بازشناسی الگو

درس ۶

طبقه‌بندی مبتنی بر نظریه‌ی تصمیم بیز

روش‌های تخمین ناپارامتری تابع چگالی احتمال مجهول

Classification Based on Bayes Decision Theory

Nonparametric Estimation Methods for Unknown Probability Density Functions

کاظم فولادی

دانشکده مهندسی برق و کامپیوتر

دانشگاه تهران

http://courses.fouladi.ir/pr

تخمین توابع چگالی احتمال مجهول

روی‌کرد ناپارامتری

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.

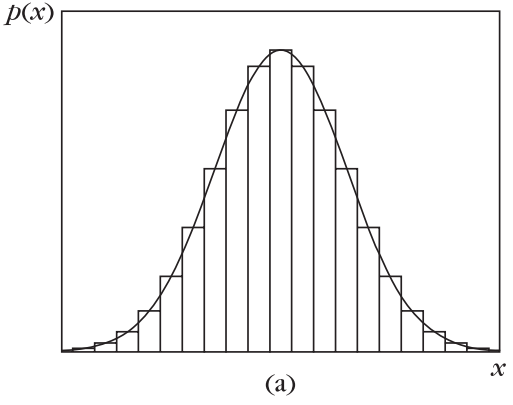
روی‌کردهای تخمین توابع چگالی احتمال مجهول

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

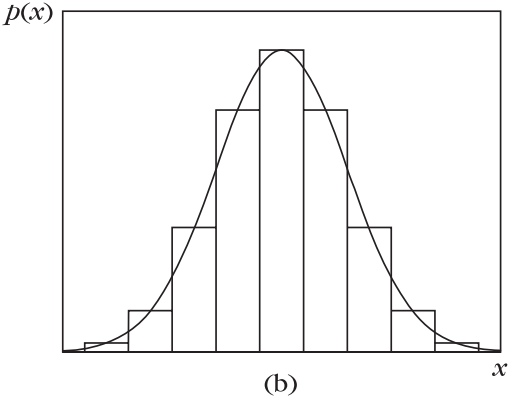
تخمین توابع چگالی احتمال مجهول

رویکرد ناپارامتری: استفاده از هیستوگرام

محور x (فضای یک‌بعدی) به binهای متوالی با طول h تقسیم می‌شود.
سپس احتمال اینکه یک نمونه‌ی x در یک bin قرار گیرد، برای همه‌ی binها برآورد می‌شود.

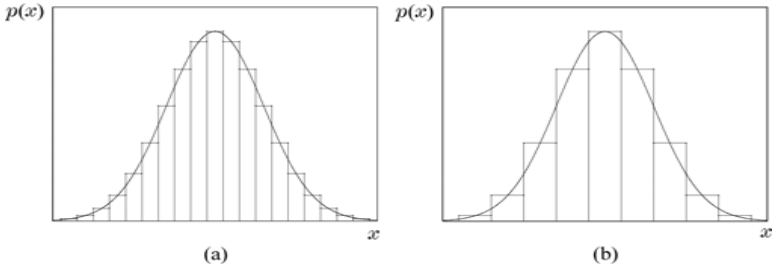


تخمین با binهای کوچک‌تر



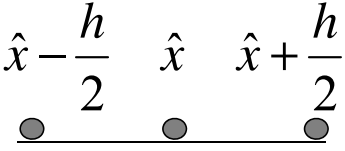
تخمین با binهای بزرگ‌تر

❖ Nonparametric Estimation



$$P \approx \frac{k_N}{N} \begin{cases} \rightarrow k_N \text{ in } h \\ \rightarrow N \text{ total} \end{cases}$$

- $\hat{p}(x) \equiv \hat{p}(\hat{x}) = \frac{1}{h}\frac{k_N}{N}, \left|x - \hat{x}\right| \le \frac{h}{2}$



- If $p(x)$ continuous, $\hat{p}(x) \rightarrow p(x)$

  as $N \rightarrow \infty$, if $h_N \rightarrow 0$, $k_N \rightarrow \infty$, $\frac{k_N}{N} \rightarrow 0$

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

۱

روش پنجره‌های پارزن

تخمین توابع چگالی احتمال مجهول

روی‌کرد ناپارامتری: روش پنجره‌های پارزن

PARZEN WINDOWS

فضای l–بعدی به ابرمکعب‌هایی با طول ضلع h و حجم h^l تقسیم می‌شود.

برای تخمین تابع چگالی احتمال در نقطه‌ی x:

- یک ابرمکعب به طول ضلع h را با مرکز نقطه‌ی x در نظر می‌گیریم.
- تعداد نقاط نمونه‌ی موجود در این ابرمکعب را می‌شماریم (k_N)
- تخمین می‌شود: خارج قسمت حاصل‌جمع k_N بر حجم ابرمکعب h^l و تعداد کل نمونه‌ها (N).

