Neparametrické odhady hustoty pravděpodobnosti

Václav Hlaváč

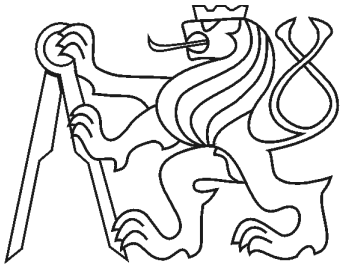
Elektrotechnická fakulta ČVUT

Katedra kybernetiky

Centrum strojového vnímání

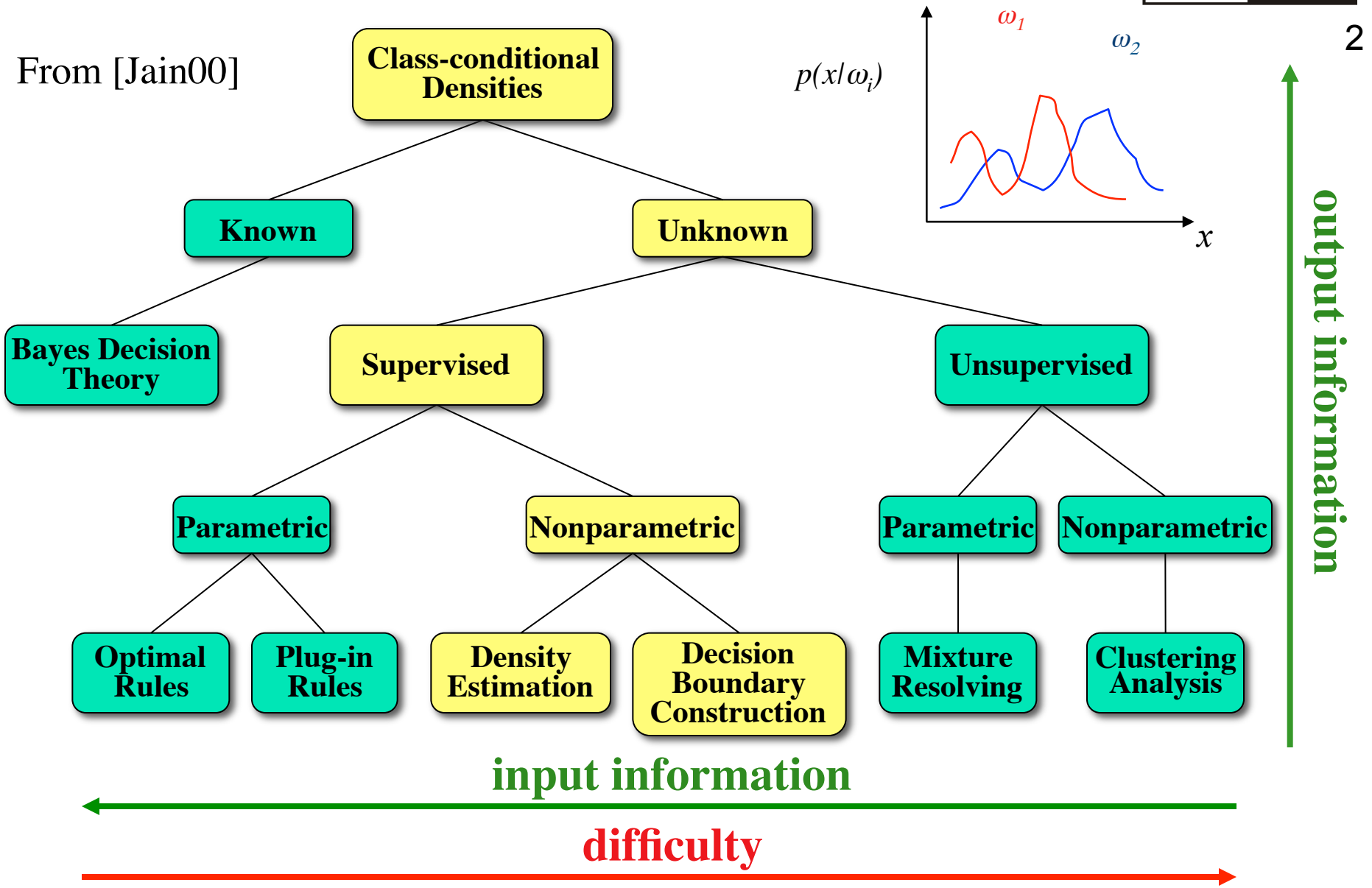
121 35 Praha 2, Karlovo nám. 13

hlavac@fel.cvut.cz



Statistické rozpoznávaní

From [Jain00]



Unimodální a vícemodální hustoty

- Parametrické metody umějí odhadovat unimodální hustoty pravděpodobnosti.
- Mnohé praktické úlohy odpovídají vícemodálním hustotám. Jen někdy (zřídka) lze vícemodální rozdělení modelovat jako směs unimodálních rozdělení.

