بازشناسی آماری الگو

جلسه پنجم

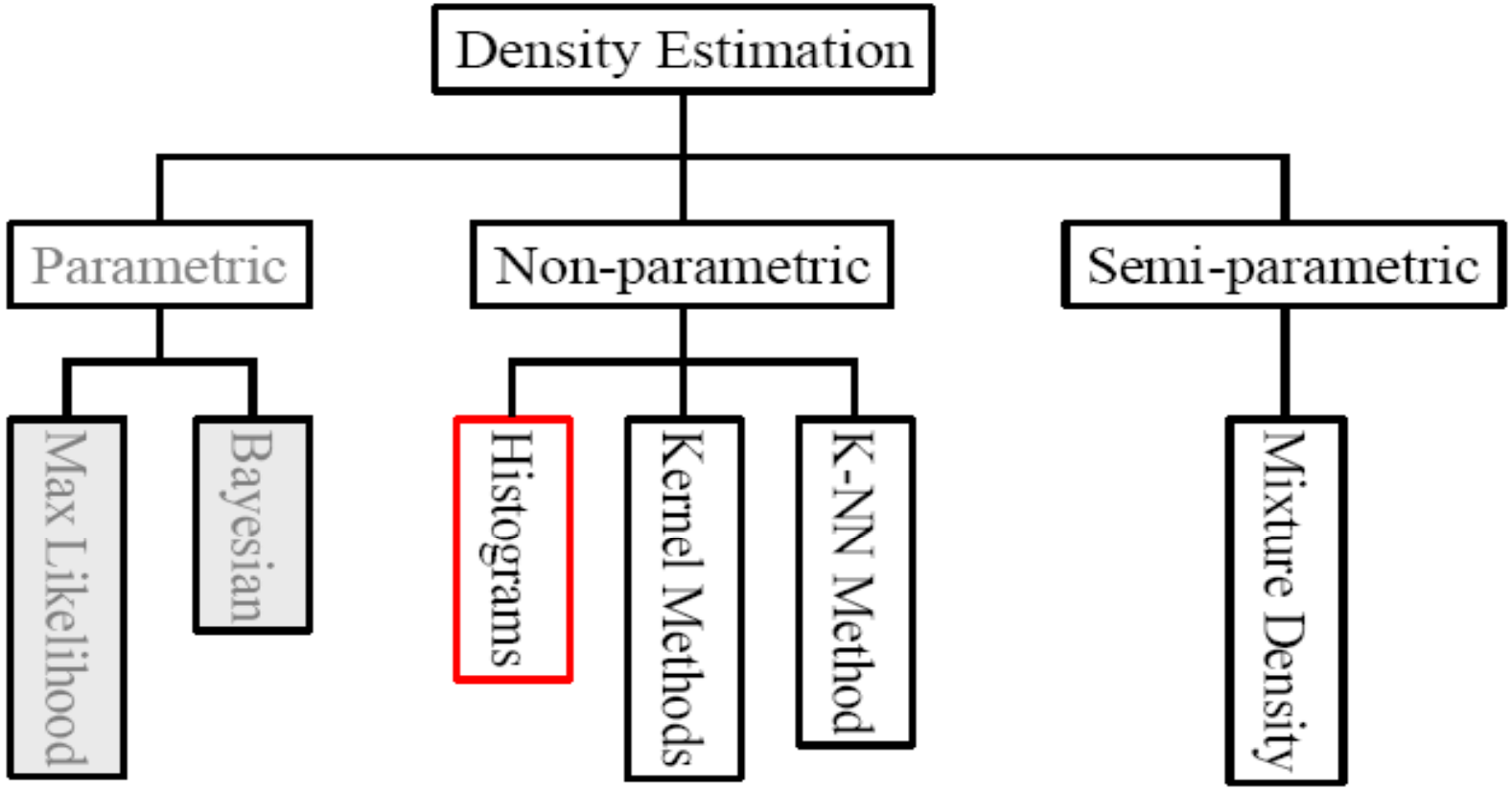
تخمین احتمال و چگالی احتمال

روش های غیر پارامتری

فهرست مطالب

- مقدمه و یادآوری
- چرا روش های غیر پارامتری
- هیستوگرام چند بعدی
- روش Parzen window
- روش KNN
- روش مدل چگالی مختلط Mixture Model Density
- خلاصه بحث

نگاهی به تقسیم بندی بحث تخمین



مقدمه و یادآوری
تخمین توابع احتمال

- دیدیم که به دلایل مختلف، در SPR ممکن است نیازمند تخمین توابع احتمال(چگالی توزیع یا احتمال جمعی) باشیم:
  - استفاده از توابع احتمال برای استفاده در فرمول های ترکیب Bayes
  - استفاده از اطلاعات توزیع برای مراحل تصمیم گیری و بعد آن
- مسئله تخمین عبارت است از یافتن تابعی که چند نقطه (نمونه) از بخش های مختلف آن در دست است.
- نخستین دسته از روش های تخمین، روش های پارامتری هستند که در آنها، فرض می شود نوع تابع معلوم است و تنها باید مقادیر پارامتر های آن محاسبه شوند.
- درروش های غیر پارامتری، پیش فرض های یاد شده (مشخص بودن نوع تابع و امکان بیان پارامتری تابع مزبور) وجود ندارد.

مقدمه و یادآوری- روش های پارامتری

- دیدیم که در روش های پارامتری، رویکرد کلی، تبدیل مسئله تخمین به یک مسئله بهینه سازی است.
- در مسئله بهینه سازی تخمین، تابع معیار بهینه سازی می تواند از مفاهیمی همچون شباهت (Likelihood) و یا احتمال پسین وقوع پارامتر در صورت وجود مجموعه نمونه ها نشات بگیرد.
- وجه دیگر بهینه سازی، یعنی روش جستجو، نوعا بااستفاده از مفهوم نقاط بینه محلی، یعنی با صفر قرار دادن تغییرات (مشتق) در نقطه بهینه بدست خواهد آمد.
- در وجه ساده تخمین پارامتری (ثابت و مجهول بودن پارامتر های بهینه- روش ML) دقت تخمین مسئله است و در وجه پیچیده آن(روش Bayesian) در بسیاری از موارد جواب به دست نمی آید.

