パターン認識
Pattern Recognition

佐藤真一
Shin'ichi Satoh

国立情報学研究所
National Institute of Informatics

May 2, 2023

本講義について

- 講義のページ: https://research.nii.ac.jp/~satoh/utpr/
- 講義資料は上記のページにあり
- 講義映像は ITC-LMS にあり
- 単位は最終レポートと宿題によって評価予定 (最終レポートの提出は必須、宿題は全 7 回のうち 3 回の提出必須)
- 第一回 (4/18) の宿題の締め切りは本日、第二回 (4/25) の宿題の締め切りは来週、本日第三回 (5/2) の宿題の締め切りは再来週 (5/16)
- 出席は取らない

前回まで

- ここまでは “パラメトリック” 法についてみてきた
- 確率密度関数（あるいは等価的には識別境界）はパラメータであらわされる
- e.g., 正規分布の場合: 平均と分散 (あるいは共分散行列)
- これらの方法は、観測の従う確率分布が機知のパラメータ表現可能な形式であることを仮定している
- しかし、これが成り立たないケースも少なくない

本日の予定

- ノンパラメトリック確率密度推定法
  - パルゼンウィンドウ法
  - k-近傍法

ノンパラメトリック法

- 単純な例: ヒストグラム
- サンプルデータが既知とすると、ある特定ビンサイズ（ヒストグラム分割数）に基づきヒストグラムを構築可能
- ヒストグラムを確率密度関数として利用

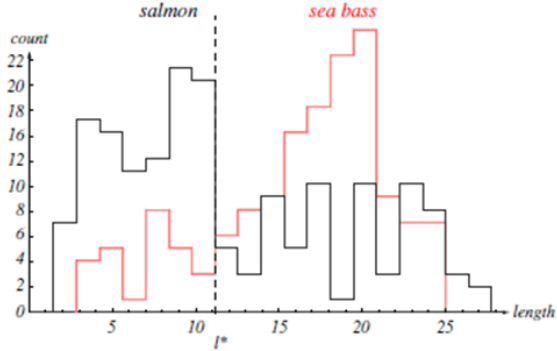
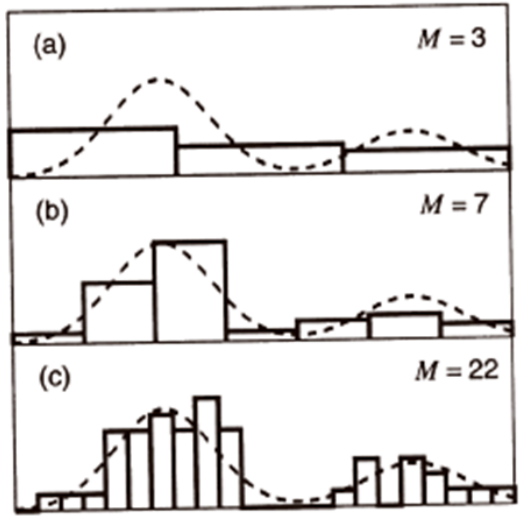


FIGURE 1.2. Histograms for the length feature for the two categories. No single threshold value of the length will serve to unambiguously discriminate between the two categories; using length alone, we will have some errors. The value marked l^* will lead to

ノンパラメトリック法

- 最適のビンサイズ M の決定法が問題
  - ビンの幅が小さい (M が大きい) と、推定された確率密度はジグザグしてしまう (雑音が多い)
  - ビンの幅が大きい (M が小さい) と、確率密度の細かい構造がなめされて見えなくなってしまう
- これら両者の間で最適な M を決定する必要がある
- 加えて、多次元データの分布推定の場合はどのように拡張するか？



ノンパラメトリック分布推定

- 確率変数 $\mathbf{x}$ が未知の確率密度 $p(\mathbf{x})$ に従うとき、$\mathbf{x}$ がある領域 R の中に入る確率:

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

- $p(\mathbf{x})$ から独立に取り出した n 点 $\{x_1, x_2, \ldots, x_n\}$ があるとき、ちょうど k 点が R に入る確率は二項分布に従う:

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

ノンパラメトリック分布推定

- k の期待値:

$$E\{k\} = nP$$

- R に入る点の占める割合の期待値:

$$E\{\frac{k}{n}\} = P$$

- その確率の分散

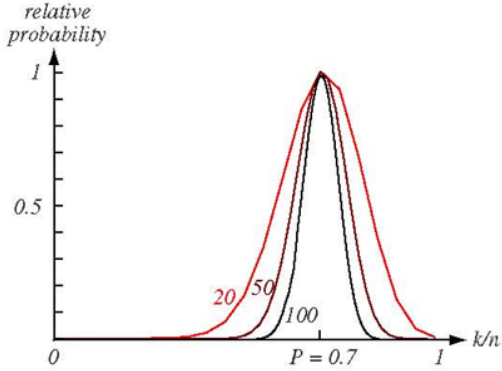
$$Var\{\frac{k}{n}\} = E\{(\frac{k}{n} - P)^2\} = \frac{P(1-P)}{n}$$

ノンパラメトリック分布推定

$n \to \inf$ の時、分布はシャープに:

$$P \approx \frac{k}{n}$$

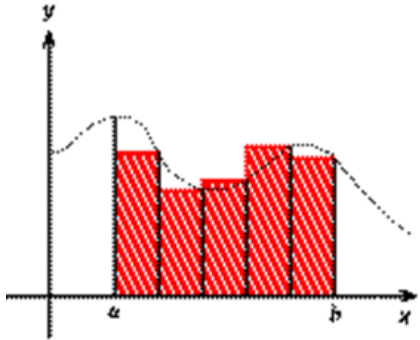
→ 近似 1



$p(\mathbf{x})$ が連続で、R の中であまり変化しないと仮定すると、P は次のように近似できる:

$$P = \int_R p(\mathbf{x}')d\mathbf{x}' \approx p(\mathbf{x})V$$

$\rightarrow$ 近似 2
ここで V は R で囲まれる体積



ノンパラメトリック分布推定

- これらの近似をあわせると:

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

- この近似は以下の相反する仮定に基づく:
  - R は中に十分多数のサンプルを含む必要があるため比較的大きい: 近似 1
  - R は比較的小さく、$p(\mathbf{x})$ はその領域内でほぼ一定: 近似 2
- 実際には最適の R をどのように決定すればよいか

ノンパラメトリック分布推定

- $\mathbf{x}$ を含む一連の領域を考える: $R_1, R_2, \ldots$
  - R_1 は k_1 個の点を含み, R_2 は k_2 個の点を含み, 等.
  - それぞれ n 個のサンプルについて考えた場合を想定
- R_i の体積は V_i で k_i 個の点を含む
- n 番目の領域の確率 $p(\mathbf{x})$ の推定値 $p_n(\mathbf{x})$ は:

$$p_n(\mathbf{x}) \approx \frac{k_n/n}{V_n}$$

以下の相反する条件を、$p_n(\mathbf{x})$ が $p(\mathbf{x})$ に収束するように決定する:

$$\lim_{n\to\infty} V_n = 0 \quad \text{近似 1}$$

$$\lim_{n\to\infty} k_n = \infty \quad \text{近似 2}$$

$$\lim_{n\to\infty} \frac{k_n}{n} = 0 \quad p_n(\mathbf{x}) \text{ の収束のため}$$

ノンパラメトリック分布推定

$$p_n(\mathbf{x}) \approx \frac{k_n/n}{V_n}$$

最適の V_n と k_n の決定法として、二つの流儀がある:

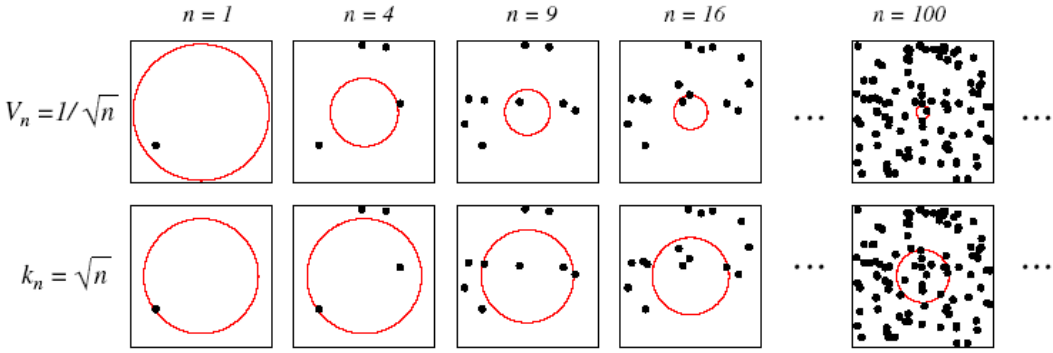
(1) 体積 V_n を固定し、データに基づき k_n を決定 (カーネル分布推定法), e.g.,