- Neparametrické metody odhadu lze použít pro vícemodální hustoty, aniž by bylo nutné předpokládat tvar jejich rozdělení.
  něco za něco: potřebují více trénovacích dat

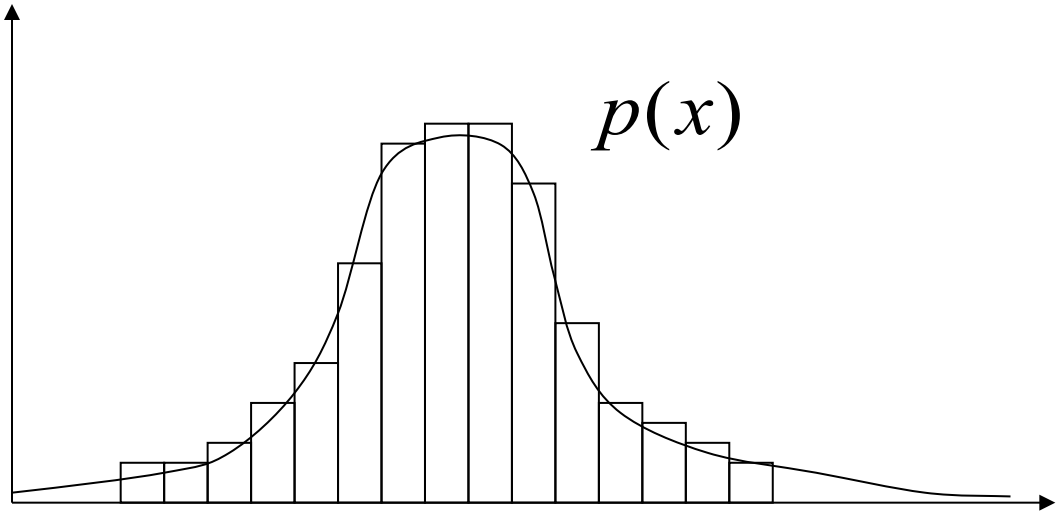
Neparametrické metody
Dva typy úloh

- Pozorování x, pravděpodobnost třídy (skrytého stavu) k z množiny tříd K.
- Zmíníme postupy pro odhad dvou pravděpodobností:
  - Hustoty pravděpodobnosti $p(x|k)$ závislé na třídě, (metoda histogramu, Parzenova okna).
  - Maximální aposteriorní pravděpodobnosti $P(k|x)$, (metody nejbližšího souseda, obcházejí odhad hustoty a odhadují přímo rozhodovací pravidlo).

Myšlenka = histogram

- Rozděl prostor jevů na přihrádky o šířce h
- Aproximuj rozdělení pomocí

$$\hat{p}(x) = \frac{1}{h} \frac{\text{počet bodů v přihrádce}}{\text{celkový počet bodů}}$$