چرا روش های غیر پارامتری؟

- روش های غیر پارامتری دشوار تر بوده و از دقت کمتری برخوردار است، پس چرا باید از آنها استفاده کنیم؟
  - این روش ها به معلومات اولیه، ملزومات و پیش فرض های کمتری نیازمندند.
    - عدم نیاز به دانستن (حدس) شکل تابع چگالی احتمال
  - برخی از توابع توزیع معروف(مانند توزیع یکنواخت) قابل بیان با تعداد محدودی از پارامتر ها نیستند.
  - مشکلات عملی در بهینه سازی (مانند وجود مینیمم های محلی) در روش های پارامتری وجود دارد.
  - عدم تطابق مدل پارامتری با مراحل دیگر پروسه (مثلا طرح کلاسیفایر) ممکن است ما را به این روش ها علاقمند سازد.

کاربرد های اصلی مسئله تخمین غیر پارامتری چگالی

- تخمین مستقیم چگالی بر اساس ایده شمارش
  - تقسیم بندی فضا و شمارش در بخش های مختلف(هیستوگرام چند بعدی)
- تخمین احتمال پسین عضویت در کلاس
  - نسبت اعضای شمارش شده یک کلاس به کل اعضا
- تخمین غیر مستقیم چگالی(بر اساس مبادرت به نشانه label و سپس محاسبه احتمال پسین وقوع کلاس ها با شمارش اعضای نشانه گذاری شده در هرکلاس)
  - معمولا در این شرایط، استفاده از اطلاعات آماری پس از نشانه گذاری مورد توجه است. (مثلا رد تصمیم هایی با دلایل ضعیف یا ترکیب چند کلاسیفایر)

تخمین چگالی بر اساس ایده هیستوگرام چند بعدی
بررسی صورت مسئله اساسی

- فرض کنیم داده های (آموزشی) به N مجموعه جدا از هم (افراز) تقسیم شده که هر یک از آنها متعلق به یکی از کلاس هاست.
- نمونه های داده شده باید به صورت بردار ها(نقطه ها) ی قرار گرفته در فضای مشخصه ها در نظر گرفته شوند.
- توزیع نمونه های هر کلاس معلوم نیست.
- ایده اساسی آن است که در ناحیه مزبور، می توان زیر فضاهایی (با توزیع نسبتا ثابت) داشت و با استفاده از شمارش در آن زیر فضا ها، توزیع ثابت مزیور قابل محاسبه است.

ایده اساسی هیستوگرام چند بعدی برای تخمین چگالی:

- احتمال اینکه یک بردار(نقطه) در ناحیه خاصی از فضای مشخصه ها مانند R قرار گیرد عبارتست از:

$$P = \int_{\Re} p(x')dx' \qquad (1)$$

- تابع P عبارت است از مقدار متوسط چگالی احتمال p در ناحیه مزبور.
  - با این فرض که بازه را بتوان آنقدر کوچک فرض کرد که چگالی احتمال در آن ناحیه چندان تغییر نکند،
  - ضمنا تعداد n نمونه دیده شده در آن ناحیه نیز قابل شمارش باشد.
- احتمال دیده شدن(قرار گرفتن) k نمونه از N نمونه اولیه در یک ناحیه با فرض معلوم بودن احتمال دیده شدن هر نمونه (با فرض مستقل بودن احتمال وقوع نمونه ها) را می توان با توزیع دوجمله ای به دست آورد.

توزیع دو جمله ای و محاسبه احتمال

By definintion:

$$P(x \in R) = P = \int_R p(x')dx'$$

If we have N i.i.d. points drawn form $p(x)$:

$$P(|\, x \in R \,| = k) = \underbrace{\frac{N!}{k!(N-k)!}}_{\text{Num. of unique splits K vs. (N-K)}} \underbrace{P^k}_{\text{Prob that } k \text{ of particular } x\text{-es are in R}} \underbrace{(1-P)^{N-k}}_{\text{Prob that the rest are not}} = B(N, P)$$

$B(N, P)$ is a *binomial* distribution of k

پارامتر های توزیع دو جمله ای

$$P_k = \binom{n}{k} P^k (1-P)^{n-k} \qquad (2)$$

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$$

مقدار متوسط یا امید ریاضی عبارت خواهد بود از

$$E(k) = nP \qquad (3)$$

- می توان نشان دادکه برای توزیع دو جمله ای، پارامتر های توزیع عبارتند از:

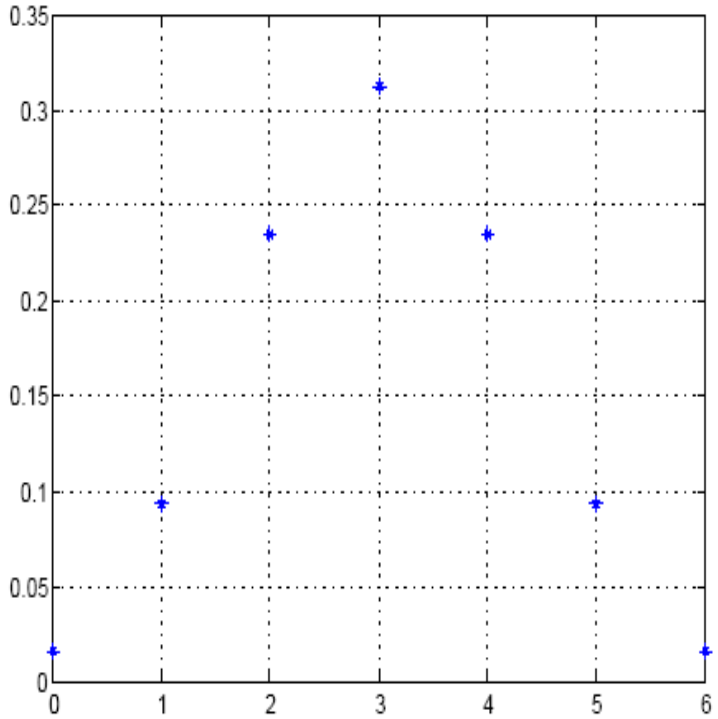
$$\mu = np$$

$$\sigma^2 = np(1-p)$$

مثال: بررسی احتمال آمدن تعدادی خاص از شیر یا خط در یک سلسله پرتاب سکه: احتمال تعداد خط آمدن در ۶ بار پرتاب سکه

- نمودار احتمال:

0	1/64
1	6/64
2	15/64
3	20/64
4	15/64
5	6/64
6	1/64



تخمین پارامتر های توزیع دو جمله ای

- حال اگر پارامتر های این توزیع رابه شکلی تخمین بزنیم (مثلا با تخمین ML):
  - مقداری برای پارامتر نشان دهنده الگو را باید انتخاب کرد که بیشترین احتمال را به دست دهد.
  - این مقدار برای توزیع دو جمله ای همان مقدار متوسط خواهد بود
- در این مثال، بیشترین مقدار برای k=3 بدست می آید.
- با بسط ایده فوق به مسئله اولیه، اساس تخمین چگالی غیر پارامتری حاصل می شود.