$$V_n = \frac{1}{\sqrt{n}}$$

(2) k_n を固定し、対応する体積 V_n をデータから推定 (k 近傍法), e.g.,

$$k_n = \sqrt{n}$$

ノンパラメトリック分布推定



パルゼンウィンドウ

$$p_n(\mathbf{x}) \approx \frac{k_n/n}{V_n}$$

- 問題: ベクトル $\mathbf{x}$ が与えられたとき $p(\mathbf{x})$ を推定
- R_n を一辺 h_n の超立方体とし、その中心を $\mathbf{x}$ とすると:

$$V_n = {h_n}^d$$

- k_n を決定するため (=超立方体内の点の数) カーネル関数を決定:

$$\varphi(u) = \begin{cases} 1 & |u_j| \leq \frac{1}{2} \ (j = 1, \ldots, d) \\ 0 & \text{otherwise} \end{cases}$$

パルゼンウィンドウ

- 超立方体内の点 x_i の数 :

$$k_n = \sum_{i=1}^{n} \varphi(\frac{\mathbf{x} - x_i}{h_n})$$

- 以下の推定値:

$$p_n(\mathbf{x}) \approx \frac{k_n/n}{V_n}$$

は以下のようになる

$$p_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{V_n} \varphi(\frac{\mathbf{x} - x_i}{h_n})$$

$\rightarrow$ パルゼンウィンドウ推定法

パルゼンウィンドウ

- 確率分布の推定は、カーネル関数とサンプル x_i の重ね合わせとなる:

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi(\frac{\mathbf{x}-x_i}{h_n})$$

- $\varphi(u)$ はサンプル間の確率分布を補間する
- 各サンプル x_i は x からの距離に基づき確率分布推定に貢献

パルゼンウィンドウ

- カーネル関数 $\varphi(u)$ はより一般的な形式をとりうる (超立方体以外)
- $p_n(\mathbf{x})$ が適正な分布となるため、$\varphi(u)$ も有効な確率分布となる必要がある。したがって:

$$\varphi(u) \geq 0$$

$$\int \varphi(u) du = 1$$

パルゼンウィンドウ

パラメータ h_n は、分布の推定をスムーズにするために働く

- h_n が大きいと、推定分布はスムーズになりすぎる (あまりに「広い」カーネル関数の重ねあわせ)
- h_n があまりに小さいと、推定結果は真の分布というよりも元のデータそのもの (あまりに「狭い」カーネル関数の重ねあわせ)

パルゼンウィンドウ

異なる h_n によるカーネル関数: $\varphi(u)$

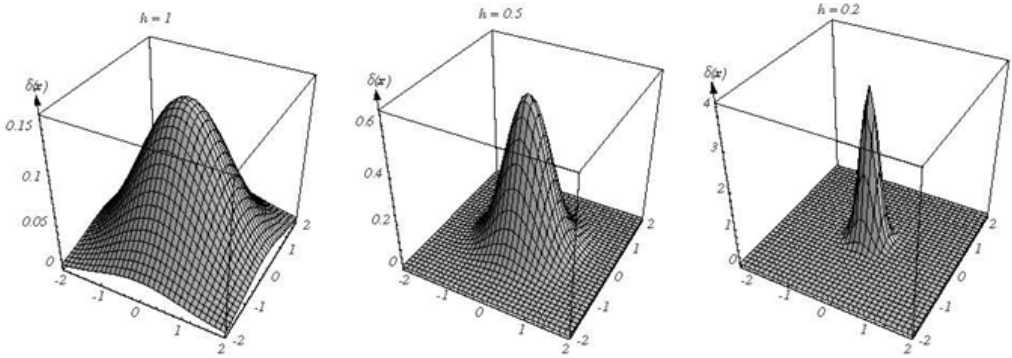


FIGURE 4.3. Examples of two-dimensional circularly symmetric normal Parzen windows for three different values of h. Note that because the $\delta(\mathbf{x})$ are normalized, different vertical scales must be used to show their structure. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

例: 5 サンプル点による $p_n(\mathbf{x})$:

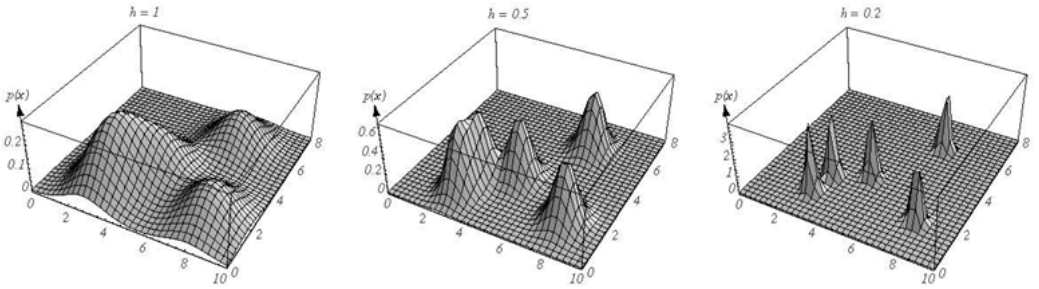
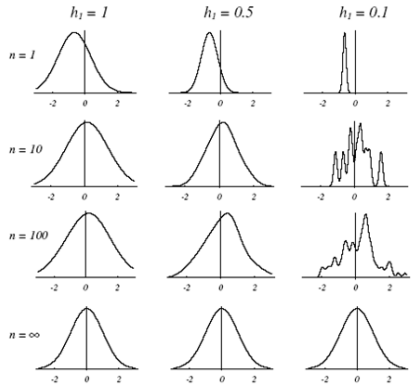


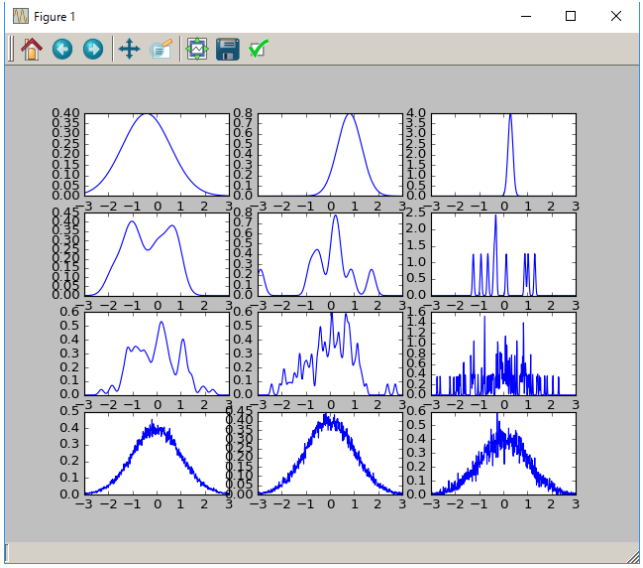
FIGURE 4.4. Three Parzen-window density estimates based on the same set of five samples, using the window functions in Fig. 4.3. As before, the vertical axes have been scaled to show the structure of each distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

例: 真の分布 $p(\mathbf{x})$ とカーネル関数 $\varphi(u)$ の両方とも正規分布の場合

$$h_n = \frac{h_1}{\sqrt{n}}$$

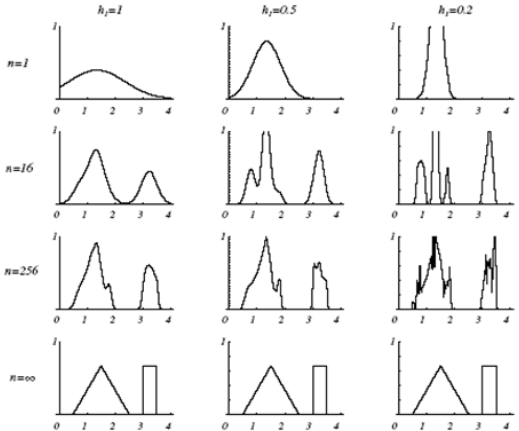


演習 (parzeng.py)

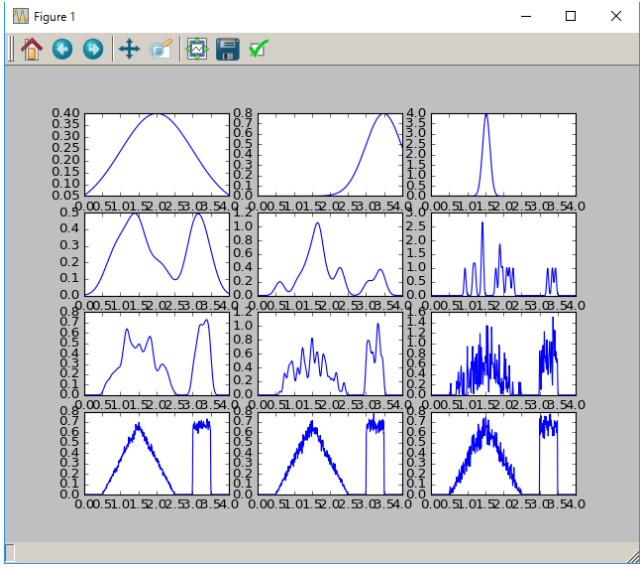


例: 真の分布 $p(\mathbf{x})$ は一様分布と三角形の組み合わせで、カーネル関数 $\varphi(u)$ は正規分布の場合

$$h_n = \frac{h_1}{\sqrt{n}}$$



演習 (parzentr.py)



