع چگالی احتمال مجهول

۵

روش مدل‌های مخلوط

تخمین پارامتری توابع چگالی احتمال مجهول

روش مدل‌های چگالی مخلوط

MIXTURE DENSITY MODELS

- A mixture model is a linear combination of m densities

$$p(\mathbf{x}|\mathbf{\Theta}) = \sum_{j=1}^{m} \alpha_j p_j(\mathbf{x}|\boldsymbol{\theta}_j)$$

where $\mathbf{\Theta} = (\alpha_1, \ldots, \alpha_m, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_m})$ such that $\alpha_j \geq 0$ and $\sum_{j=1}^{m} \alpha_j = 1$.

- $\alpha_1, \ldots, \alpha_m$ are called the mixing parameters.
- $p_j(\mathbf{x}|\boldsymbol{\theta}_j),\ j = 1, \ldots, m$ are called the component densities.

فرض می‌شود یک توزیع pdf مجهول، از طریق ترکیب خطی m تابع چگالی ساخته می‌شود.

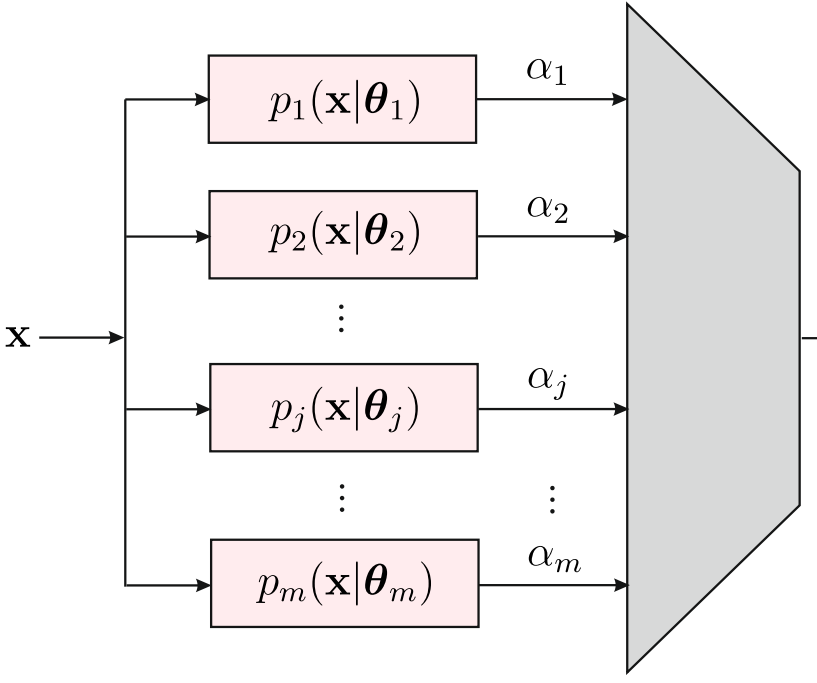
* این مدل به‌طور ضمنی فرض می‌کند که
هر نقطه‌ی x می‌تواند از هر یک از m توزیع با احتمال α_j بیرون کشیده شود.

* این مدل می‌تواند هر تابع چگالی پیوسته را با تعداد کافی مؤلفه و با پارامترهای مناسب تقریب بزند.

تخمین پارامتری توابع چگالی احتمال مجهول

روش مدل‌های چگالی مخلوط

MIXTURE DENSITY MODELS



$$p(\mathbf{x}|\boldsymbol{\Theta}) = \sum_{j=1}^{m} \alpha_j p_j(\mathbf{x}|\boldsymbol{\theta}_j)$$