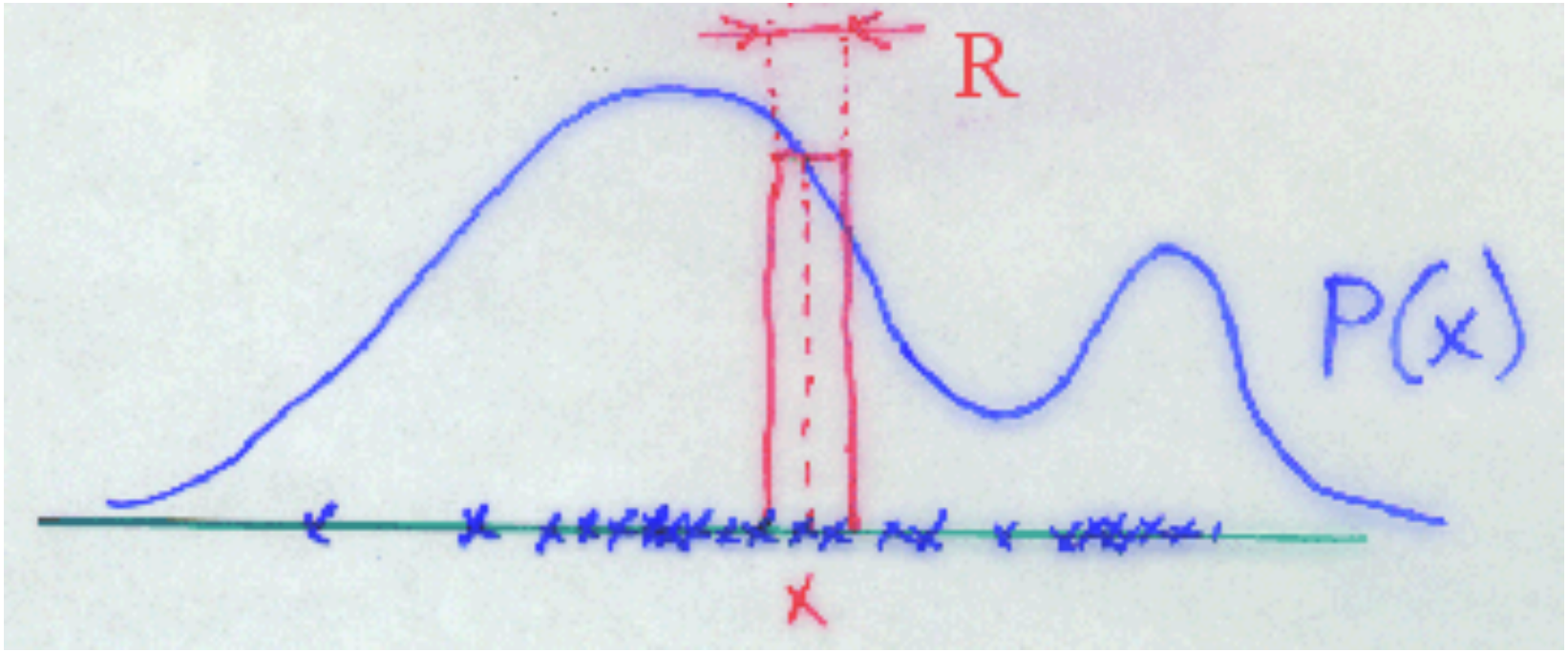


NEVÝHODY HISTOGRAMU

- Nespojitosti v odhadu hustoty závisí na kvantizaci přihrádek, nikoliv na hustotě.
- Prokletí dimenzionality:
  - Jemný popis vyžaduje mnoho přihrádek.
  - Počet přihrádek roste exponenciálně s počtem dimenzí.
  - Dat není dost, a tak je většina přihrádek prázdných.
- Tyto nevýhody činí metodu histogramu prakticky nepoužitelnou až na případ rychlé vizualizace dat v dimenzi 1 nebo 2.

Myšlenka neparametrických odhadů (1)

Trénovací množina $x = \{x_1, ..., x_n\}$



Myšlenka neparametrických odhadů (2)

Pravděpodobnost, že x padne do přihrádky o rozměru R

$$P = \int_R p(x')\ \mathrm{d}\ x'$$

Pravděpodobnost P je vyhlazenou verzí $p(x)$.

Obráceně, hodnotu $p(x)$ lze odhadnout z pravděpodobnosti P.

Myšlenka neparametrických odhadů (3)

Předpokládejme, že jsme vytáhli nezávisle m vzorků ze stejného rozdělení $p(x)$.

Pravděpodobnost, že m vzorků je z n je dána binomickým rozdělením

$$P_m = \left(\begin{array}{c} n \\ m \end{array} \right) P^m \left(1 - P\right)^{(n-m)}$$

Myšlenka neparametrických odhadů (4)

Očekávaná hodnota m rozdělení je

$$\mathcal{E}(m) = nP$$

Binomické rozdělení je velmi špičaté kolem své očekávané hodnoty, a proto lze očekávat, že m/n bude dobrým odhadem P, a tudíž i hustoty p.

Myšlenka neparametrických odhadů (5)

Objem přihrádky o šířce R

$$V = \int_R 1 \mathrm{d}n$$

Odhad pravděpodobnosti

$$\int_R p(n) \mathrm{d}n \approx p(x) \cdot V$$

Myšlenka neparametrických odhadů (6)