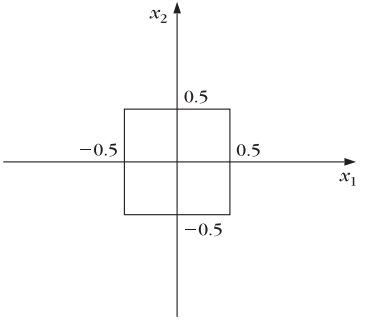
تخمین توابع چگالی احتمال مجهول

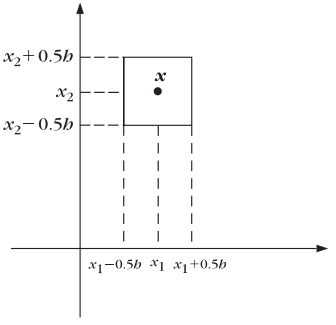
روی‌کرد ناپارامتری: روش پنجره‌های پارزن

PARZEN WINDOWS

برای محاسبه، تابع ϕ به صورت زیر تعریف می‌شود:

$$\phi(\mathbf{x}_i) = \begin{cases} 1 & , |x_{ij}| \leq \frac{1}{2} \\ 0 & , \text{otherwise} \end{cases}$$





در نتیجه فرمول $\hat{p}(\mathbf{x}) = \frac{1}{h}\frac{k_N}{N}$ به صورت زیر بازنویسی می‌شود:

$$\hat{p}(\mathbf{x}) = \frac{1}{h^l}\left(\frac{1}{N}\sum_{i=1}^{N}\phi\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)\right)$$

مشکل:

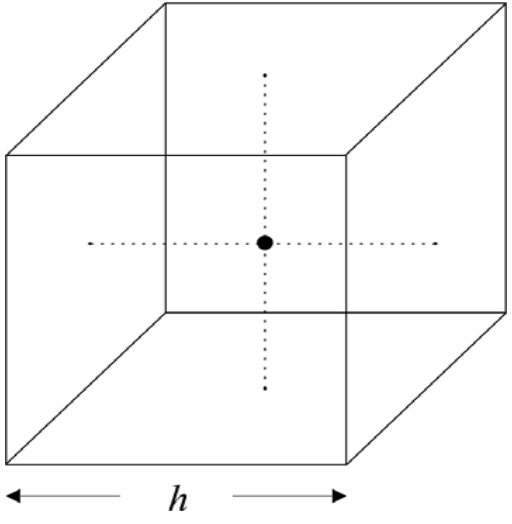
تخمین یک تابع پیوسته $p(x)$ از طریق بسط آن بر حسب توابع پله‌ای ناپیوسته $\phi(.)$

⇐ برآورد حاصل از این مورد تأثیر می‌پذیرد.

راه‌حل: جایگزینی یک تابع هموار برای $\phi(.)$ و تعمیم با فرض pdf بودن $\phi(.)$

❖ Parzen Windows

- ➢ Divide the multidimensional space in hypercubes



➢ Define $$\varphi(\underline{x}_i) = \begin{cases} 1 & \left|x_{ij}\right| \le \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

- That is, it is 1 inside a unit side hypercube centered at 0

$$\hat{p}(\underline{x}) = \frac{1}{h^l}\left(\frac{1}{N}\sum_{i=1}^{N}\varphi\left(\frac{\underline{x}_i - \underline{x}}{h}\right)\right)$$

$$\frac{1}{\text{volume}} \times \frac{1}{N} \times \text{number of points inside an } h\text{-side hypercube centered at } \underline{x}$$

➢ The problem: $p(\underline{x})$ continuous

$\varphi(.)$ discontinuous

➢ **Parzen windows**: (kernels, potential functions)

$$\varphi(\underline{x}) \text{ is smooth}, \quad \varphi(\underline{x}) \ge 0, \quad \int_{\underline{x}} \varphi(\underline{x}) d\underline{x} = 1$$

➢ Mean value

$$E[\hat{p}(\underline{x})] = \frac{1}{h^l}\left(\frac{1}{N}\sum_{i=1}^{N} E[\varphi(\frac{\underline{x}_i - \underline{x}}{h})]\right) = \int_{\underline{x}'} \frac{1}{h^l}\varphi(\frac{\underline{x}' - \underline{x}}{h}) p(\underline{x}')d\underline{x}'$$