$$\sum_{j=1}^{m} \alpha_j = 1$$

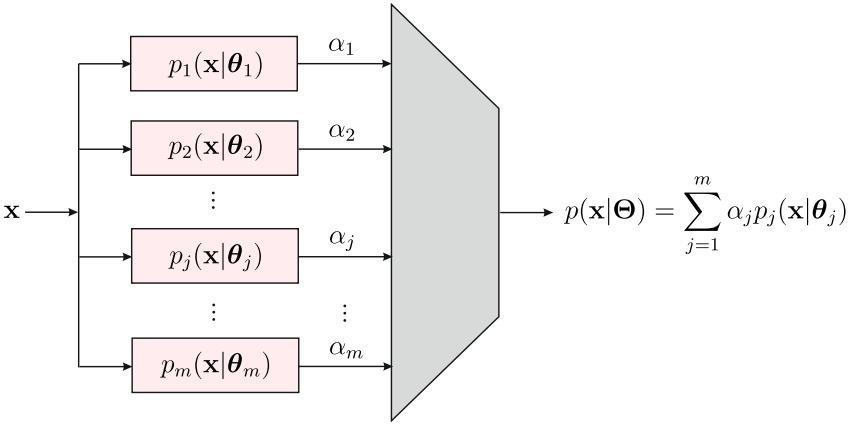
$$\forall j \quad \alpha_j = 1$$

$$\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_j, \ldots \boldsymbol{\theta}_m, \alpha_1, \alpha_2, \ldots, \alpha_j \ldots, \alpha_m]$$

تخمین پارامتری توابع چگالی احتمال مجهول

روش مدل‌های چگالی مخلوط: تخمین پارامترها

MIXTURE DENSITY MODELS



چون مشخص نیست هر داده‌ی آموزشی $\mathbf{x}$ از کدام مؤلفه‌ی مخلوط می‌آید،
استفاده از ML برای تخمین پارامترها به یک مسئله‌ی بهینه‌سازی غیرخطی پیچیده منجر می‌شود.
(اگر مشخص بود، m مسئله‌ی ML جداگانه حل می‌شد.)

⇓

با یک مسئله با مجموعه داده‌ی آموزشی ناکامل سر و کار داریم (اطلاعات برچسب ناقص است)

⇓

از روش ماکزیمم‌سازی امید (EM) استفاده می‌کنیم.

❖ Mixture Models

➢ $$p(\underline{x}) = \sum_{j=1}^{J} p(\underline{x}|j) P_j \qquad \sum_{j=1}^{M} P_j = 1, \int_{\underline{x}} p(\underline{x}|j) d\underline{x} = 1$$

➢ Assume parametric modeling, i.e., $p(\underline{x}|j;\underline{\theta})$

➢ The goal is to estimate $\underline{\theta}$ and $P_1, P_2, ..., P_j$
given a set $X = \{\underline{x}_1, \underline{x}_2, ..., \underline{x}_N\}$

➢ Why not ML? As before? $\max_{\underline{\theta}, P_i, ..., P_j} \prod_{k=1}^{N} P(\underline{x}_k; \underline{\theta}, P_i, ..., P_j)$

- This is a nonlinear problem due to the missing label information.
  This is a typical problem with an **incomplete** data set.

Solution??

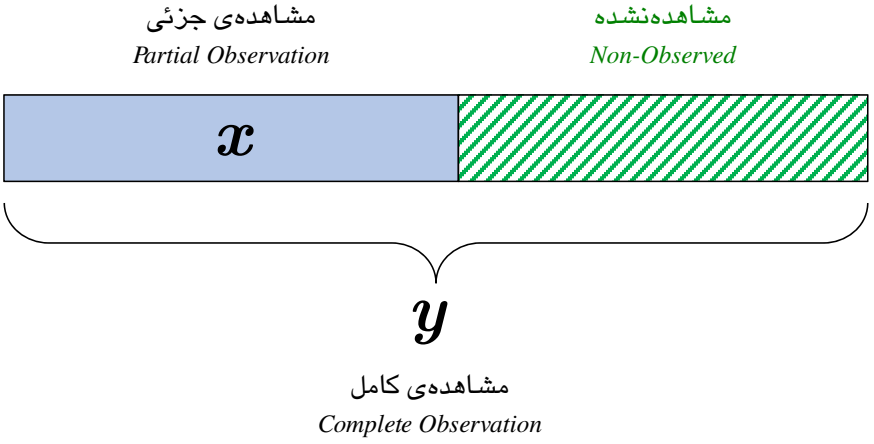
- The **Expectation-Maximization (EM)** algorithm.