تخمین غیر پارامتری بر اساس هیستوگرام چند بعدی

- بر اساس این ایده باید:
  - فضا را به زیر فضا هایی تقسیم کنیم که در آنها بتوان به شمارش تعداد نمونه ها پرداخت
  - احتمال قرار گرفتن یک نمونه در این زیر فضا، عبارت خواهد بود از نسبت تعداد نمونه های شمرده شده در آن زیر فضا به تعدادکل نمونه های در اختیار
  - در صورتی که زیر فضای مزبور (حد اقل نسبت به کل فضا) کوچک باشد، نسبت یاد شده را می توان چگالی احتمال فرض نمود.

ML estimation of $P = \theta$

$$\underset{\theta}{Max}(P_k \mid \theta)$$

$$\hat{\theta} = \frac{k}{n} \cong P$$

نسبت k/n تقریب خوبی برای احتمال بوده و برای مقادیر مختلف ورودی با شمارش قابل بدست آمدن است و از روی آن برای نواحی افراز شده مقدار p(x) نیز قابل محاسبه است.

حال اگر حجم معینی از فضای ورودی R مورد توجه قرار گیرد(V) و x نقطه ای در این فضا باشد:

$$\int_{\Re} p(x')dx' \cong p(x)V \qquad (4)$$

محاسبه چگالی از روی احتمال پسین

- اگرمقدار P معلوم بوده و بتوان در ناحیه R چگالی p را ثابت فرض کرد، ممکن است آن را به عنوان یک مقدار ثابت از انتگرال یاد شده خارج نمود:

$$P = \int_{\Re} p(x')dx' \cong p(x)V$$

- در این شرایط خواهیم داشت:

$$p = P/V$$

- مقدار P بستگی به" نسبت اعضای کلاس مورد نظر به کل نمونه ها" و مقدار V به ابعاد انتخاب شده برای زیر فضای مورد نظر دارد.

بطور خلاصه:

- از نتیجه ترکیب روابط و بحث های بدست آمده می توان به یک رابطه اساسی در تخمین چگالی توزیع کلاس در ناحیه معین R با حجم v رسید:

$$p(x) \cong \frac{k/n}{V}$$

- در زیر فضای با حجم v ، که چگالی احتمال در آن ثابت فرض می شود، k نمونه وجود دارد که متعلق به کلاس مورد بررسی هستند
- مشکل اساسی: انتخاب مقدار مناسب برای v بطوریکه:
  - بتوان p را در آن محاسبه کرد
  - بتوان p را در آن ثابن فرض کرد

همگرایی و دقت تقریب
اشکال کوچک گرفتن حجم فضا

دقت تقریب زمانی قابل قبول است که حجم مورد بررسی به سمت صفر میل کند(تئوری مشتق)

$$\lim_{V \to 0\,,\, k=0} p(x) = 0 \quad (if \quad n = fixed)$$

در این شرایط این نگرانی بوجود می آید که عملا هیچ نمونه ای در ناحیه R مشاهده نخواهد شد

$$\lim_{V \to 0\,,\, k\neq 0} p(x) = \infty$$

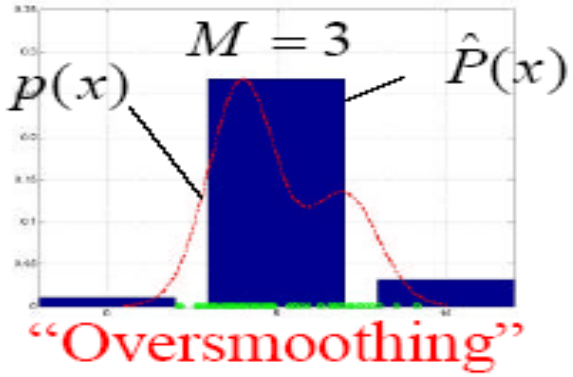
بدیهی است این جواب بلا استفاده است

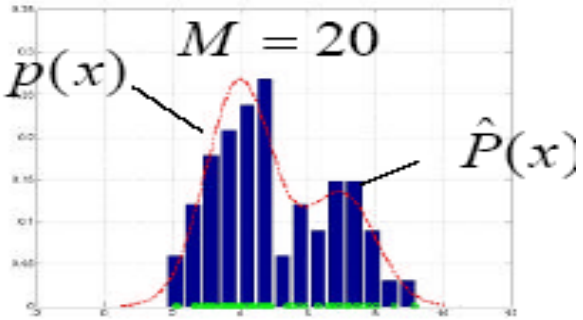
ضمنا اشکالات ناشی از توزیع گسسته نیز به وجود خواهد آمد.

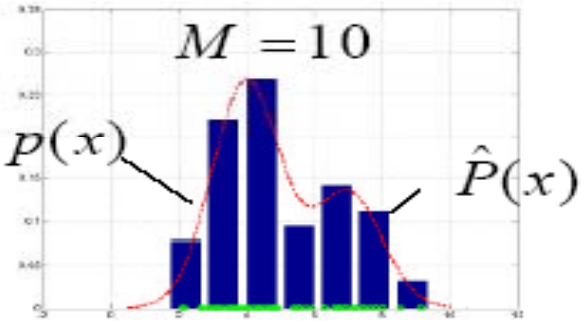
همگرایی و دقت تقریب
اشکال بزرگ گرفتن حجم فضا

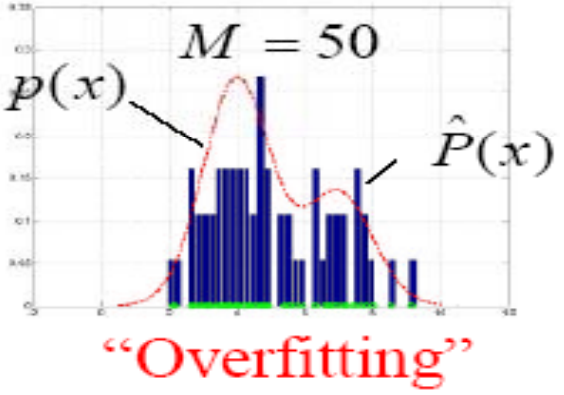
- اگر حجم فضا را بزرگ بگیریم (در تحلیل مجانبی، ∞) در این صورت:
  - در شکل مجانبی مقدار چگالی همیشه صفر است (چون مخرج کسر یعنی V مقدار بسیار بزرگی است) یعنی در عمل نتوانسته ایم چگالی توزیع معین پیدا کنیم.
  - در شکل های عملی،(V محدود ولی ثابت و بزرگ) دقت تقریب کم است (به عبارتی توزیع برای همه جا مقدار ثابت در نظر گرفته شده است مانند توزیع یکنواخت)