- m – počet x_i, které spadly do R

  n – počet přihrádek

  $$\int_R p(n)\ \mathrm{d}\, n \approx \frac{m}{n}$$

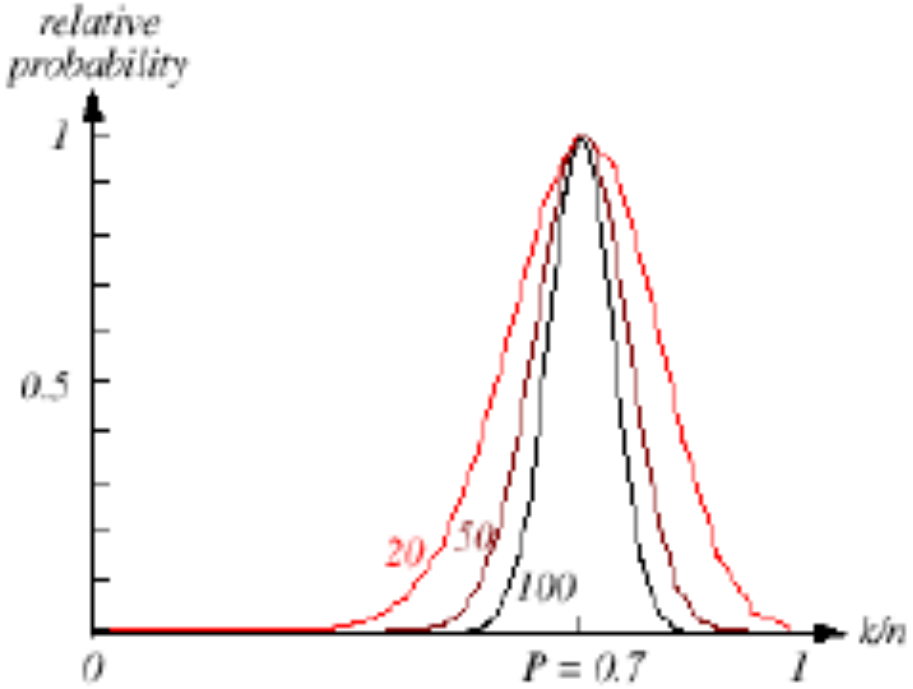
- Kombinací předchozích dvou vztahů získáme odhad pravděpodnosti

$$p(x) \approx \hat{p}(x) = \frac{m/n}{V}$$

ILUSTRACE

Skutečná hodnota hustoty rozdělení,
z něhož se v bodě x vybíralo je 0,7.

Normalizováno na stejnou hodnotu.



Praktická potíž

- Když zvolíme pevnou velikost přihrádky, potom m/n konverguje k vyhlazené hodnotě $p(x)$.
- Když zmenšujeme přihrádku do nekonečna, nepadne nám do ní žádný vzorek a náš odhad bude $p(x) \sim 0$.
- Musíme být připraveni, že prakticky náš odhad bude vždy vyhlazen.

Jak obejít problémy ?

- Potížím se teoreticky vyhneme, když budeme mít nekonečně vzorků rozdělení.
- Můžeme uvažovat posloupnost přihrádek různé velikosti kolem x. První přihrádka obsahuje 1 vzorek m=1, druhá m=2, atd.
- Odhad hodnoty rozdělení

$$p_n(x) = \frac{m_n/n}{V_n}$$

Tři podmínky

$$\lim_{n\to\infty} V_n = 0$$

$$\lim_{n\to\infty} m_n = 0\,,\; p(x) \neq 0$$

$$\lim_{n\to\infty} \frac{m_n}{n} = 0$$

Dva způsoby vytvoření posloupností přihrádek

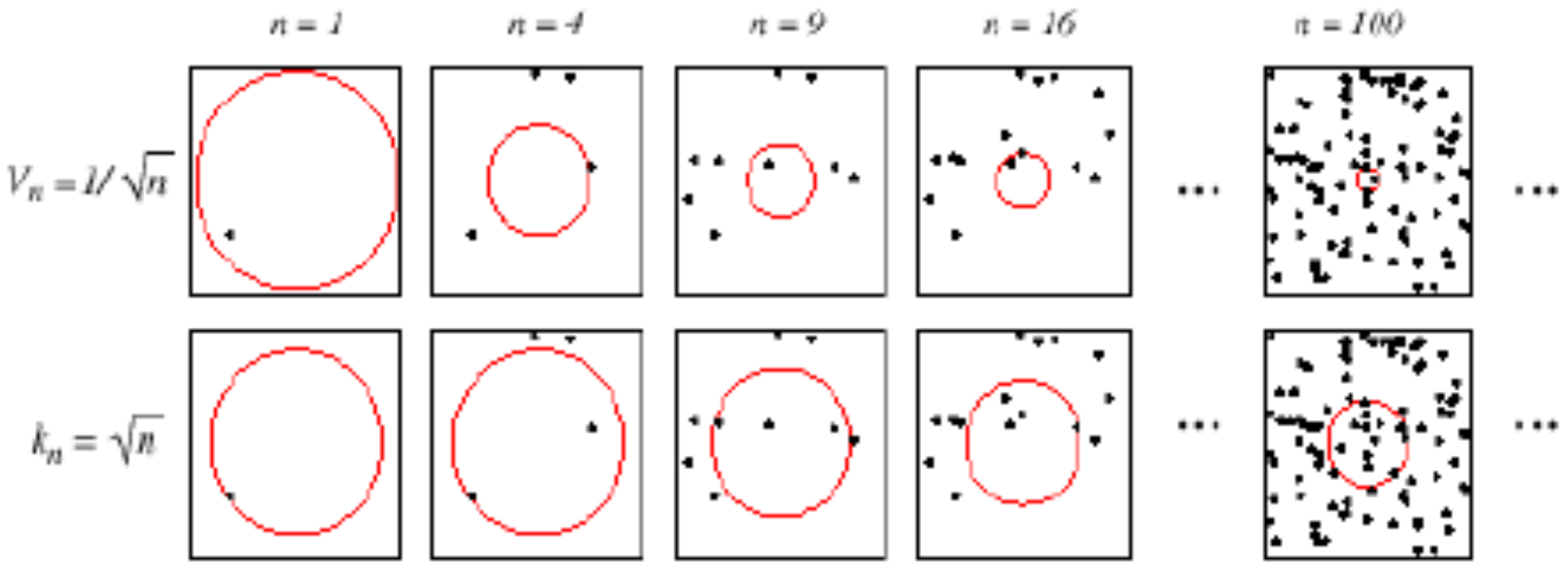
1. Objem přihrádky je funkcí n, např.