تخمین پارامتری توابع چگالی احتمال مجهول

ماکزیمم‌سازی امید

EXPECTATION MAXIMIZATION (EM)

EM روشی کارآمد برای حل عددی مسئله‌ی ML است، وقتی که مشاهدات جزئی در دست است.



الگوریتم EM مقدار امید تابع درست‌نمایی را به‌شرط نمونه‌های مشاهده شده و تخمین فعلی θ ماکزیمم می‌کند:

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} E\{\sum_{k} \log p_y(\mathbf{y}_k|\boldsymbol{\theta})\}$$

که در آن امید ریاضی بر روی بخش تصادفی $\mathbf{y}$ (قسمت مشاهده نشده) محاسبه می‌شود.

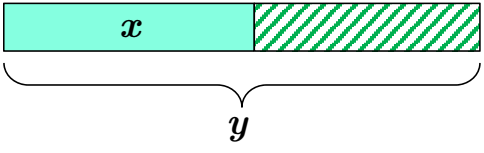
➢ The Expectation-Maximization (EM) algorithm

- General formulation
  - $\underline{y}$ the complete data set $\underline{y} \in Y \subseteq \Re^m$, with $p_{\underline{y}}(\underline{y};\underline{\theta})$, which are not observed directly.

We observe

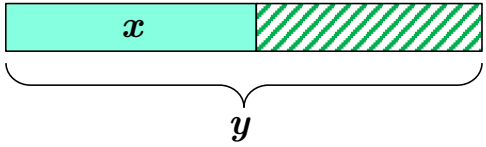
$$\underline{x} = g(\underline{y}) \in X_{ob} \subseteq \Re^l, l < m \text{ with } P_x(\underline{x};\underline{\theta}),$$

a many to one transformation



- Let $Y(\underline{x}) \subseteq Y$ all $\underline{y}'s \rightarrow$ to a specific $\underline{x}$

$$p_{\underline{x}}(\underline{x};\underline{\theta}) = \int_{Y(\underline{x})} p_{\underline{y}}(\underline{y};\underline{\theta}) d\underline{y}$$



- What we need is to compute

$$\hat{\underline{\theta}}_{ML} : \sum_k \frac{\partial \ln(p_{\underline{y}}(\underline{y}_k;\underline{\theta}))}{\partial \underline{\theta}} = \underline{0}$$

- But $\underline{y}_k$'s are not observed.
  Here comes the EM.
  Maximize the expectation of the log-likelihood
  conditioned on the observed samples
  and the **current** iteration estimate of $\underline{\theta}$.

➢ The algorithm:

- E-step:

$$Q(\underline{\theta};\underline{\theta}(t)) = E[\sum_{k} \ln(p_{\underline{y}}(\underline{y}_k;\underline{\theta}|X;\underline{\theta}(t))]$$

- M-step:

$$\underline{\theta}(t+1) = \arg\max_{\underline{\theta}} Q(\underline{\theta};\underline{\theta}(t))$$

$$\underline{\theta}(t+1): \frac{\partial Q(\underline{\theta};\underline{\theta}(t))}{\partial \underline{\theta}} = \underline{0}$$

تخمین پارامتری توابع چگالی احتمال مجهول

ماکزیمم‌سازی امید: الگوریتم

EXPECTATION MAXIMIZATION (EM)

Algorithm (Expectation-Maximization)

1 **begin initialize** $\boldsymbol{\theta}^0, T, i = 0$
2 **do** $i \leftarrow i + 1$
3 **E step** : compute $Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^i)$
5 **M step** : $\boldsymbol{\theta}^{i+1} \leftarrow \arg\max_{\theta} Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^i)$
6 **until** $Q(\boldsymbol{\theta}^{i+1};\ \boldsymbol{\theta}^i) - Q(\boldsymbol{\theta}^i;\ \boldsymbol{\theta}^{i-1}) \leq T$
7 **return** $\hat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^{i+1}$
8 **end**

تخمین پارامتری توابع چگالی احتمال مجهول

ماکزیمم‌سازی امید: نکته‌ها

EXPECTATION MAXIMIZATION (EM)

در الگوریتم EM تخمین‌های پی‌درپی $\boldsymbol{\theta}(t)$ هرگز مقدار تابع درست‌نمایی را کاهش نمی‌دهد.