$$h \to 0,\ \frac{1}{h^l} \to \infty$$

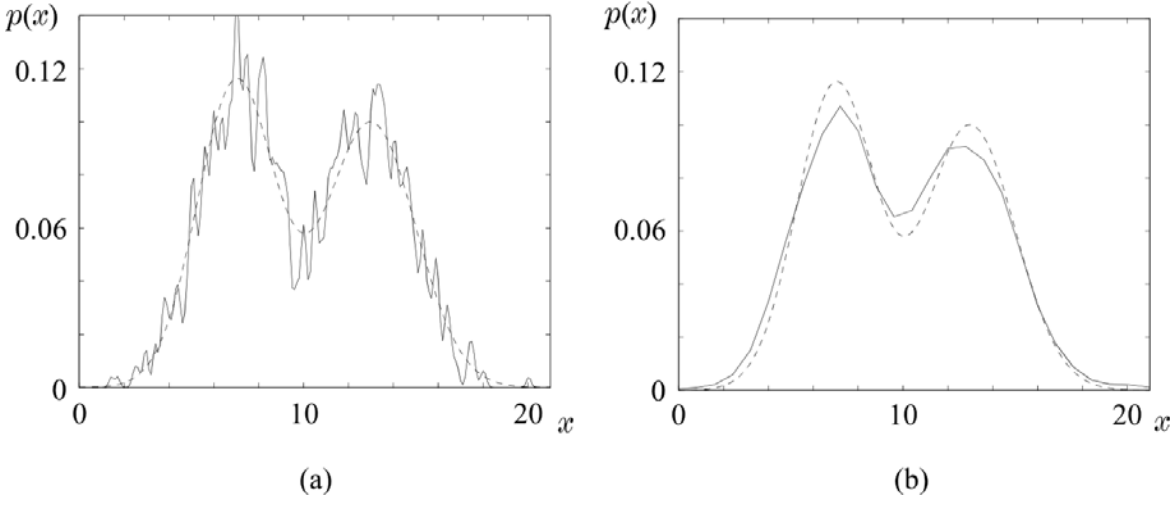
- $h \to 0$ the width of $\varphi(\frac{\underline{x}' - \underline{x}}{h}) \to 0$
- $\int \frac{1}{h^l}\varphi(\frac{\underline{x}' - \underline{x}}{h})d\underline{x} = 1$
- $h \to 0 \qquad \frac{1}{h^l}\varphi(\frac{\underline{x}}{h}) \to \delta(\underline{x})$
- $E[\hat{p}(x)] = \int_{\underline{x}'} \delta(\underline{x}' - \underline{x}) p(\underline{x}')d\underline{x}' = p(\underline{x})$

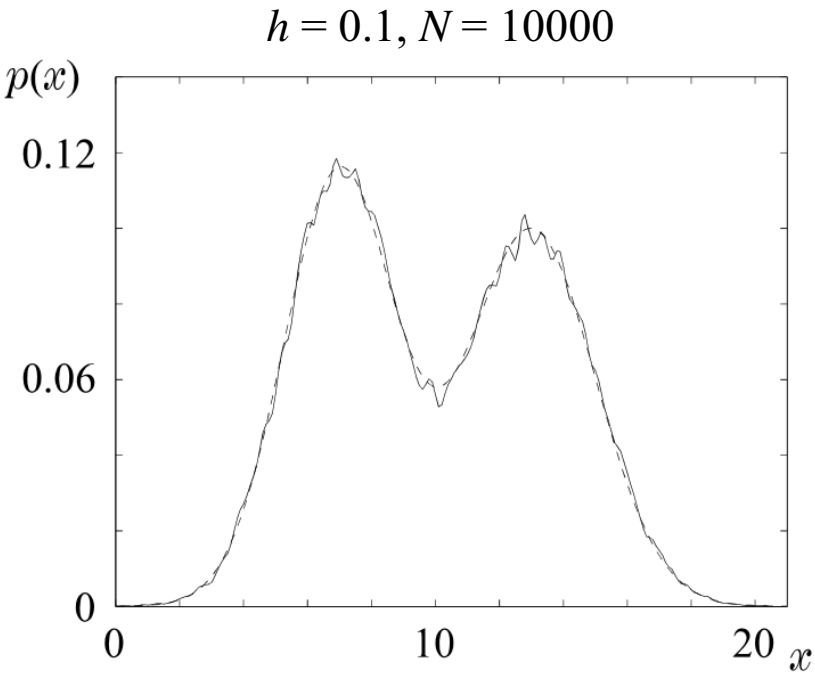
Hence **unbiased** in the limit

➢ Variance

- The smaller the h the higher the variance

$h = 0.1, N = 1000$ $h = 0.8, N = 1000$





(b)

➢The higher the N the better the accuracy

➢ If

- $h \to 0$
- $N \to \infty$
- $h_N \to \infty$

asymptotically unbiased

➢ The classification method:

- Remember: $l_{12} \equiv \dfrac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)} \gtrless \dfrac{P(\omega_2)}{P(\omega_1)} \dfrac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}} \equiv \theta$

- $$\frac{\frac{1}{N_1 h^l} \sum_{i=1}^{N_1} \varphi\left(\frac{\underline{x}_i - \underline{x}}{h}\right)}{\frac{1}{N_2 h^l} \sum_{i=1}^{N_2} \varphi\left(\frac{\underline{x}_i - \underline{x}}{h}\right)} \gtrless \theta$$

❖ CURSE OF DIMENSIONALITY

- In all the methods, so far, we saw that the highest the number of points, N, the better the resulting estimate.