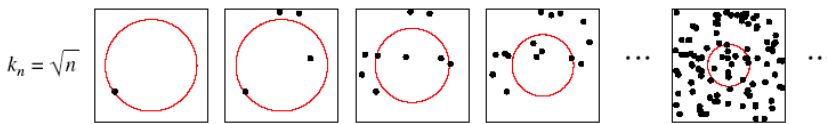
k-近傍推定法

k_n を固定し、V_n を変化させる

- $\mathbf{x}$ を中心とした超球を考える
- 超球の半径を、ちょうど k_n 個の点を含むまで広げる
- V_n は超球の体積

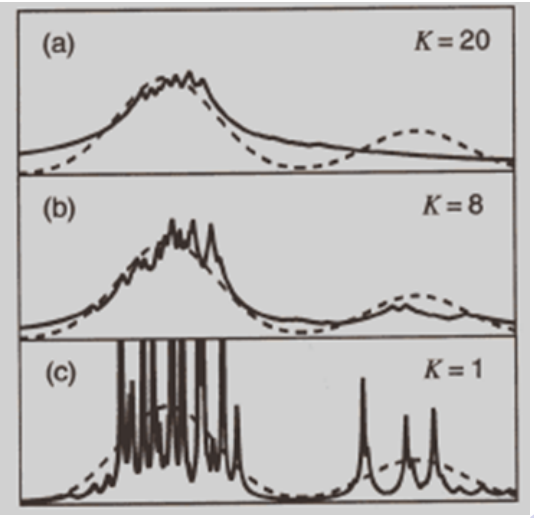
$$p_n(\mathbf{x}) \approx \frac{k_n/n}{V_n}$$