تابع درست‌نمایی صعودی می‌ماند تا به یک ماکزیمم (محلی/سراسری) برسد و EM همگرا شود.

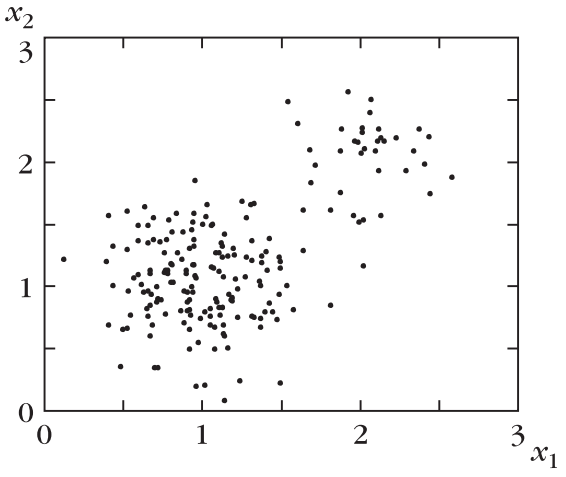
همگرایی الگوریتم EM کندتر از همگرایی الگوریتم‌های جستجو مانند نیوتن است.

اما همگرایی آن ملایم است و نسبت به ناپایداری آسیب‌پذیری ندارد (قوام بالاتر).

تخمین پارامتری توابع چگالی احتمال مجهول

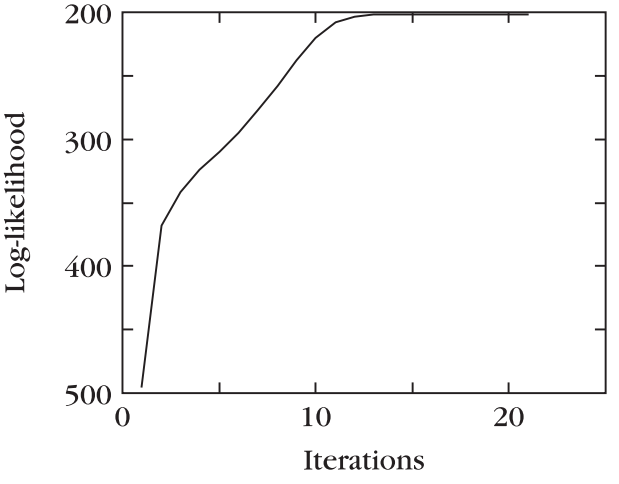
ماکزیمم‌سازی امید: مثال

EXPECTATION MAXIMIZATION (EM)



(a)

مجموعه‌ی داده‌های آموزشی (دوبعدی)



(b)

تابع log-likelihood محاسبه شده در تکرارهای پی‌درپی الگوریتم EM (همیشه غیرنزولی)

تخمین پارامتری توابع چگالی احتمال مجهول

مدل‌های مخلوط گاوسی

GAUSSIAN MIXTURE MODELS (GMM)

مدل مخلوط گاوسی
Gaussian Mixture Model

یک مدل چگالی مخلوط که مؤلفه‌های آن گاوسی باشد.

یک راه برآورد تعداد مؤلفه‌ها، تعداد قله‌های موجود در pdf است.