- If in the one-dimensional space an interval, filled with N points, is adequately (for good estimation), in the two-dimensional space the corresponding square will require N^2 and in the ℓ-dimensional space the ℓ-dimensional cube will require N^ℓ points.

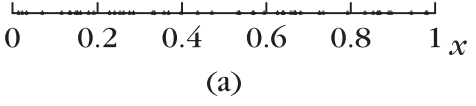
- The exponential increase in the number of necessary points in known as the curse of dimensionality. This is a major problem one is confronted with in high dimensional spaces.

تخمین توابع چگالی احتمال مجهول

روی‌کرد ناپارامتری: روش پنجره‌های پارزن: مشکل بعدیت

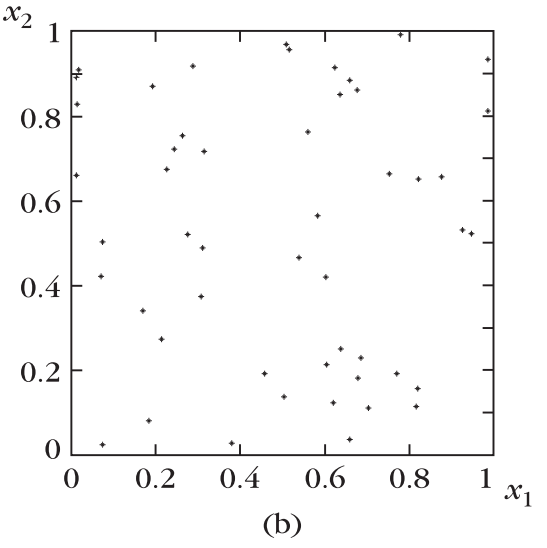
PARZEN WINDOWS: THE CURSE OF DIMENSIONALITY

تعداد نقاط کافی برای تخمین مناسب، با افزایش بعد به‌صورت نمایی افزایش می‌یابد.



(a)

۵۰ نقطه تولید شده از توزیع یکنواخت
واقع بر یک فضای یک–بعدی



(b)

۵۰ نقطه تولید شده از توزیع یکنواخت
واقع بر یک فضای دو–بعدی

نقاط در فضای دوبعدی پراکنده‌تر هستند.

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

۲

روش تخمین چگالی چند نزدیک‌ترین همسایه

تخمین توابع چگالی احتمال مجهول

روی‌کرد ناپارامتری: روش تخمین چند نزدیک‌ترین همسایه

k Nearest Neighbor Density Estimation (k NN)

در تخمین پارزن، حجم پیرامون نقطه‌ی x ثابت فرض می‌شد (h^l)
و تعداد نقاط واقع درون این حجم به طور تصادفی از یک نقطه به نقطه‌ی دیگر متغیر بود (k_N)

در تخمین kNN، تعداد نقاط $(k_N = k)$ را ثابت می‌گیریم
و اندازه‌ی حجم پیرامون نقطه‌ی x را به‌گونه‌ای تغییر می‌دهیم که k نقطه را در برگیرد.

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

$\hat{p}(\mathbf{x})$ یک تخمین‌گر ناریب مجانبی برای $p(\mathbf{x})$ خواهد بود (وقتی $k \to \infty, N \to \infty, k/N \to 0$)

نتیجه:
در نواحی کم چگالی، حجم بزرگ خواهد شد
و
در نواحی پرچگالی، حجم کوچک خواهد شد

نکته: حجم پیرامونی، لزوماً ابرمکعب نیست.

تخمین توابع چگالی احتمال مجهول

روی‌کرد ناپارامتری: روش تخمین چند نزدیک‌ترین همسایه

k NEAREST NEIGHBOR DENSITY ESTIMATION (*k* NN)

از دیدگاه عملی:
در برخورد با یک بردار ویژگی $\mathbf{x}$،
فاصله‌ی آن را تا همه‌ی بردارهای ویژگی آموزشی طبقه‌های مختلف محاسبه می‌کنیم (d)