   $V_n = 1/\sqrt{n}.$

   Metoda Parzenova okna.

2. Počet vzorků mn je určen jako funkce n, např. $k_n = \sqrt{n}.$
   Zde přihrádka roste, až obsahuje mn sousedů vzorku x.
   Metoda mn-nejbližších sousedů.

ILUSTRACE růstu přihrádek

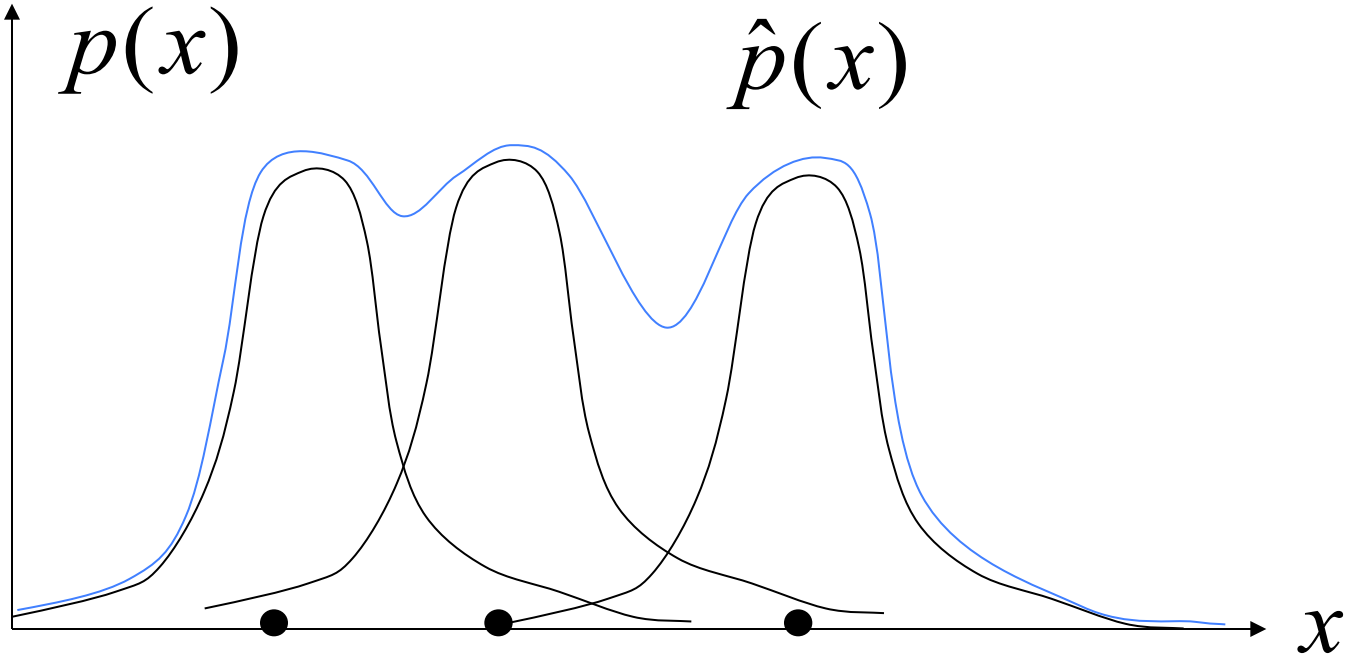


Metoda Parzenových oken

- Také nazývaná jádrová metody odhadu hustoty.

- Parzen E. (1962). *On estimation of a probability density function and mode*, Ann. Math. Stat. **33**, pp. 1065-1076.
- Duda R.O., Hart P.E., Stork D.G. (2001). *Pattern Classification.* John Willey & Sons, New York.

Neformálně Parzenova okna



Myšlenka: každý bod z trénovací množiny přispívá jednou Parzenovou jádrovou funkcí (nebo oknem) k vytvoření hustoty pravděpodobnosti.

Parzenovo okno

- Přihrádka je d-rozměrná nadkrychle o straně h se středem ve vzorku x přispívajícím odhadu.
- Počet vzorků v přihrádce je dán jádrovou funkcí

$$\varphi(u) = \begin{cases} 1 & |u_j| \leq 1; \quad j = 1,\ldots,d \\ 0 & \text{jindy,} \end{cases}$$

které se říká Parzenovo okno nebo naivní estimátor.