➢ Application to the mixture modeling problem

- Complete data

$$(\underline{x}_k, j_k), k = 1,2,\ldots,N$$

- Observed data

$$\underline{x}_k, k = 1,2,\ldots,N$$

$$p(\underline{x}_k, j_k; \underline{\theta}) = p(\underline{x}_k | j_k; \underline{\theta}) P_{jk}$$

- Assuming mutual independence, the log-likelihood is

$$L(\underline{\theta}) = \sum_{k=1}^{N} \ln(p(\underline{x}_k | j_k; \underline{\theta}) P_{jk})$$

➢ Unknown parameters

$$\underline{\Theta}^T = [\underline{\theta}^T, \underline{P}^T]^T, \quad \underline{P} = [P_1, P_2, \ldots, P_J]^T$$

➢ E-step

$$Q(\underline{\Theta}; \underline{\Theta}(t)) = E[\sum_{k=1}^{N} \ln(p(\underline{x}_k | j_k; \underline{\theta}) P_{j_k})] = \sum_{k=1}^{N} E[\quad] =$$

$$\sum_{k=1}^{N} \sum_{j_k=1}^{J} P(j_k | \underline{x}_k; \underline{\Theta}(t)) \ln(p(\underline{x}_k | j_k; \underline{\theta}) P_{j_k})$$

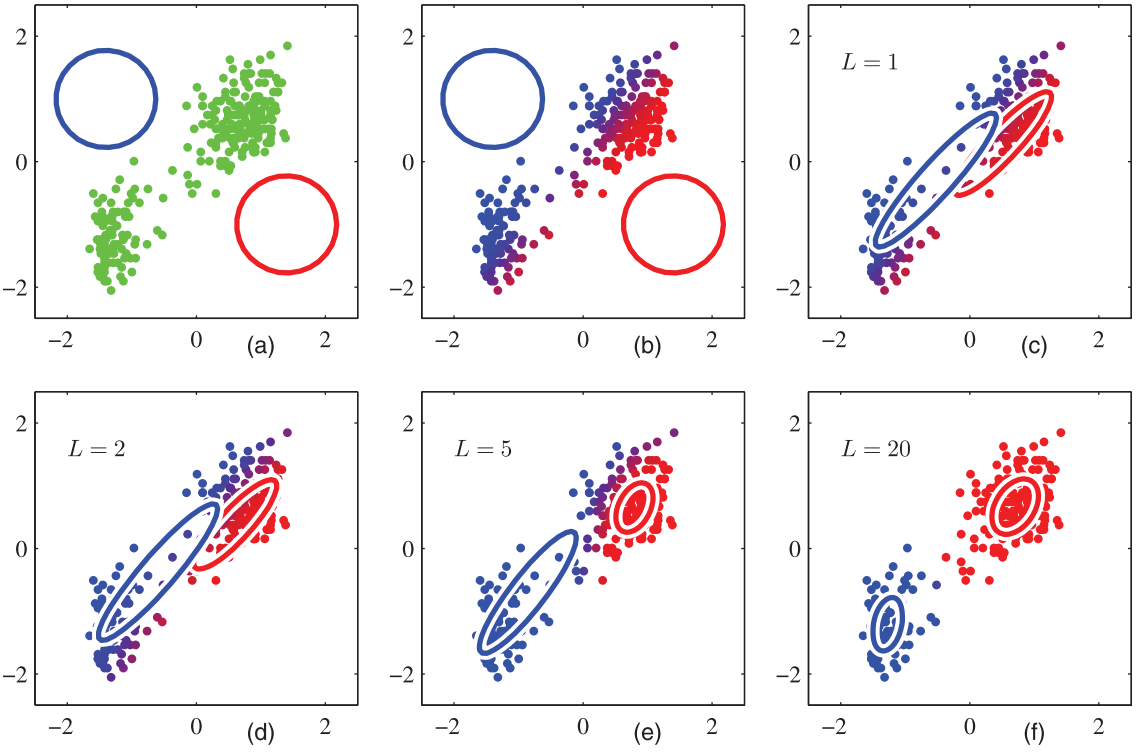
➢ M-step $\dfrac{\partial Q}{\partial \underline{\theta}} = \underline{0} \qquad \dfrac{\partial Q}{\partial P_{jk}} = 0, \qquad j_k = 1, 2, \ldots, J$

$$P(j | \underline{x}_k; \Theta(t)) = \frac{p(\underline{x}_k | j; \underline{\Theta}(t)) P_j}{P(\underline{x}_k; \underline{\Theta}(t))} \qquad p(\underline{x}_k; \underline{\Theta}(t)) = \sum_{j=1}^{J} p(\underline{x}_k | j; \underline{\theta}(t)) P_j$$

تخمین پارامتری توابع چگالی احتمال مجهول

ماکزیمم‌سازی امید: مثال

EXPECTATION MAXIMIZATION (EM)



مدل‌های چگالی مخلوط

کاربرد در خوشه‌بندی

MIXTURE DENSITY MODELS: APPLICATION IN CLUSTERING

می‌توان از مدل‌های مخلوط در کاربرد **خوشه‌بندی** استفاده کرد.