ابربیضی‌گون
Hyperellipsoid

$$V = V_0 |\Sigma|^{l/2} r^l$$

$$V_0 = \frac{\pi^{l/2}}{\Gamma(1+l/2)}$$

فاصله‌ی ماهالونوبیس
حجم: ابربیضی‌گون

ابرکره
Hypersphere

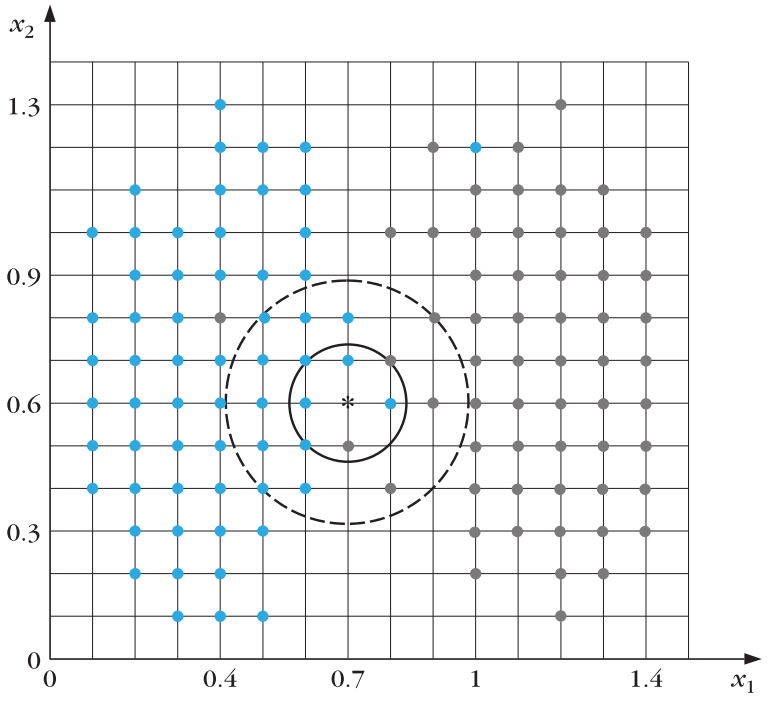
$$V = \frac{\pi^{l/2} r^l}{\Gamma(1+l/2)}$$

فاصله‌ی اقلیدسی
حجم: ابرکره

تخمین توابع چگالی احتمال مجهول

روی‌کرد ناپارامتری: روش تخمین چند نزدیک‌ترین همسایه: مثال

k NEAREST NEIGHBOR DENSITY ESTIMATION (k NN)



می‌خواهیم جرم احتمال تعلق * به طبقه‌های آبی و طوسی را مشخص کنیم:

۱) برای $k = 5$ دایره را آن‌قدر بزرگ می‌گیریم تا حداقل 5 نمونه آبی در آن قرار بگیرد.

۲) برای $k = 5$ دایره را آن‌قدر بزرگ می‌گیریم تا حداقل 5 نمونه طوسی در آن قرار بگیرد.

تخمین توابع چگالی احتمال مجهول

روی‌کرد ناپارامتری: روش تخمین چند نزدیک‌ترین همسایه: مثال (دو طبقه)

k NEAREST NEIGHBOR DENSITY ESTIMATION (k NN)

از دیدگاه عملی:
در برخورد با یک بردار ویژگی $\mathbf{x}$،
فاصله‌ی آن را تا همه‌ی بردارهای ویژگی آموزشی طبقه‌های مختلف محاسبه می‌کنیم (d)

برای حالت دو طبقه‌ای:

r_1 شعاع ابرکره به مرکز $\mathbf{x}$ شامل k نقطه از طبقه ω_1 (با حجم V_1)
r_2 شعاع ابرکره به مرکز $\mathbf{x}$ شامل k نقطه از طبقه ω_2 (با حجم V_2)

(k می‌تواند برای طبقه‌های مختلف، متفاوت باشد)

$$l_{12} \equiv \frac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)} = \frac{kN_2V_2}{kN_1V_1} \gtrless \frac{P(\omega_2)}{P(\omega_1)} \frac{\lambda_{21}-\lambda_{22}}{\lambda_{12}-\lambda_{11}} \equiv \theta$$

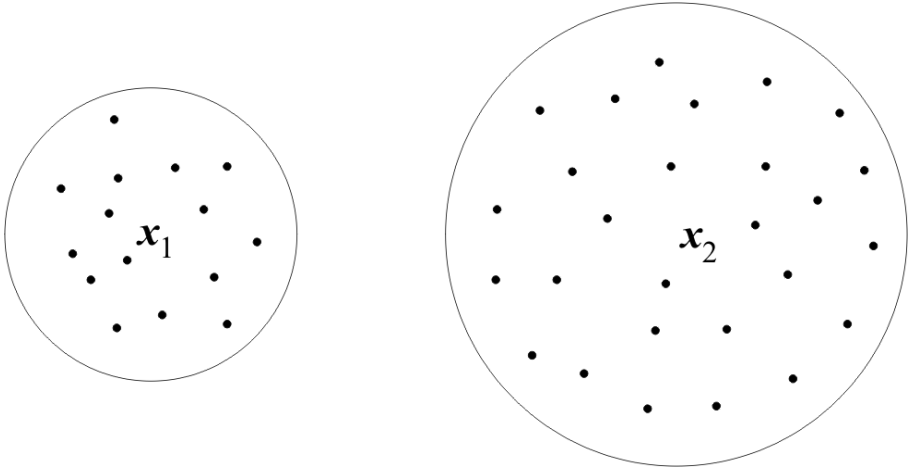
❖ K Nearest Neighbor Density Estimation

➢ In Parzen:

- The volume is constant
- The number of points in the volume is varying

➢ Now:

- Keep the number of points $k_N = k$ constant
- Leave the volume to be varying
- $$\hat{p}(\underline{x}) = \frac{k}{NV(\underline{x})}$$



$$\frac{\frac{k}{N_1 V_1}}{\frac{k}{N_2 V_2}} = \frac{N_2 V_2}{N_1 V_1} \gtrless \theta$$

تخمین توابع چگالی احتمال مجهول

روی‌کرد ناپارامتری

کارآیی روش‌های تخمین ناپارامتری،
به‌عنوان تخمین‌گر تابع چگالی احتمال در فضاهای با بعد بالا تنزل می‌کند.
(به دلیل فقدان داده‌ی آموزشی کافی)

اما کارآیی آنها به‌عنوان طبقه‌بندی کننده به اندازه‌ی کافی خوب است.
(هر چند فقدان داده‌ی آموزشی کافی بر همه‌ی روش‌ها اثر می‌گذارد)

شبکه‌های عصبی احتمالاتی یک روش کارآمد برای پیاده‌سازی محاسبات طبقه‌بندی‌کننده است

بازشناسی الگو

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

۳

قاعده‌ی نزدیک‌ترین همسایه

قاعده‌ی نزدیک‌ترین همسایه

قاعده‌ی نزدیک‌ترین همسایه

k NEAREST NEIGHBOR RULE

با یک تغییر در تخمین‌گر چگالی $k\text{NN}$ به یک طبقه‌بندی‌کننده‌ی غیرخطی زیربهینه اما متداول در عمل می‌رسیم.

الگوریتم (قاعده‌ی نزدیک‌ترین همسایه)

با داشتن یک بردار ویژگی $\mathbf{x}$ و یک معیار فاصله:

۱) از میان N بردار ویژگی آموزشی، k نزدیک‌ترین همسایه را مشخص می‌کنیم (بدون توجه به برچسب طبقه)
(در حالت دوطبقه‌ای: k باید فرد باشد؛ در حالت M طبقه‌ای، k باید مضرب M نباشد)

۲) از میان این k نمونه، تعداد بردارهای k_i که متعلق به طبقه‌ی ω_i است را مشخص می‌کنیم $(\sum_{i=1}^{M} k_i = k)$

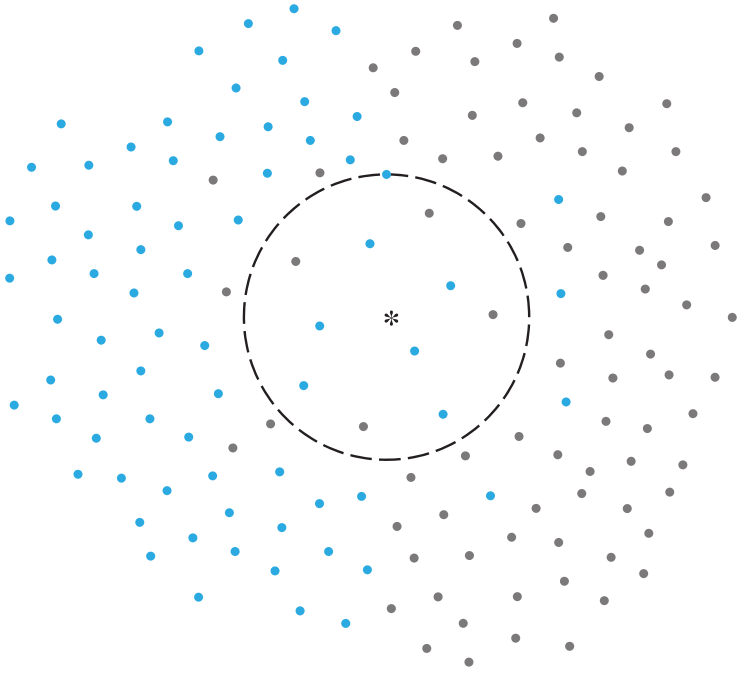
۳) $\mathbf{x}$ را به طبقه‌ای نسبت می‌دهیم که ماکزیمم نمونه‌ها (k_i) را دارد:

$$x \in \omega_{i^*} \quad : \quad i^* = \arg\max_i k_i$$

تخمین توابع چگالی احتمال مجهول

قاعده‌ی نزدیک‌ترین همسایه: مثال

k Nearest Neighbor Density Estimation (k NN)



می‌خواهیم مشخص کنیم ✻ به کدام طبقه تعلق دارد:

برای $k = 11$ دایره را آن‌قدر بزرگ می‌گیریم تا حداقل 11 نمونه در آن قرار بگیرد.
تعداد دایره آبی: ۷
تعداد دایره خاکستری: ۴

پس ✻ به طبقه‌ی آبی متعلق است.

