データ解析
第十回「k-近傍法」

鈴木　大慈
理学部情報科学科
西八号館 W707 号室
s-taiji@is.titech.ac.jp

休講情報

7/7 は休講

今日の講義内容

- k-近傍法による密度推定
- k-近傍法による判別

構成

k-近傍法による確率密度推定の目的

ノンパラメトリック密度推定法の一種
カーネル密度推定と類似

k-近傍密度推定法の長所・短所

- （長所） 計算が簡単．k-近傍への距離を計算すればよい．
- （長所） 漸近的な性質が理論的に導出可能．
- （長所） 分布に対する仮定が少ない．滑らかさくらい．
- （短所） 推定した密度が積分して 1 になるとは限らない．
- （短所） 推定した密度関数は不連続になりえる．
- （短所） ノイズに弱い．
- （短所） 推定された分布は裾が重い．

構成

k-近傍密度推定量

$\{X_i\}_{i=1}^n$: データ (d 次元とする)
k-近傍密度推定量 (k-nearest neighbor)

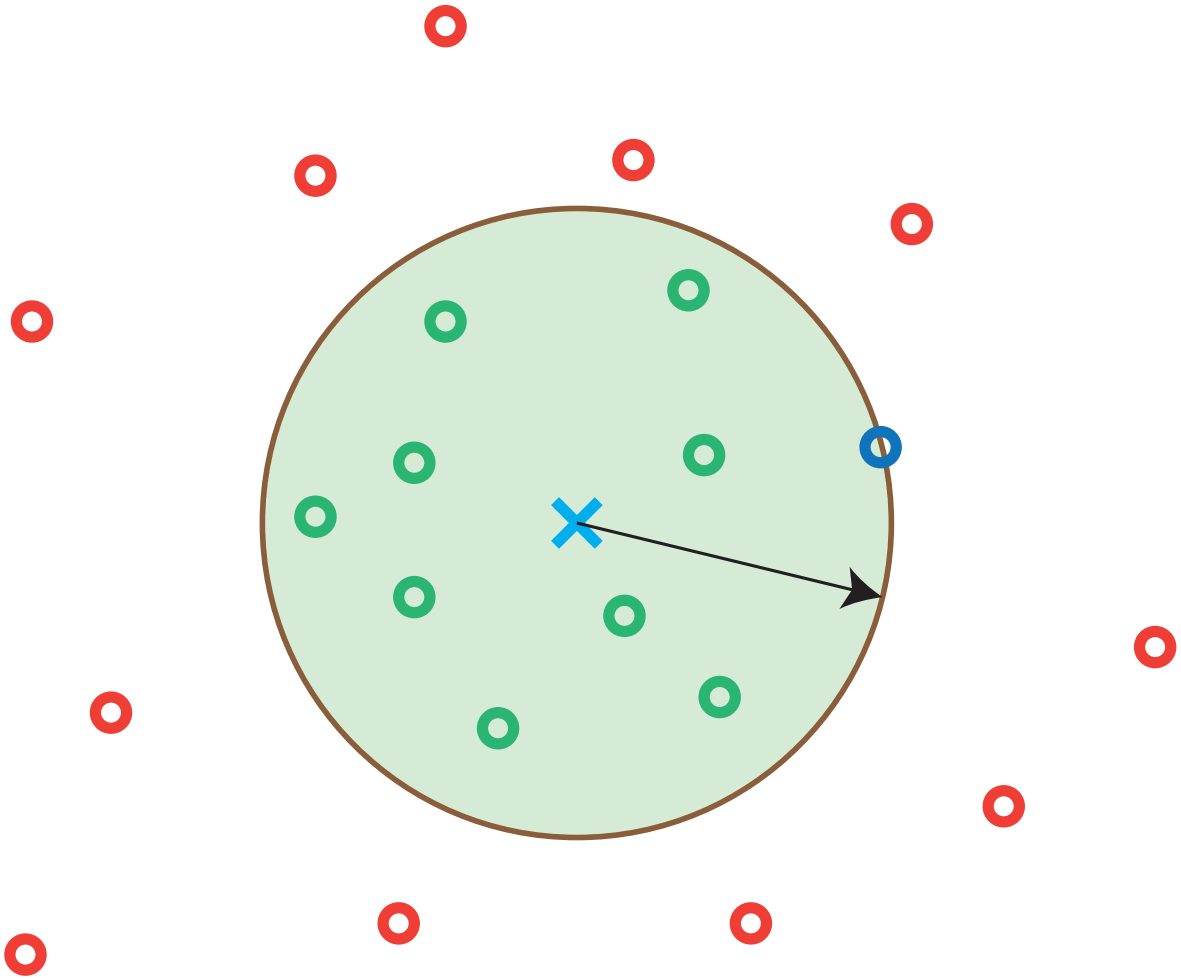
1. $k \geq 1$ を一つ固定.
2. 点 x に一番近い k 個のサンプル点 $\{X_{(1)}, X_{(2)}, \ldots, X_{(k)}\}$ を持ってくる. (距離が近い順に並んでいるとする. $X_{(k)}$ が一番遠い)
3. 密度を次のように推定:

$$\hat{p}(x) = \frac{k}{nV_d \|x - X_{(k)}\|^d}.$$

ただし, V_d は d 次元超球 ($\{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$) の体積. $V_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$.