Odhad hustoty

- Počet vzorků v nadkrychli je

$$m = \sum_{j=1}^{n} \varphi\left(\frac{x - x_j}{h}\right)$$

- Odhad hustoty je

$$p_n(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{Vn} \varphi\left(\frac{x - x_j}{h_n}\right)$$

Odhad $p(x)$ jako suma δ-funkcí (1)

- Odhad je interpolací založené na oknové (jádrové) funkci $\varphi()$. Mimo přihrádku má δ-funkce nulovou hodnotu.
- Aby byl odhad pravděpodobností, musí platit

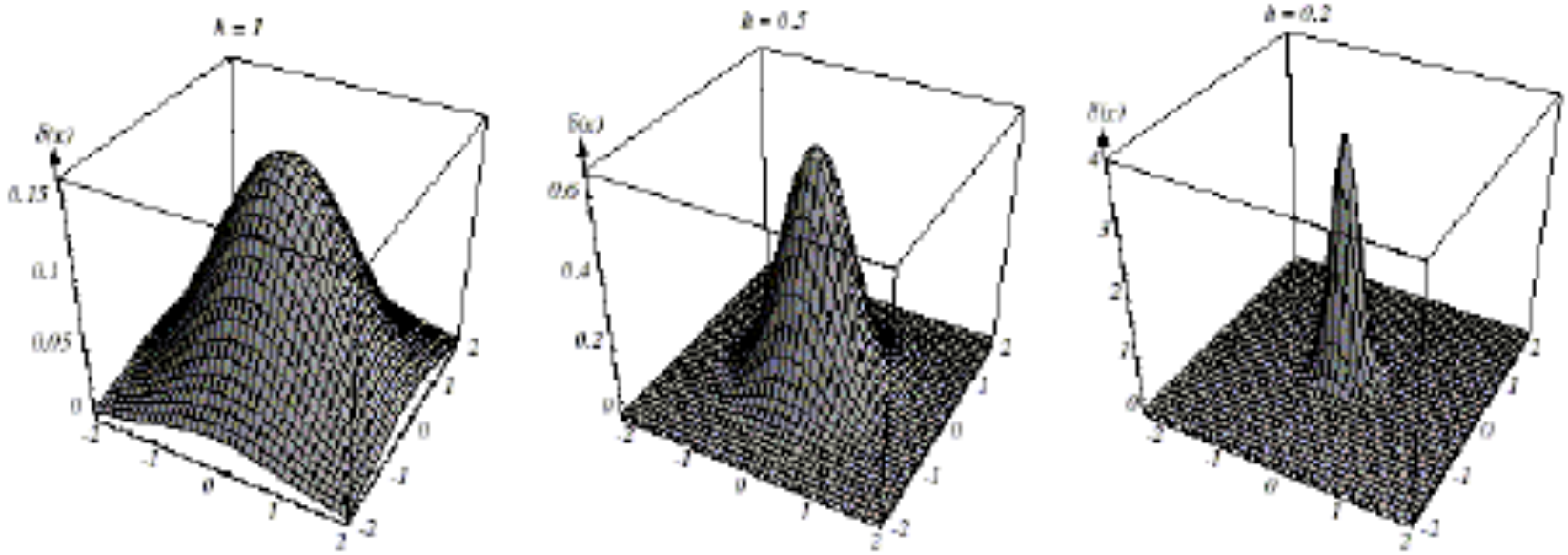
$$\varphi(x) \geq 0, \qquad \int \varphi(u)\,\mathrm{d}u = 1$$

- Zkoumejme vliv šířky okna h_n na $p_n(x)$. Uvažujme provizorně, že odhad je superpozicí Diracových pulsů δ

$$\delta_n(x) = \frac{1}{V_n}\varphi\left(\frac{x}{h_n}\right) \qquad p_n(x) = \frac{1}{n}\sum_{j=1}^{n}\delta_n(x - x_j)$$