قاعده‌ی نزدیک‌ترین همسایه

ساده‌ترین حالت

k NEAREST NEIGHBOR RULE

قاعده‌ی نزدیک‌ترین همسایه (NN rule): در ساده‌ترین حالت $(k = 1)$
$\mathbf{x}$ به طبقه‌ی نزدیک‌ترین همسایه‌ی خود نسبت داده می‌شود.

اگر تعداد نمونه‌های آموزشی به اندازه‌ی کافی بزرگ باشد،
کارآیی این روش ساده، بالا خواهد بود.

❖ The Nearest Neighbor Rule

- Choose k out of the N training vectors, identify the k nearest ones to $\underline{x}$

- Out of these k identify k_i that belong to class ω_i

- Assign $\underline{x} \rightarrow \omega_i : \; k_i > k_j \;\; \forall i \neq j$

- The simplest version

  $k = 1$!!!

- For large N this is not bad. It can be shown that: if P_B is the optimal Bayesian error probability, then:

$$P_B \leq P_{NN} \leq P_B\left(2 - \frac{M}{M-1} P_B\right) \leq 2P_B$$

- $P_B \leq P_{kNN} \leq P_B + \sqrt{\frac{2P_{NN}}{k}}$

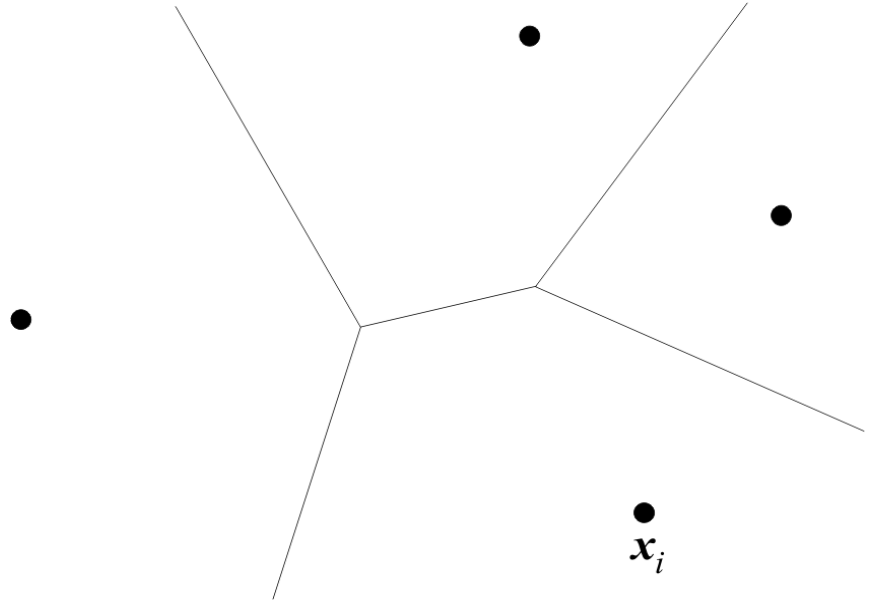
- $\boxed{k \to \infty,\ P_{kNN} \to P_B}$

- For small P_B:

$$P_{NN} \cong 2P_B$$

$$P_{3NN} \cong P_B + 3(P_B)^2$$

❖ Voronoi tesselation



$$R_i = \left\{ \underline{x} : d(\underline{x}, \underline{x}_i) < d(\underline{x}, \underline{x}_j), i \neq j \right\}$$

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

۴

طبقه‌بندی بیز ساده

طبقه‌بندی‌کننده‌ی بیز ساده

NAÏVE-BAYES CLASSIFIER

اگر $\mathbf{x} \in \mathbb{R}^l$ و هدف تخمین

$$p(\mathbf{x}|\omega_i), \qquad i = 1,2, \dots, M$$

باشد، برای یک تخمین خوب این تابع چگالی احتمال، به N^l نقطه نیاز داریم.

اما اگر $x_1, x_2, \dots, x_l$ دوبه‌دو مستقل باشند، آن‌گاه داریم:

$$p(\mathbf{x}|\omega_i) = \prod_{j=1}^{l} p(x_j|\omega_i)$$

در این صورت برای هر $p(x_j|\dots)$ به N نقطه نیاز داریم
پس در مجموع می‌شود $N \times l$ نقطه.

طبقه‌بندی در روش Naïve-Bayes حتی زمانی که شرط استقلال نقض می‌شود، هم خوب عمل می‌کند.

❖ NAIVE-BAYES CLASSIFIER

➢ Let $\underline{x} \in \Re^{\ell}$ and the goal is to estimate $p(\underline{x} \mid \omega_i)$
$i = 1, 2, \ldots, M$.
For a "good" estimate of the pdf one would need, say, N^{ℓ} points.