x のまわりに多くのサンプルがあれば $\hat{p}(x)$ は大きく, 逆に少なければ $\hat{p}(x)$ は小さい.

k-近傍



$k = 10$ での k-近傍

k-近傍密度推定法の簡単な導出

x のまわりの半径 R の領域を $\mathcal{S}$ とする:

$$\mathcal{S} := \{x' \in \mathbb{R}^d \mid \|x - x'\| \leq R\}.$$

$\mathcal{S}$ に点が含まれる確率 P は以下のように与えられる:

$$P = \int_{\mathcal{S}} p(x')\mathrm{d}x'.$$

すると，n サンプルのうち k 個が $\mathcal{S}$ に落ちる確率は

$$P(|\{X_i \in \mathcal{S}\}| = k) = \binom{n}{k} P^k (1-P)^{n-k}$$

である．二項分布の平均と分散より

$$\mathrm{E}\left[\frac{k}{n}\right] = P, \quad \mathrm{Var}\left[\frac{k}{n}\right] = \frac{P(1-P)}{n}.$$

よって，サンプルサイズ n が十分大きければ,

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

である．ただし，k は領域 $\mathcal{S}$ に含まれるサンプルの数である．

一方，R が十分小さければ，

$$P = \int_{\mathcal{S}} p(x')\mathrm{d}x' \simeq V(\mathcal{S})p(x)$$

である (テイラー展開).
これらを合わせて，

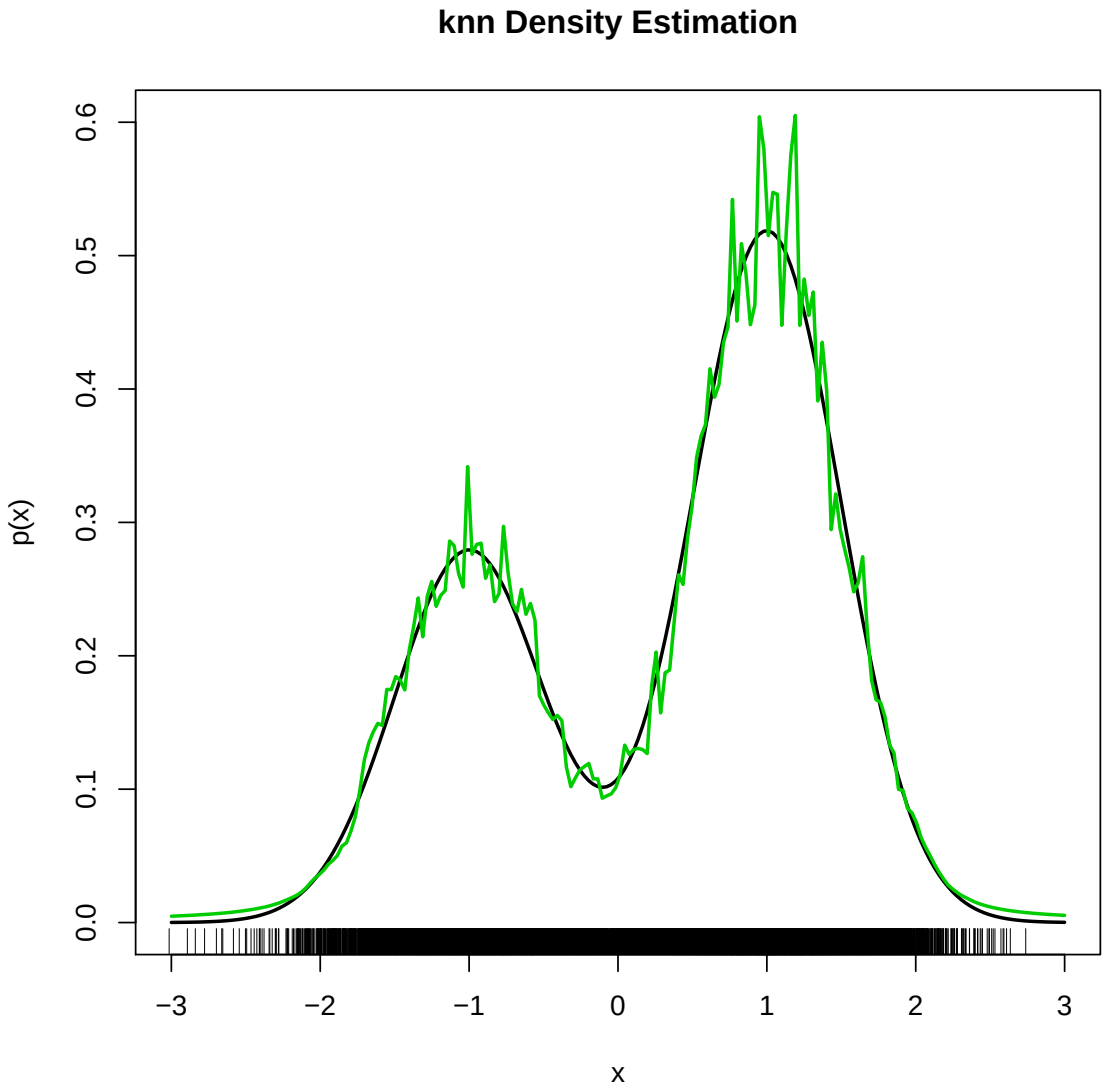
$$V(\mathcal{S})p(x) \simeq \frac{k}{n} \quad \Rightarrow \quad p(x) \simeq \frac{k}{nV(\mathcal{S})}.$$