بررسی چند حالت مختلف
تقسیم بندی فضای مشخصه به تعداد متفاوت زیر فضا









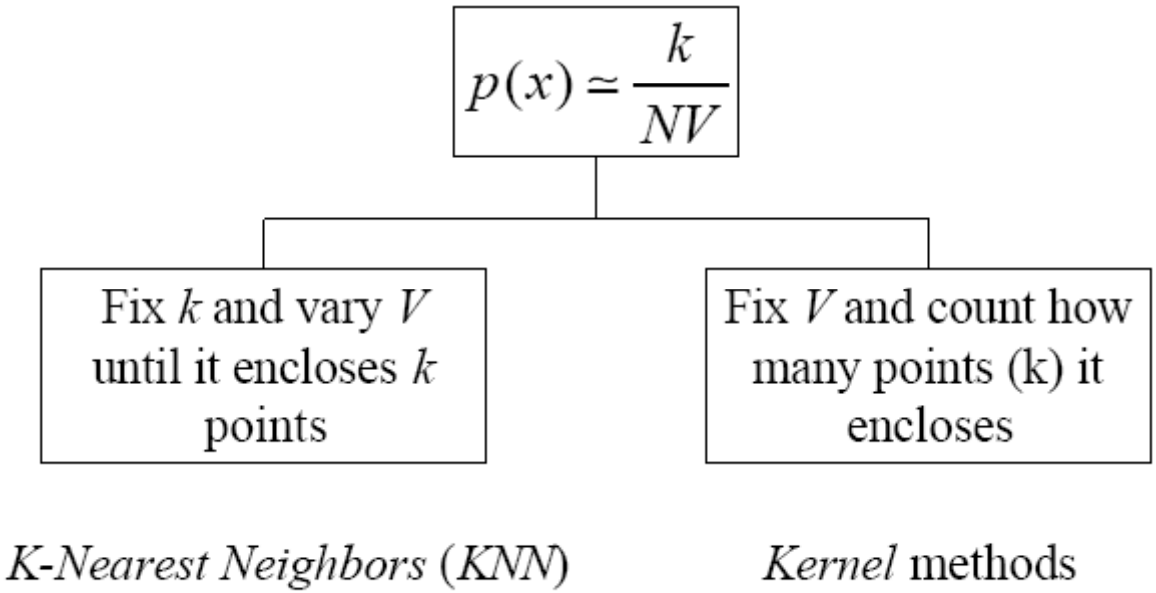
مشکل عملی هیستو گرام چند بعدی
نحوست ابعاد Curse of dimensionality

- فرض کنیم در یک مسئله ۱ بعدی بخواهیم هیستوگرام را برای ناحیه ای به حجم ۱۰٪ بدست بیاوریم. تعدادحداقل نمونه ها باید ۱۰ تا باشد(که یکی از آنها در هر یک از نواحی بیفتد)
- برای یک مسئله دو بعدی (که هر بعد آن به ۱۰ بخش تقسیم می شود) حد اقل تعداد نمونه های مورد نیاز ۱۰۰ نمونه خواهد بود.
- در فضای ۳ بعدی، این تعداد ۱۰۰۰ نمونه و در فضاهای با ابعاد بیشتر تعدا دبسیار بالاتر خواهد بود.
- این مسئله نشان می دهد که ایده هیستوگرام چند بعدی به شکل ساده می تواند در ابعاد بالا ، که در بسیاری از مسائل اجتناب ناپذیر است، به مشکل نحسی ابعاد curse of dimensionality دچار شود.

رویکرد های عملی برای محاسبه چگالی توزیع

- برای حل مسئله انتخاب حجم مناسب زیر فضا ها (در هیستو گرام چند بعدی)، ممکن است به دو طریق ممکن عمل شود:
  - حجم های ثابت به گونه ای در نظر گرفته شود (که در تمام آنها بتوان عمل شمارش و یافتن نسبت را انجام داد) و به تدریج تا جائیکه ممکن است حجم ها را کم می کنیم تا بهترین گزینه برای حجم پیدا شود. (فشرده سازی فضا)
  - به جای اینکه به حجم ثابت بیاندیشیم(و اینکه در این حجم چه تعداد نمونه شمارش می شوند) به این بیاندیشیم که با چه حجمی، تعداد معینی از نمونه ها را خواهیم داشت. بنابراین از حجم های کمی شروع کرده و به تدریج بر آن می افزائیم تا به تعداد دلخواه نمونه ها برسیم .(رشد نواحی)

رویکرد های معمول در
فشردن فضا و رشد فضا



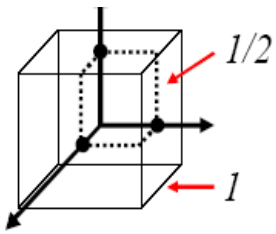
رویکرد استفاده از Hard kernel پنجره Parzen
در یافتن چگالی احتمال

- پنجره (مکعب d بعدی) به ابعاد h را به گونه ای انتخاب می کنیم که حجم درون آن v مورد نظر باشد.

$$V = h^d$$

- سپس با استفاده از یک تابع هسته kernel که باید در المان حجم ضرب شود، به شمارش تعداد نمونه ها، مبادرت می شود.

$$H(\mathbf{y}) = \begin{cases} 1 & |y_j| < 1/2 \quad j = 1,\ldots,d \\ 0 & \text{otherwise} \end{cases}$$



kernel function:

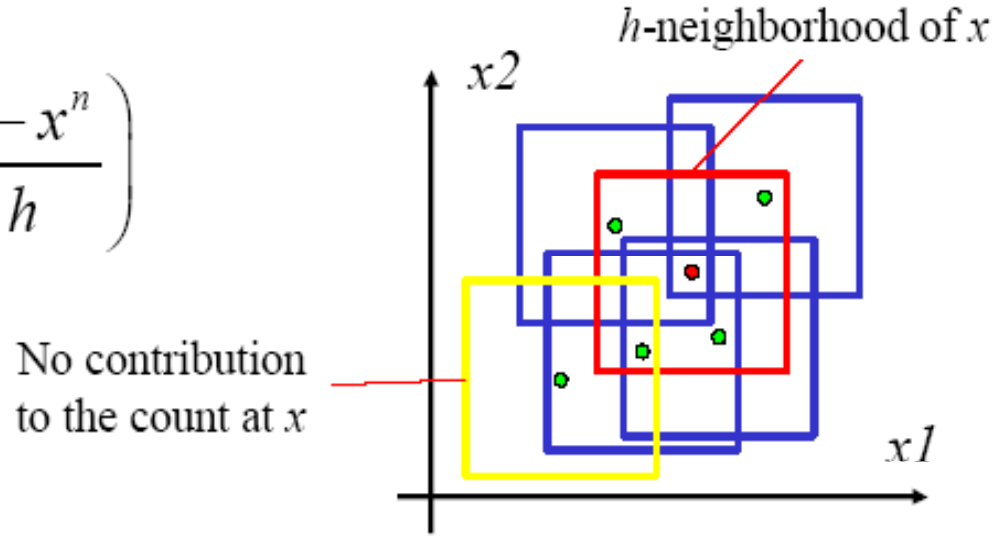
$$H(\mathbf{y}) \geq 0, \quad \forall \mathbf{y} \quad \text{and} \quad \int H(\mathbf{y}) d\mathbf{y} = 1$$

روش پنجره Paezen...

- گام بعدی انتقال مرکز این پنجره به نقاط دلخواه است
- حال به کمک این پنجره ها می توان به شمارش پرداخت

$$H\left(\left(\mathbf{x}-\mathbf{x}^n\right)/h\right)$$

$$k(x)=\sum_{n=1}^{N} H\left(\frac{x-x^n}{h}\right)$$



روش پنجره Paezen ....

$$k(x) = \sum_{n=1}^{N} H\left(\frac{x - x^n}{h}\right)$$

… is easily converted to the density estimate:

$$\tilde{p}(x) = \frac{k(x)}{NV} = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^d} H\left(\frac{x - x^n}{h}\right)$$

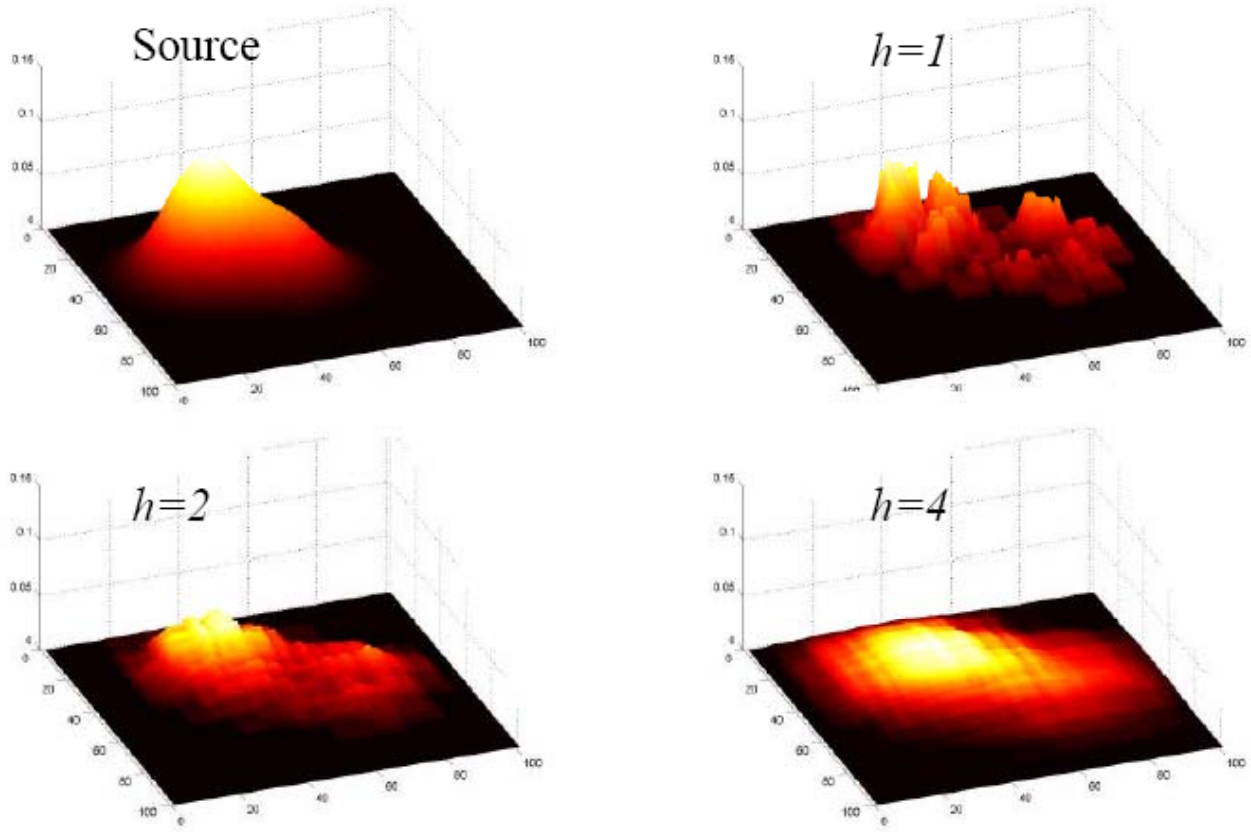
$K\left(x, x^n\right)$ Integrates to 1

Subtle point:

$$\int \left[\frac{1}{N}\sum_{n=1}^{N} K\left(x, x^n\right)\right] dx = \frac{1}{N}\sum_{n=1}^{N}\left[\int K\left(x, x^n\right) dx\right] = 1$$