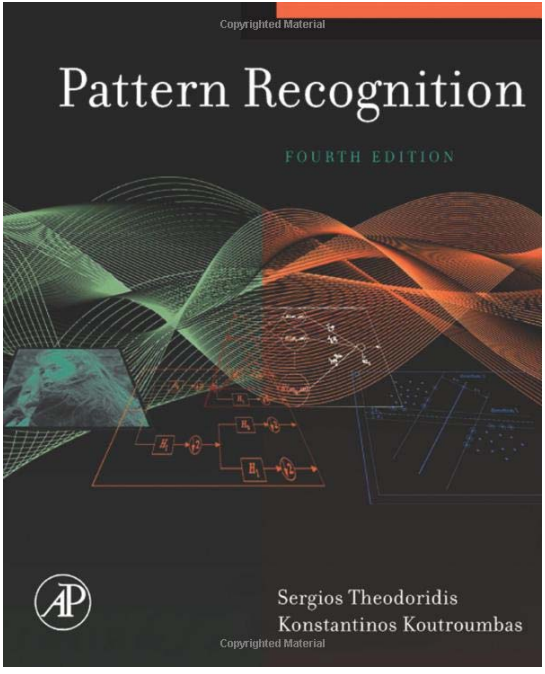
هر مؤلفه را می‌توان یک خوشه در نظر گرفت:
وقتی مشخص شود که احتمال تولید هر $\mathbf{x}$ توسط کدام مؤلفه (j) بیشتر است،
عملاً خوشه‌ی آن $\mathbf{x}$ مشخص می‌شود.

طبقه‌بندی مبتنی بر نظریه‌ی تصمیم بیز:
روش‌های تخمین پارامتری
تابع چگالی احتمال مجهول

۶

منابع

منبع اصلی



S. Theodoridis, K. Koutroumbas,
Pattern Recognition,
Fourth Edition, Academic Press, 2009.

Chapter 2

CHAPTER 2

Classifiers Based on Bayes Decision Theory

2.1 INTRODUCTION

This is the first chapter, out of three, dealing with the design of the classifier in a pattern recognition system. The approach to be followed builds upon probabilistic arguments stemming from the statistical nature of the generated features. As has already been pointed out in the introductory chapter, this is due to the statistical variation of the patterns as well as to the noise in the measuring sensors. Adopting this reasoning as our kickoff point, we will design classifiers that classify an unknown pattern in the most probable of the classes. Thus, our task now becomes that of defining what "most probable" means.

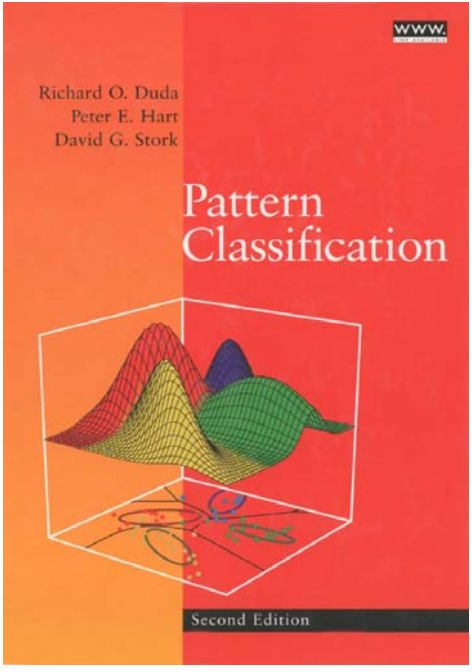
Given a classification task of M classes, $\omega_1, \omega_2, \ldots, \omega_M$, and an unknown pattern, which is represented by a feature vector $\boldsymbol{x}$, we form the M conditional probabilities $P(\omega_i|\boldsymbol{x}), i = 1, 2, \ldots, M$. Sometimes, these are also referred to as *a posteriori probabilities*. In words, each of them represents the probability that the unknown pattern belongs to the respective class ω_i, given that the corresponding feature vector takes the value $\boldsymbol{x}$. Who could then argue that these conditional probabilities are not sensible choices to quantify the term *most probable*? Indeed, the classifiers to be considered in this chapter compute either the maximum of these M values or, equivalently, the maximum of an appropriately defined function of them. The unknown pattern is then assigned to the class corresponding to this maximum.