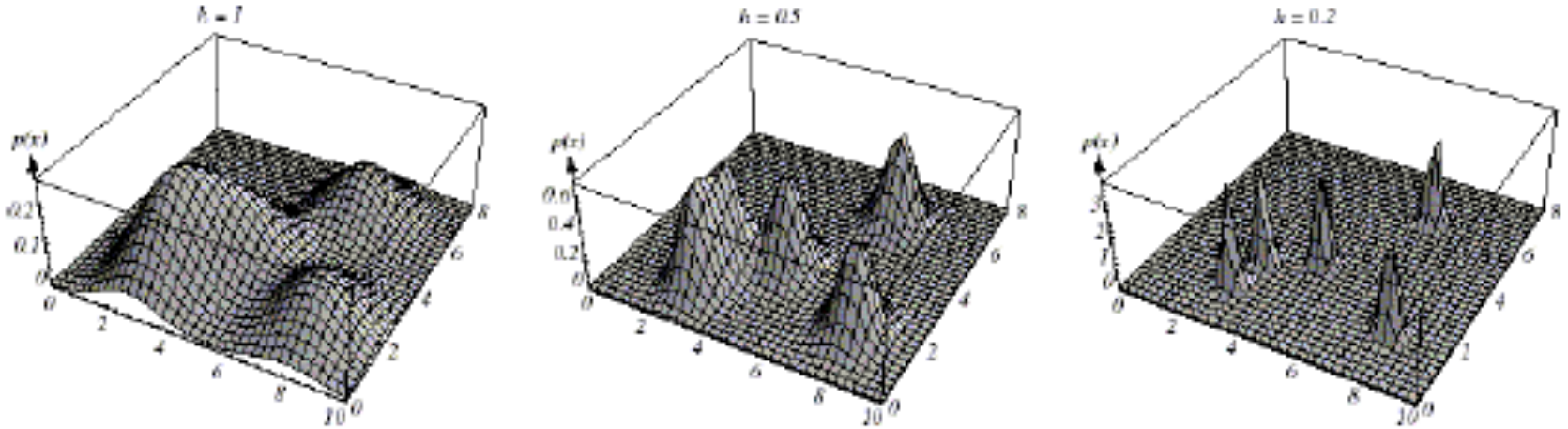
Odhad $p(x)$ jako suma δ-funkcí (2)

- Odhad $p(x)$ je sumou δ-funkcí.
- Rozměr přihrádky h ovlivňuje jak amplitudu tak i šířku $\delta(x)$, protože objem zůstává konstantní.



Ilustrace vlivu velikosti okna

- Při malém h je odhad $p(x)$ je superpozicí velmi pozvolně se měnících funkcí. Je „rozmazaný“.
- Při velkém h je odhad $p(x)$ superpozicí úzkých špiček v místě vzorků.



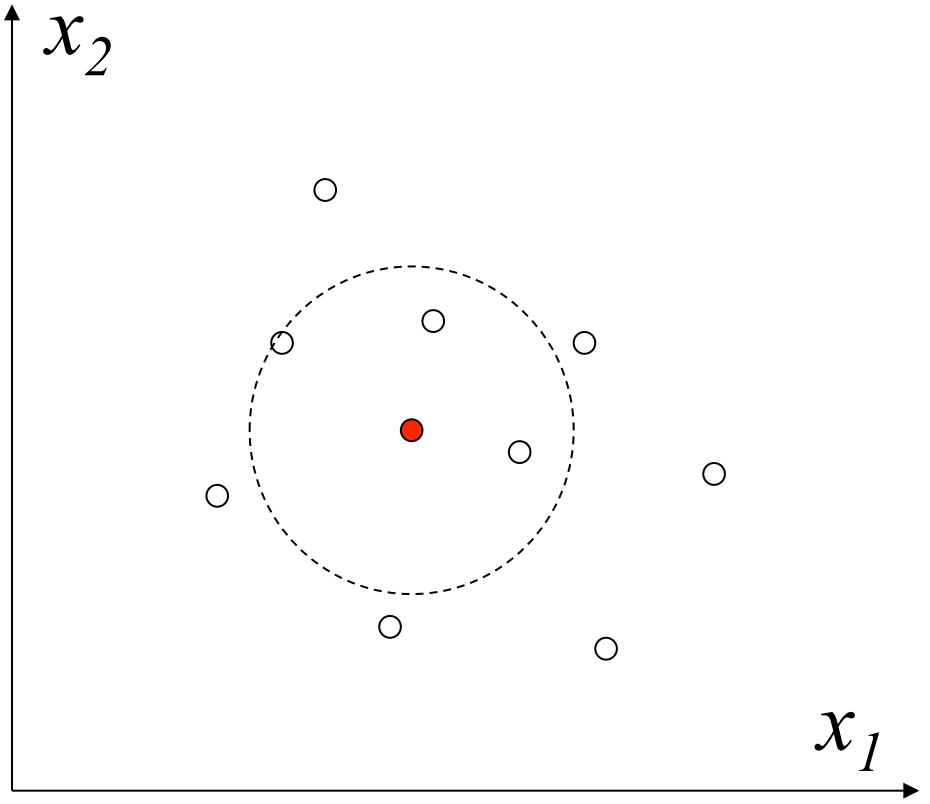
Volba jádrové funkce (okna)

- Vyhlazování je nutné. Superpozice Diracových pulsů by vedla k nespojitému odhadu $p(x)$.
- Doporučená volba: Gaussián