$$\Rightarrow \int \tilde{p}(x) dx = 1$$

مثال نتایج تخمین با پنجره Parzen با کرنل های Hard limiter



ایرادات تابع کرنل Hard

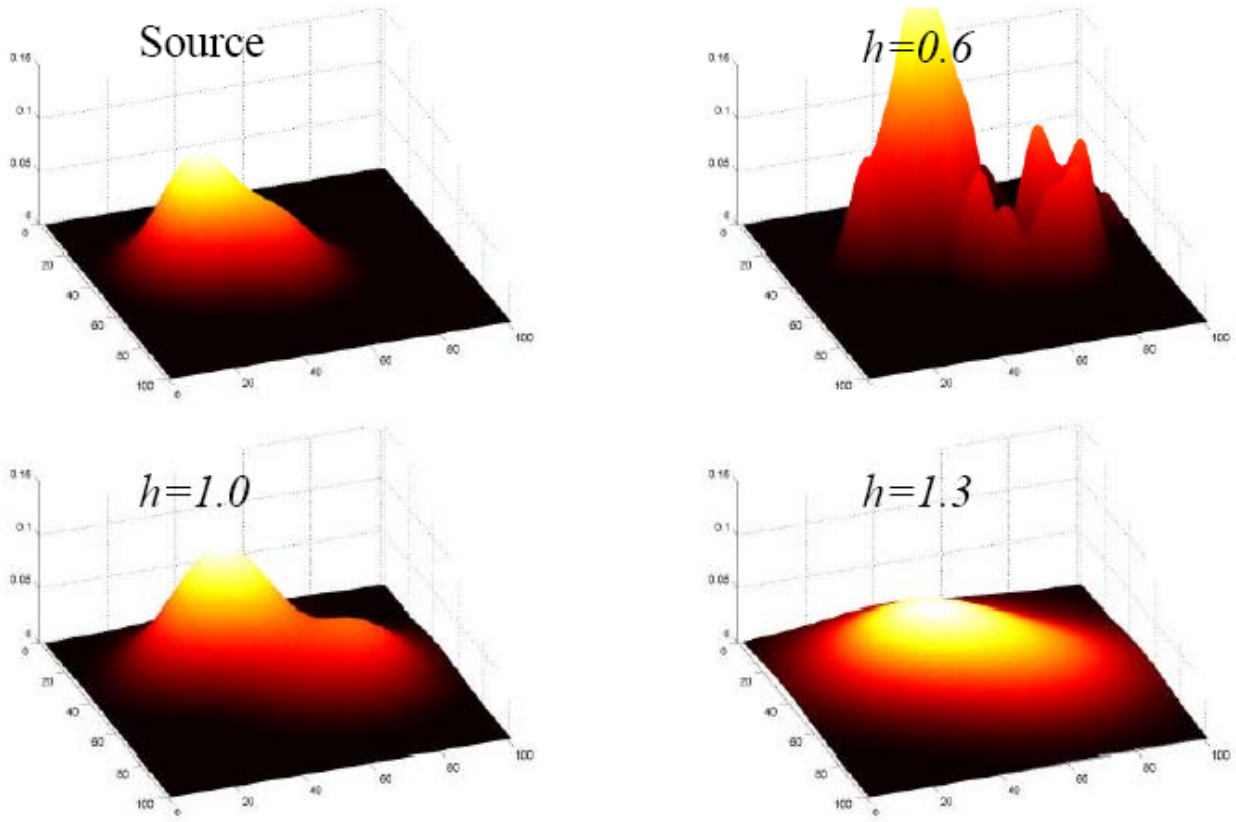
- دو ایراد :
  - غیر پیوسته بودن
  - ارزش یکسان برای همه نمونه ها
- مثال جایگزین : استفاده از توابع گوسی به عنوان کرنل در پنجره Parzen

$$K(x,x^n)=\frac{1}{(\sqrt{2\pi}h)^d}\exp\left(-\frac{\|x-x^n\|}{2h^2}\right)$$

.. so:

$$\tilde{p}(x)=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{(\sqrt{2\pi}h)^d}\exp\left(-\frac{\|x-x^n\|}{2h^2}\right)$$

مثال نتایج تخمین با کرنل گوسی



طراحی کلاسیفایر با parzen windowing

- در این رویکرد، بعد از تخمین چگالی احتمال کلاس ها، هر نمونه به کلاسی نسبت داده می شود که احتمال پسین تعلق نمونه به آن بالاتر باشد.
- بدیهی است ناحیه (تصمیم) مربوط به کلاس های مختلف، و نیز مرز های تصمیم کاملا بستگی به انتخاب تابع کرنل در تخمین دارد.
- در مواردی، در یک مسئله نیز ممکن است بهتر باشد که توابع کرنل دچار تغییر شوند

مثال: کاسیفایر دو کلاسه با استفاده از پنجره Parzen

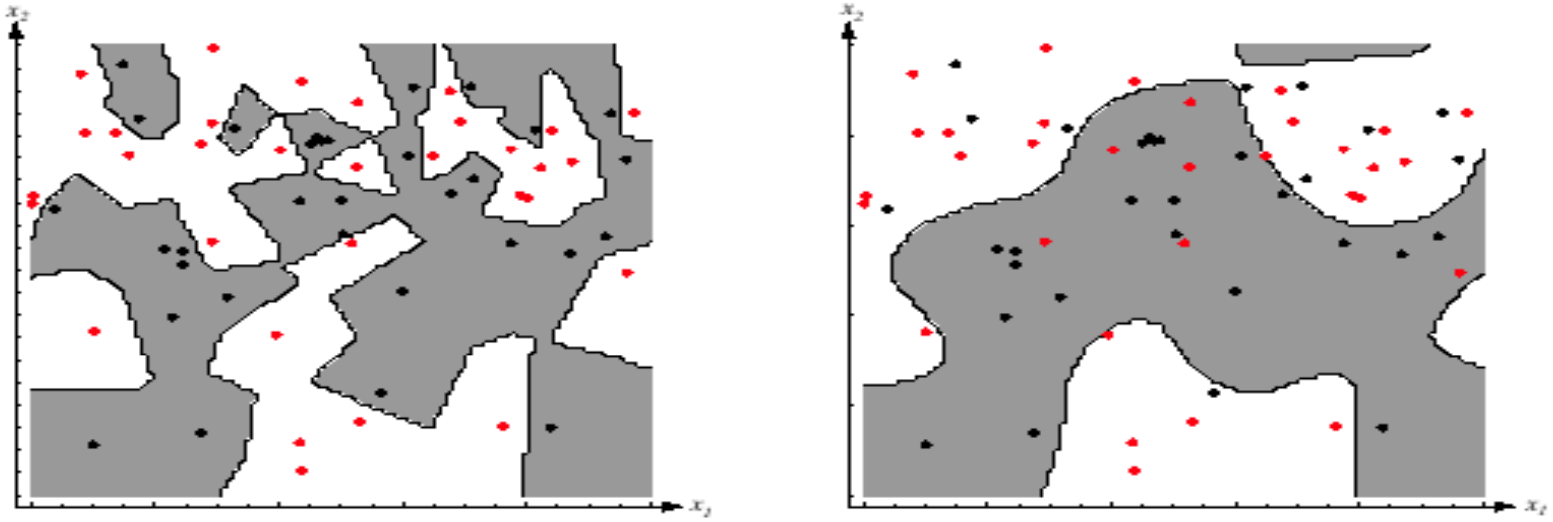


FIGURE 4.8. The decision boundaries in a two-dimensional Parzen-window dichotomizer depend on the window width h. At the left a small h leads to boundaries that are more complicated than for large h on same data set, shown at the right. Apparently, for these data a small h would be appropriate for the upper region, while a large h would be appropriate for the lower region; no single window width is ideal overall. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

روش KNN

$$\tilde{p}(x) = \frac{k}{NV}$$

Now we fix k (typically $k = \sqrt{N}$) and expand V to contain k points

This is not a true density!

