宿題 1

2 つのクラス c_1, c_2 からそれぞれ得た標本 x_1, x_2 について，条件付確率密度 $p(x|c_i)$ 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 について）で求め，図示する。また，その事後確率 $p(c_i|x)$ も図示する。

図 1，2 にそれぞれ x_1, x_2 のヒストグラムを示す。

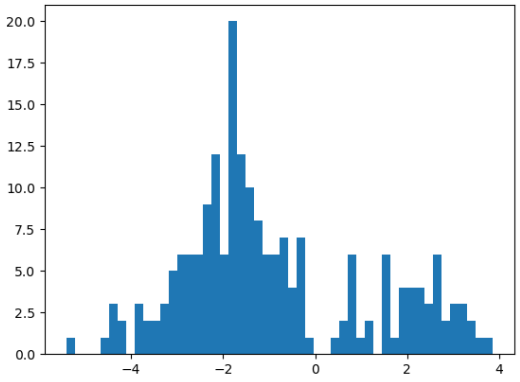

図 1: x_1 のヒストグラム

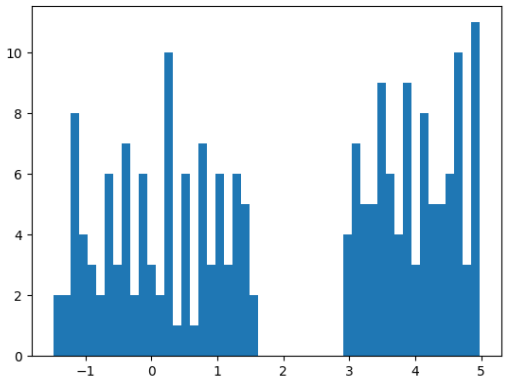

図 2: x_2 のヒストグラム

プログラム

プログラムの本体は**??**ページの Listing **??**に示した。以下に簡単なプログラムの説明を示す。

- load
  .mat ファイルからデータを読み出し，x_1, x_2 のデータを返す。
- plot_data
  x_1, x_2 それぞれのヒストグラムをプロットする。

- normal_distribution
  正規分布の確率密度を返す。parzen 法におけるカーネル関数として用いる。
- hypercube
  超立方体の確率密度を返す。parzen 法におけるカーネル関数として用いる。
- conditional_probability_parzen
  parzen 法における条件付確率を求める。
- conditiona_probability_kmeans
  k 近傍法における条件付確率を求める。距離関数には絶対値を用いる。
- _nonparametric_method
  parzen 法と k 近傍法で x_1, x_2 の条件付確率と事後確率を求めてプロットするときの，共通する部分をまとめた関数。
- parzen
  parzen 法で x_1, x_2 の条件付確率と事後確率を求めてプロットする。
- kmeans
  k 近傍法で x_1, x_2 の条件付確率と事後確率を求めてプロットする。
- main
  上記の関数をまとめて実行する関数。

結果

parzen 法にて正規分布をカーネル関数として用いたときの結果を図 3 に，超立方体をカーネル関数として用いたときの結果を図 4 に，k 近傍法の結果を図 5 に示す。バンド幅 h とクラスタに含まれるサンプル数 k は，図に示した値について実験した。