➢ Assume $x_1, x_2, \ldots, x_{\ell}$ mutually independent. Then:

$$p(\underline{x} \mid \omega_i) = \prod_{j=1}^{\ell} p(x_j \mid \omega_i)$$

➢ In this case, one would require, roughly, N points for each pdf.
Thus, a number of points of the order $N \cdot \ell$ would suffice.

➢ It turns out that the **Naïve – Bayes classifier** works reasonably well even in cases that violate the independence assumption.

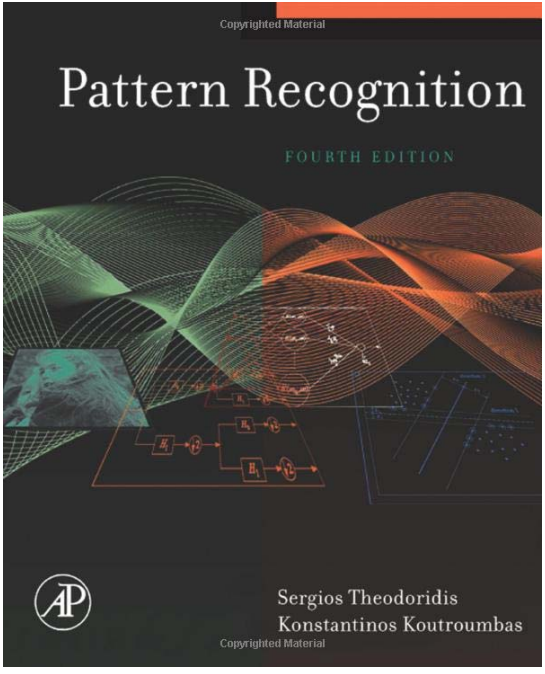
بازشناسی الگو

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

۴

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.

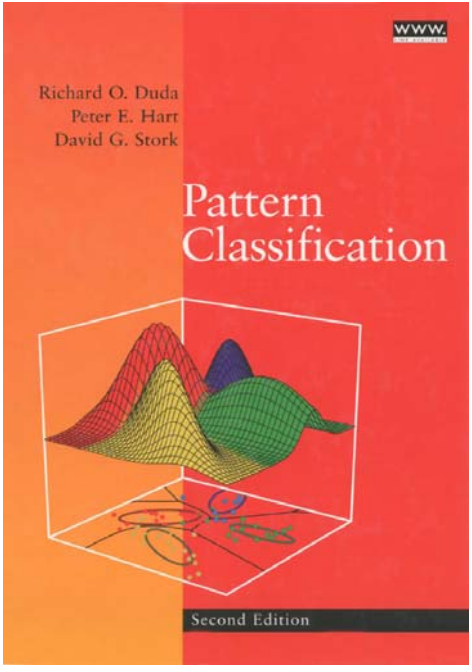
Given a classification task of M classes, $\omega_1, \omega_2, \ldots, \omega_M$, and an unknown pattern, which is represented by a feature vector $\boldsymbol{x}$, we form the M conditional probabilities $P(\omega_i|\boldsymbol{x}), i = 1, 2, \ldots, M$. Sometimes, these are also referred to as *a posteriori 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 4

CHAPTER
4

NONPARAMETRIC TECHNIQUES

4.1 INTRODUCTION

In Chapter 3 we treated supervised learning under the assumption that the forms of the underlying density functions were known. Alas, in most pattern recognition applications this assumption is suspect; the common parametric forms rarely fit the densities actually encountered in practice. In particular, all of the classical parametric densities are unimodal (have a single local maximum), whereas many practical problems involve multimodal densities. Furthermore, our hopes are rarely fulfilled that a high-dimensional density might be accurately represented as the product of one-dimensional functions. In this chapter we shall examine *nonparametric* procedures that can be used with arbitrary distributions and without the assumption that the forms of the underlying densities are known.

There are several types of nonparametric methods of interest in pattern recognition. One consists of procedures for estimating the density functions $p(\mathbf{x}|\omega_j)$ from sample patterns. If these estimates are satisfactory, they can be substituted for the true densities when designing the classifier. Another consists of procedures for directly estimating the *a posteriori* probabilities $P(\omega_j|\mathbf{x})$. This is closely related to nonparametric design procedures such as the nearest-neighbor rule, which bypass probability estimation and go directly to decision functions.

4.2 DENSITY ESTIMATION

The basic ideas behind many of the methods of estimating an unknown probability density function are very simple, although rigorous demonstrations that the estimates converge require considerable care. The most fundamental techniques rely on the fact that the probability P that a vector $\mathbf{x}$ will fall in a region $\mathcal{R}$ is given by

$$P = \int_{\mathcal{R}} p(\mathbf{x}')\, d\mathbf{x}'. \qquad (1)$$

161