Eg.: choose $N=1$, $k=1$. Then:

$$\tilde{p}(x) = \frac{1}{1 \cdot \|x - x_1\|}$$

Oops!

روش KNN...

Data: N - total points

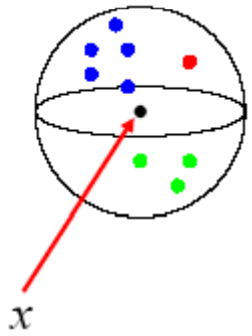
N_j - points in class ω_j

Need to find the class label for a query, x

Expand a sphere from x to include K points

K - number of neighbors of x

K_j - points of class ω_j among K



محاسبه احتمال:

Then class priors are given by: $p(\omega_j) = \frac{N_j}{N}$

We can estimate conditional and marginal densities around any x:

$$p(x \mid \omega_j) = \frac{K_j}{N_j V} \qquad p(x) = \frac{K}{NV}$$

By Bayes rule: $p(\omega_j \mid x) = \frac{K_j}{N_j V}\frac{N_j}{N}\frac{NV}{K} = \frac{K_j}{K}$

مثال کلاسیفایر Knn

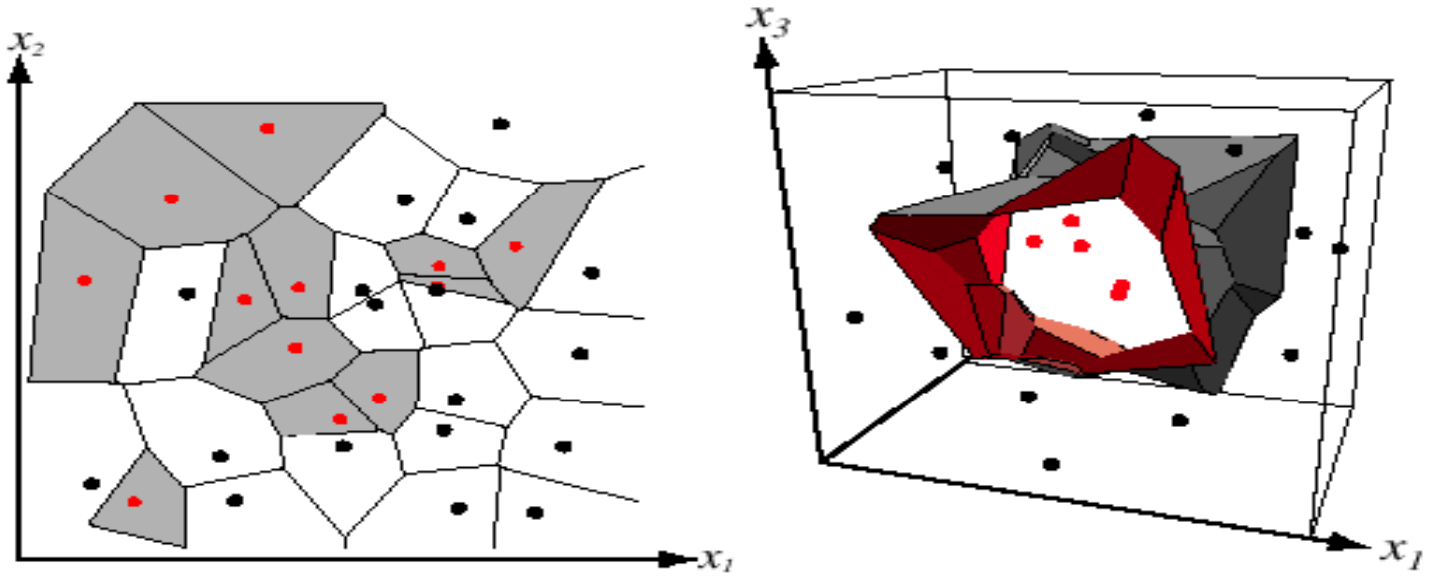


FIGURE 4.13. In two dimensions, the nearest-neighbor algorithm leads to a partitioning of the input space into Voronoi cells, each labeled by the category of the training point it contains. In three dimensions, the cells are three-dimensional, and the decision boundary resembles the surface of a crystal. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

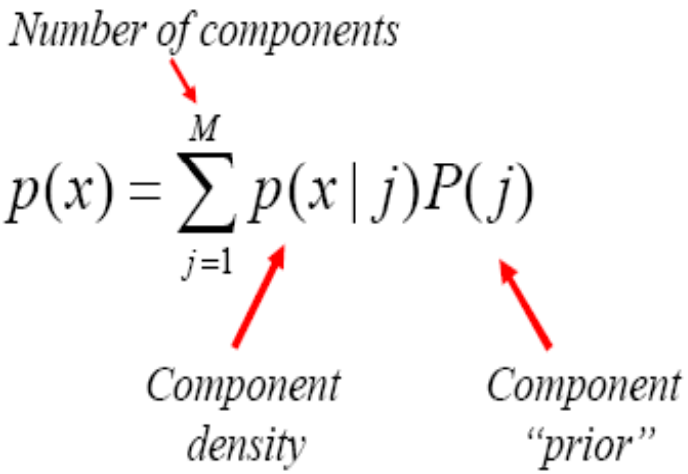
بررسی مثال

- در ابتدای کار در فضای اطراف هر نمونه بخشی در نظر گرفته می شود. (در حالت دو بعدی شبیه موزائیک های Voronoi)
- سپس با رشد این فضا، پروسه تا به آنجا پیش می رود که به تعداد معینی نمونه در زیر فضای یاد شده برسد.
- بدیهی است آن نگرانی که قبلا در مورد انتخاب حجم ثابت داشتیم، تبدیل می شود به نگرانی در مورد تعداد مناسب نمونه ها
- بر اساس یافته های تجربی عموما به عنوان اولین انتخاب، از قاعده زیر استفاده می شود:

$$K=\sqrt{N}$$

مدل های مختلط
Mixture Model

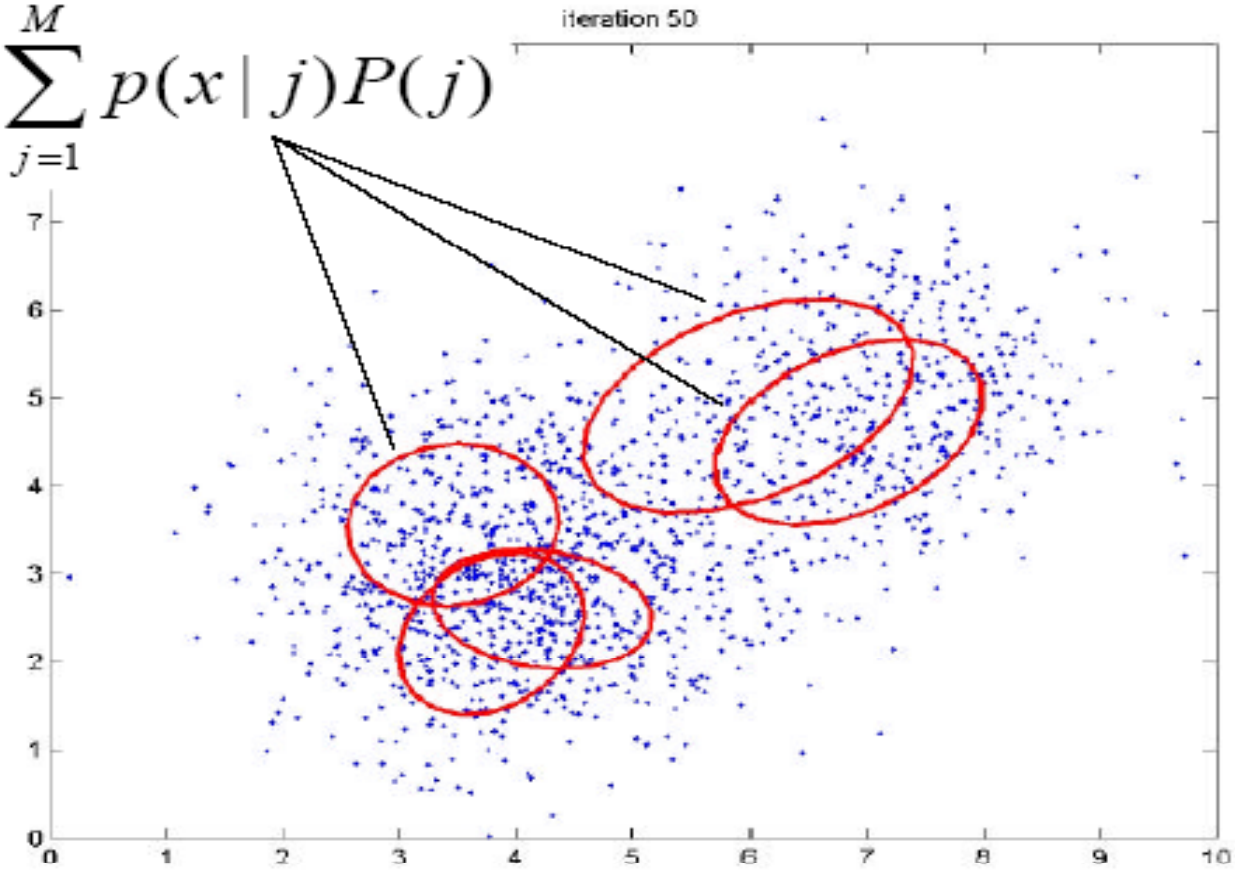
- یک تابع چگالی پیچیده که تخمین آن با روش های پارامتری یا غیر پارامتری عملی نیست را ممکن است با یک ترکیب(خطی) از تعدادی مدل ساده نمایش داد.
- این شیوه نمایش را تخمین با مدل مختلط می نامند.



$$P(j) \geq 0, \quad \forall j \qquad \text{and} \qquad \sum_{j=1}^{M} P(j) = 1$$

مثال

$$p(x) = \sum_{j=1}^{M} p(x \mid j) P(j)$$



مثال- مدل مختلط گوسی

- فرض کنیم توزیع n نمونه را بخواهیم با ترکیب خطی M تابع گوسی با مقادیر mean و واریانس مختلف نشان دهیم پارامتر های این توابع باید چگونه باشند؟
- روش معمول بر اساس بهینه سازی است.
- در این جا نیز می توان به جای خود توابع چگالی، از لگاریتم آنها استفاده نمود تا به مسئله ساده تری منجر گردد.

....

$$l(\theta) \equiv \log \prod_{n=1}^{N} p(x^n) = \sum_{n=1}^{N} \log \left\{ \sum_{j=1}^{M} p(x^n \mid j) P(j) \right\}$$

Differentiate w.r.t. parameters:

$$\nabla_{\theta_j} l(\theta) = \sum_{n=1}^{N} \frac{\partial}{\partial \theta_j} \log \left\{ \sum_{k=1}^{M} p(x^n \mid k) P(k) \right\}$$

$$= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{M} p(x^n \mid k) P(k)} \frac{\partial}{\partial \theta_j} p(x^n \mid j) P(j)$$

mean برای هر یک از مولفه ها

$$\frac{\partial l(\theta)}{\partial \mu_j} = \sum_{n=1}^{N} P(j \mid x^n)\left[\Sigma_j^{-1}(x^n - \hat{\mu}_j)\right]$$

Setting it to *0* and solving for μ_j:

$$\hat{\mu}_j = \frac{\sum_{n=1}^{N} P(j \mid x^n) x^n}{\sum_{n=1}^{N} P(j \mid x^n)}$$

- convex sum of all data

و برای کوواریانس

$$\frac{\partial l(\theta)}{\partial \sigma_j^2} = \sum_{n=1}^{N} P(j \mid x^n)\left[\hat{\mathbf{S}}_j^{-1} - \hat{\mathbf{S}}_j^{-1}(x^n - \hat{\mu}_j)(x^n - \hat{\mu}_j)^T \hat{\mathbf{S}}_j^{-1}\right]$$

Setting it to *0* and solving for $\boldsymbol{\Sigma}_i$:

$$\hat{\mathbf{S}}_j = \frac{\sum_{n=1}^{N} P(j \mid x^n)\left(x^n - \hat{\mu}_j\right)\left(x^n - \hat{\mu}_j\right)^T}{\sum_{n=1}^{N} P(j \mid x^n)}$$

خلاصه بحث

- روش های غیر پارامتری تخمین تابع
  - روش مبتنی بر هیستوگرام چند بعدی
  - روش های مبتنی بر کرنل(parzen window)
  - روش KNN
  - روش چگالی مختلط