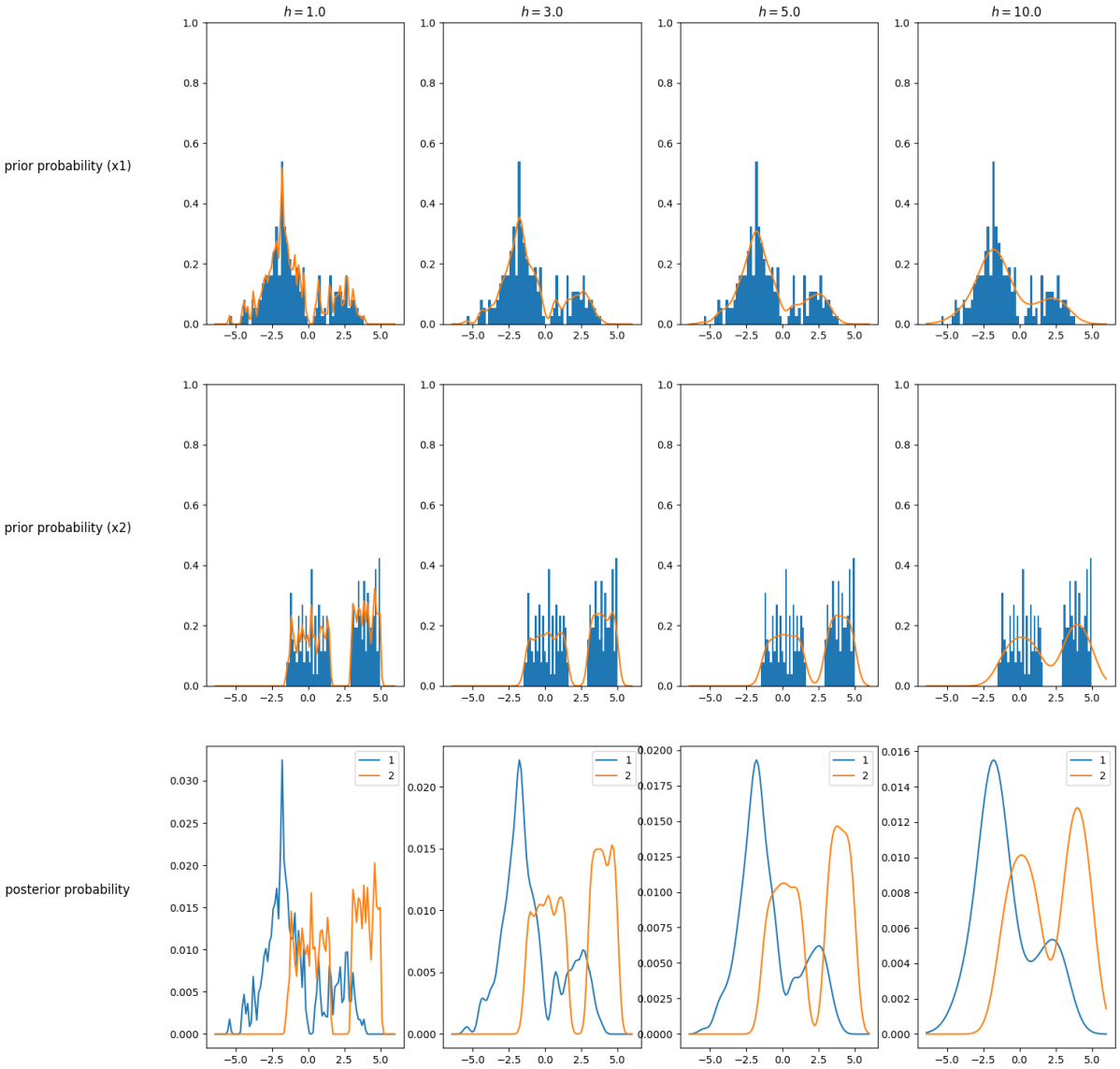

図 3: parzen 法にて正規分布をカーネル関数として用いたときの結果

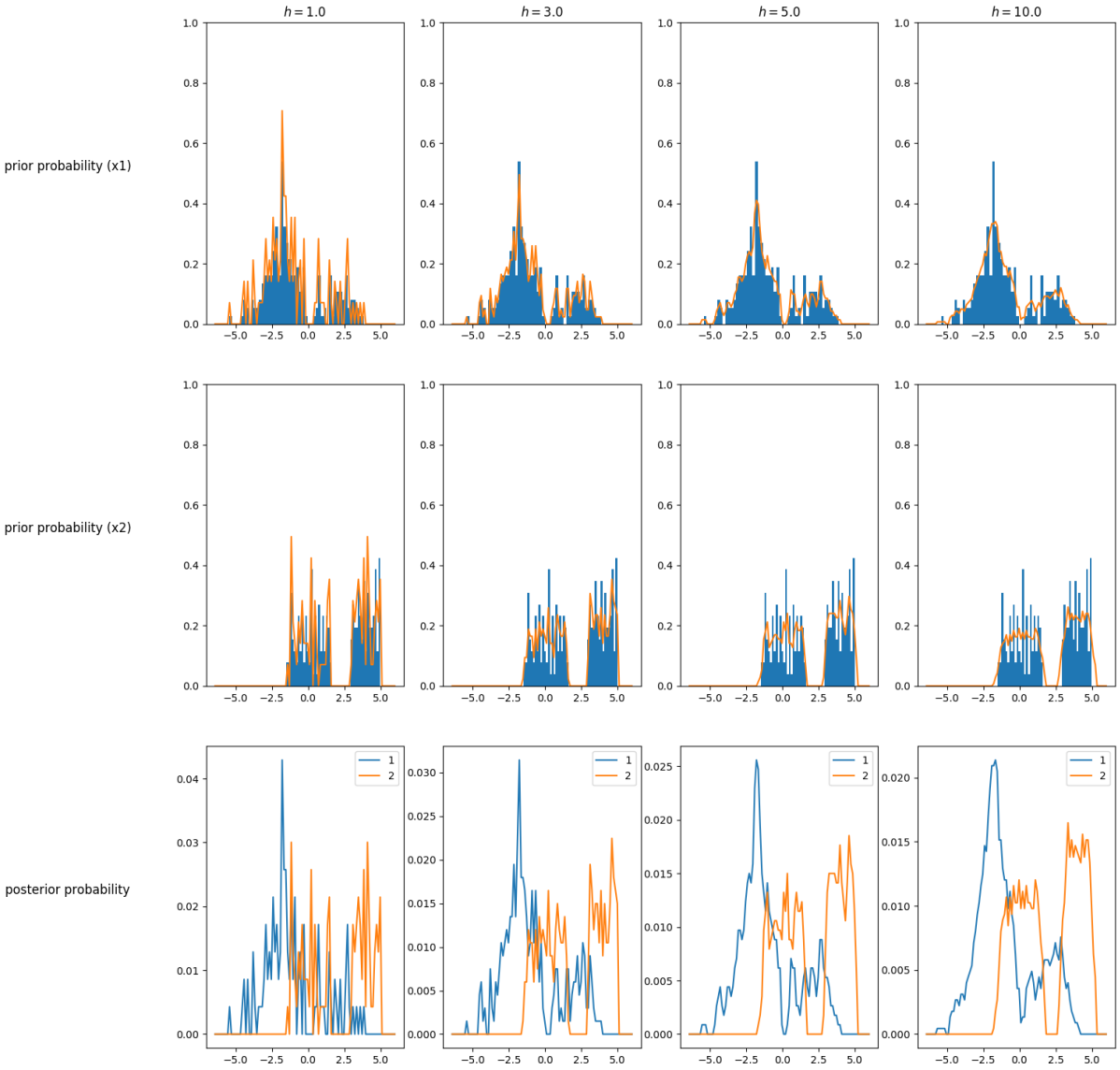

図 4: parzen 法にて超立方体をカーネル関数として用いたときの結果

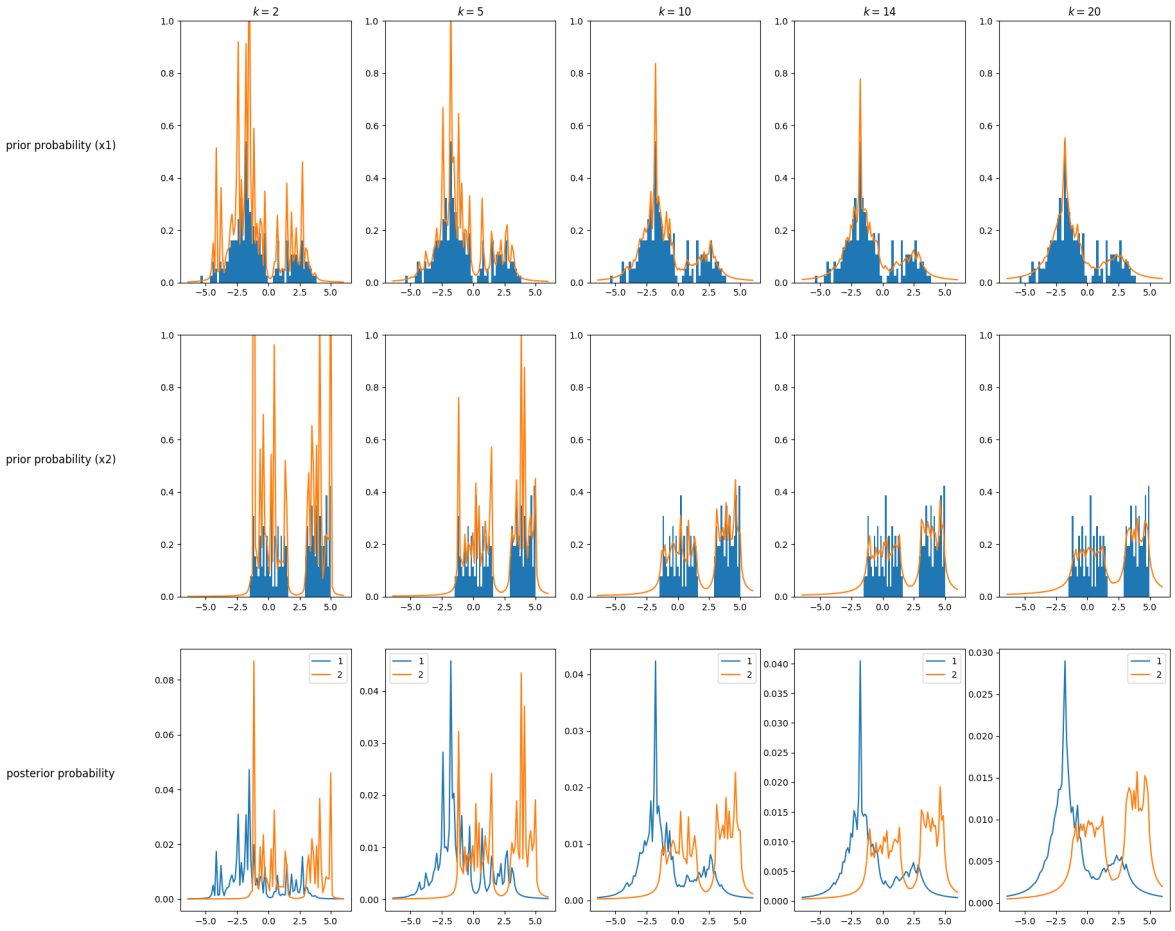

図 5: k 近傍法の結果

考察

parzen 法において正規分布をカーネル関数として用いた場合は，バンド幅 $h = 3$ でもかなり滑らかな確率密度関数が得られていたが，超立方体を用いた場合は $h = 10$ でもあまり滑らかにはならなかった。これは，正規分布の領域は分散の影響で x に対し滑らかにしか変化しないのに対し，超立方体はその領域の内と外をきっぱりと分けてしまう，つまり x に対し不連続に変化するため，当然ともいえる。実際，x_2 のような，不連続（に見えるよう）な分布について，バンド幅 3 のものを見れば，正規分布ではデータが存在しない部分でも確率密度関数の値が大きくなってしまっているが，超立方体ではそのような部分はほとんど見られない。逆に，x_1 のような，x に対し滑らかに変化しているような分布については，正規分布はうまく真の分布を近似しているように見えるが，超立方体ではノイズの影響が大きく出てしまっている。

k 近傍法については，k が小さいときは値が発散してしまっていて，ある程度大きくすると（$k \geq 10$），真の分布に近づけている。