Odhad hustoty metodou nejbližšího souseda

- Najdi n nejbližších sousedů k hodnotě x.
- V_n je objem (např. koule) obsahujici těchto n vzorků.
- Odhadni hodnotu hustoty jako

$$\hat{p}(x) = \frac{1}{V}\frac{k}{N}$$

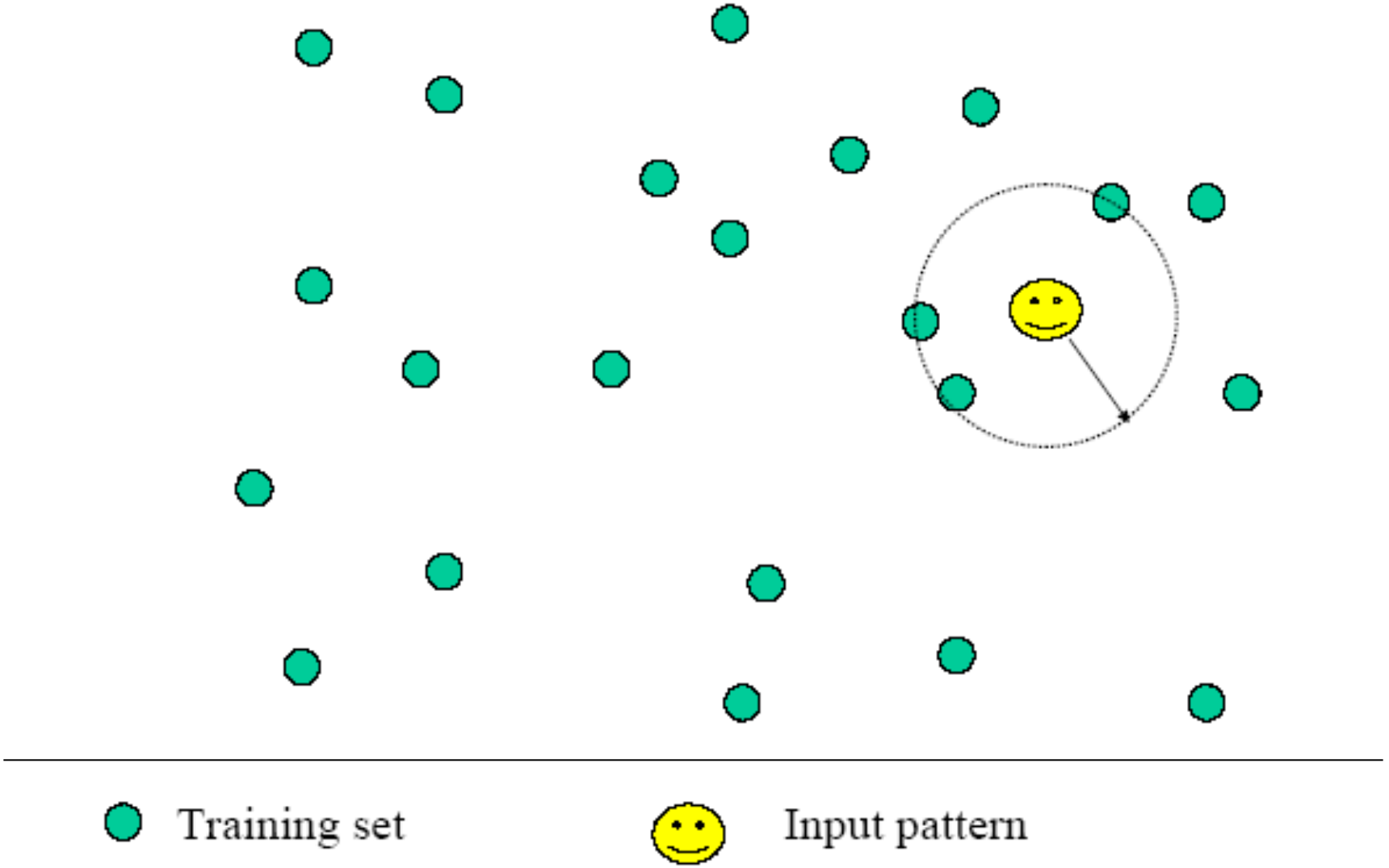


Nejbližší sousedé

Základní myšlenka je jednoduchá :

- Učení
  - Zapamatují se všechny dvojice {vstup, rozhodnutí o třídě}.
- Odhadování, rozpoznávání
  - Odpověď se vybere podle *n* nejbližších tréninkových příkladů.

Nejbližší sousedé, ilustrace



Nejbližší sousedé, vlastnosti

Výhody

- Po uchování trénovací množiny není potřebné další učení.
- Neparametrická metoda.

Nevýhody

- Potřebuje mnoho paměti pro uložení trénovací množiny.
- Pro mohutnější trénovací množiny je rozpoznávání pomalé.