The first task we are faced with is the computation of the conditional probabilities. The Bayes rule will once more prove its usefulness! A major effort in this chapter will be devoted to techniques for estimating probability density functions (pdf), based on the available experimental evidence, that is, the feature vectors corresponding to the patterns of the training set.

2.2 BAYES DECISION THEORY

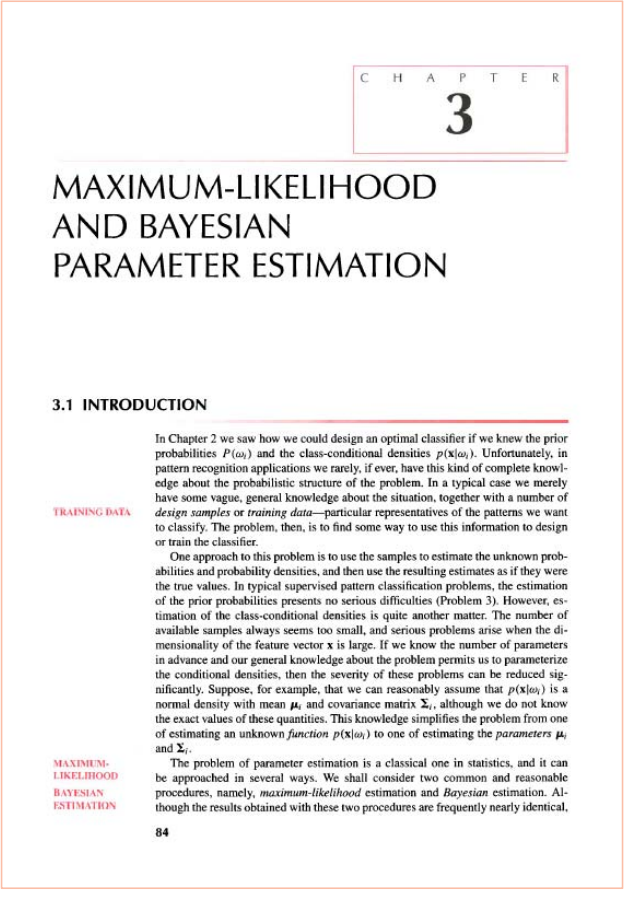
We will initially focus on the two-class case. Let ω_1, ω_2 be the two classes in which our patterns belong. In the sequel, we assume that the *a priori probabilities*

منبع کمکی



R.O. Duda, P.E. Hart, and D.G. Stork,
Pattern Classification,
Second Edition, John Wiley & Sons, Inc., 2001.

Chapter 3



CHAPTER 3

MAXIMUM-LIKELIHOOD AND BAYESIAN PARAMETER ESTIMATION

3.1 INTRODUCTION

TRAINING DATA

In Chapter 2 we saw how we could design an optimal classifier if we knew the prior probabilities $P(\omega_i)$ and the class-conditional densities $p(\mathbf{x}|\omega_i)$. Unfortunately, in pattern recognition applications we rarely, if ever, have this kind of complete knowledge about the probabilistic structure of the problem. In a typical case we merely have some vague, general knowledge about the situation, together with a number of *design samples* or *training data*—particular representatives of the patterns we want to classify. The problem, then, is to find some way to use this information to design or train the classifier.