超球のサイズは密度に依存



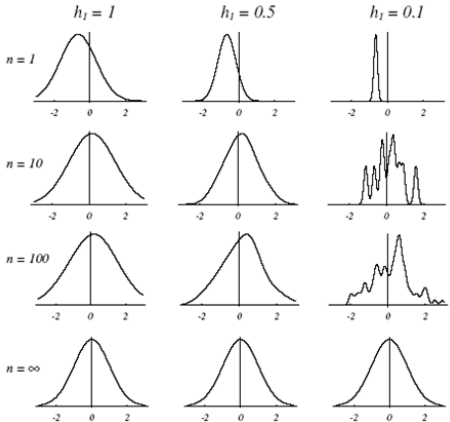
k-近傍推定法

k_n が分布のスムーズさを決定する

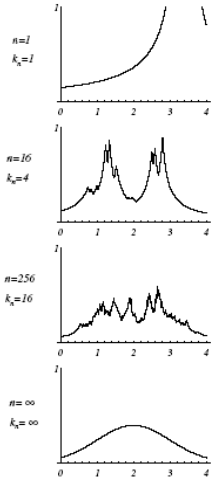


k-近傍推定法

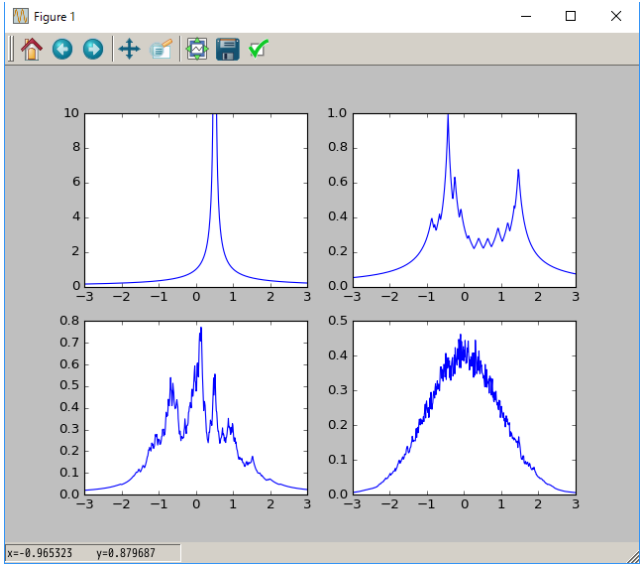
パルゼンウィンドウ



k-近傍法

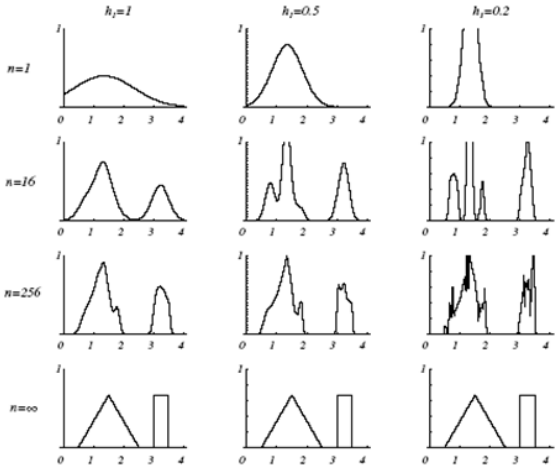


演習 (knng.py)

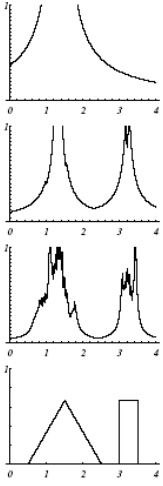


k-近傍推定法

パルゼンウィンドウ



k-近傍法



演習 (knntr.py)

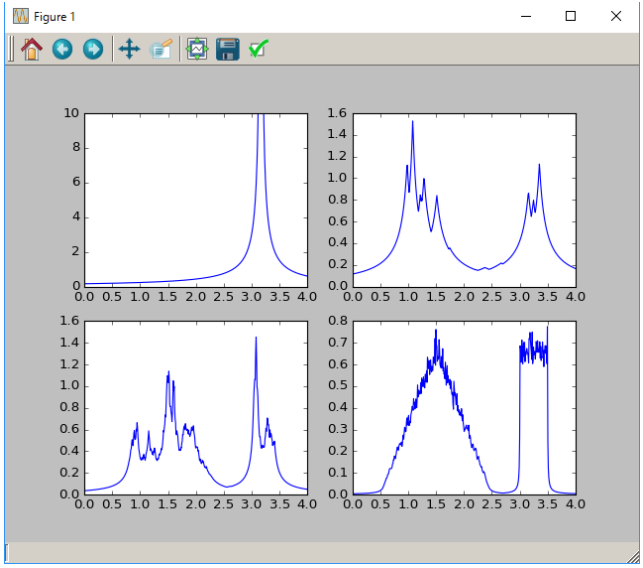


Figure 1
10
8
6
4
2
0
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0

宿題

- プログラミング課題と非プログラミング課題を課する
- プログラミング課題か非プログラミング課題かのいずれかを解いて提出すること
- プログラミング課題を解くことを推奨する
- もちろん両方解いてもらえると嬉しい
- 締切は 5/16

宿題 1 (プログラミング課題)

- 講義ページからデータファイルをダウンロードせよ。x1 と x2 という二つの変数を含む
- (データは Matlab フォーマットで与えられている。読み込み方は 4/13 の講義資料参照)
- これらは二つのクラス c_1 と c_2 から得た標本と仮定せよ
- 条件付確率密度 $p(\mathbf{x}|c_i)$ を、パルゼンウィンドウ法 (カーネル関数として正規分布と超立方体両方) と k-近傍法 (様々な k) で求め、図示せよ
- 事後確率 $P(c_i|x)$ を図示せよ。ただし事前確率は $P(c_1) = P(c_2) = \frac{1}{2}$ とせよ

宿題 2 (非プログラミング課題)

- 確率変数 $\mathbf{x}$ が未知の確率密度 $p(\mathbf{x})$ に従うとき、$\mathbf{x}$ がある領域 R の中に入る確率を P とする
- $p(\mathbf{x})$ から独立に取り出した n 点 $\{x_1, x_2, \ldots, x_n\}$ があるとき、ちょうど k 点が R に入るとする
- k の期待値が以下であることを示せ

$$E\{k\} = nP$$

- R に入る点の占める割合の期待値が以下であることを示せ

$$E\{\frac{k}{n}\} = P$$

- その確率の分散が以下であることを示せ

$$Var\{\frac{k}{n}\} = E\{(\frac{k}{n} - P)^2\} = \frac{P(1-P)}{n}$$