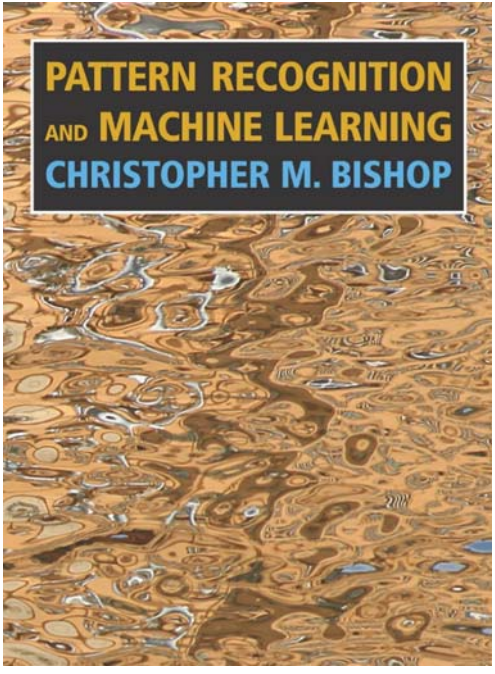
One approach to this problem is to use the samples to estimate the unknown probabilities and probability densities, and then use the resulting estimates as if they were the true values. In typical supervised pattern classification problems, the estimation of the prior probabilities presents no serious difficulties (Problem 3). However, estimation of the class-conditional densities is quite another matter. The number of available samples always seems too small, and serious problems arise when the dimensionality of the feature vector $\mathbf{x}$ is large. If we know the number of parameters in advance and our general knowledge about the problem permits us to parameterize the conditional densities, then the severity of these problems can be reduced significantly. Suppose, for example, that we can reasonably assume that $p(\mathbf{x}|\omega_i)$ is a normal density with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$, although we do not know the exact values of these quantities. This knowledge simplifies the problem from one of estimating an unknown *function* $p(\mathbf{x}|\omega_i)$ to one of estimating the *parameters* $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$.

MAXIMUM-LIKELIHOOD
BAYESIAN ESTIMATION

The problem of parameter estimation is a classical one in statistics, and it can be approached in several ways. We shall consider two common and reasonable procedures, namely, *maximum-likelihood* estimation and *Bayesian* estimation. Although the results obtained with these two procedures are frequently nearly identical,

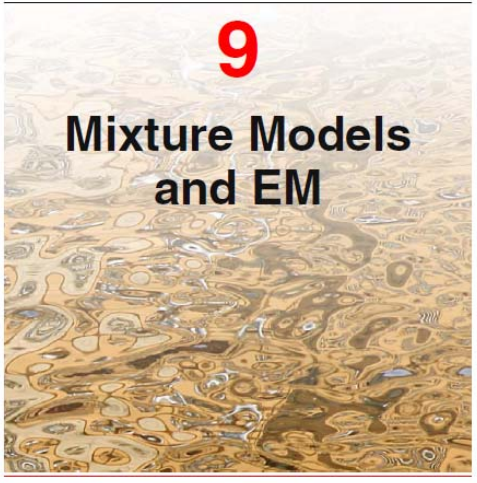
84

منبع کمکی



C.M. Bishop,
Pattern Recognition and Machine Learning,
Springer, 2006.

Chapter 9



If we define a joint distribution over observed and latent variables, the corresponding distribution of the observed variables alone is obtained by marginalization. This allows relatively complex marginal distributions over observed variables to be expressed in terms of more tractable joint distributions over the expanded space of observed and latent variables. The introduction of latent variables thereby allows complicated distributions to be formed from simpler components. In this chapter, we shall see that mixture distributions, such as the Gaussian mixture discussed in Section 2.3.9, can be interpreted in terms of discrete latent variables. Continuous latent variables will form the subject of Chapter 12.

As well as providing a framework for building more complex probability distributions, mixture models can also be used to cluster data. We therefore begin our discussion of mixture distributions by considering the problem of finding clusters in a set of data points, which we approach first using a nonprobabilistic technique called the K-means algorithm (Lloyd, 1982). Then we introduce the latent variable

Section 9.1

423