特に，R として $\|x - X_{(k)}\|$ としたものが k-近傍法である．V_d を単位球の体積とすると，

$$V(\mathcal{S}) = V_d R^d$$

なので，密度推定量を得る．

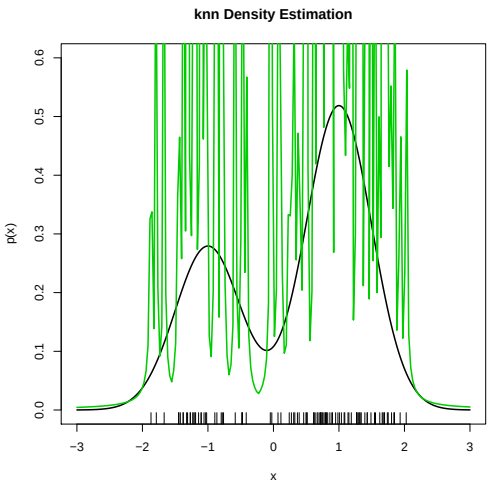
k-近傍法の様子



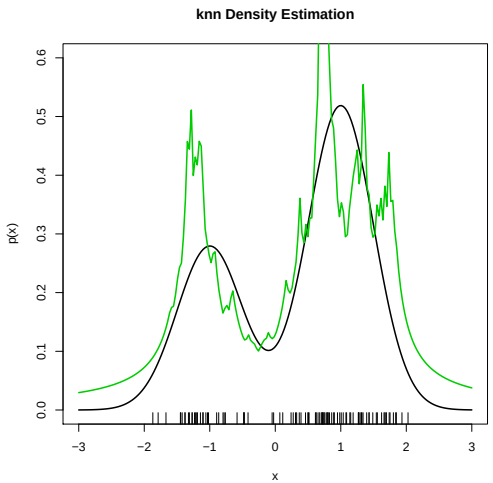
$k = 100,\ n = 10000$

k-近傍法の k と推定結果

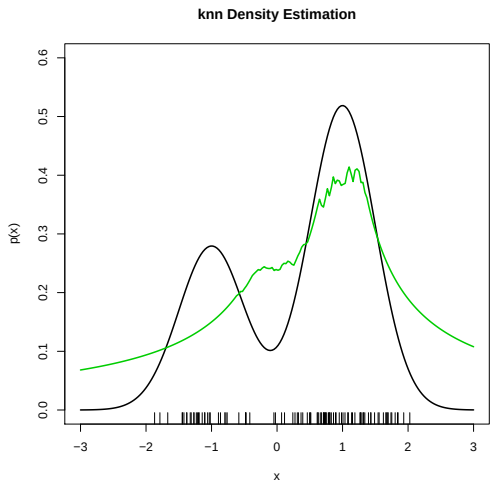
$n = 100$



$k = 1$



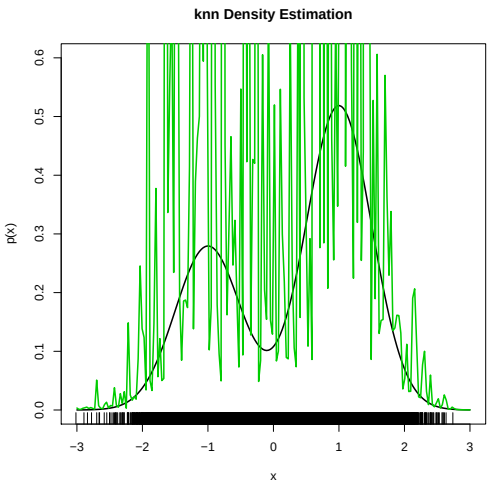
$k = 10$



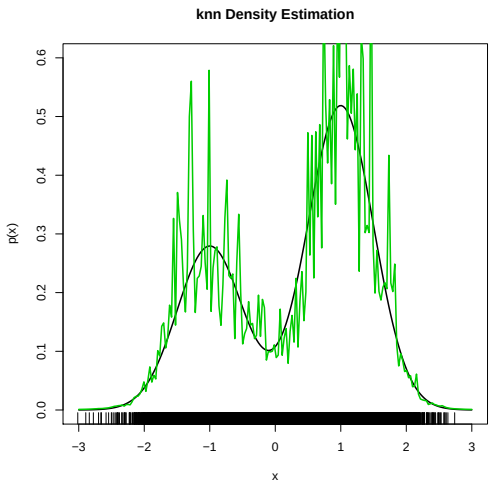
$k = 50$

k-近傍法の k と推定結果

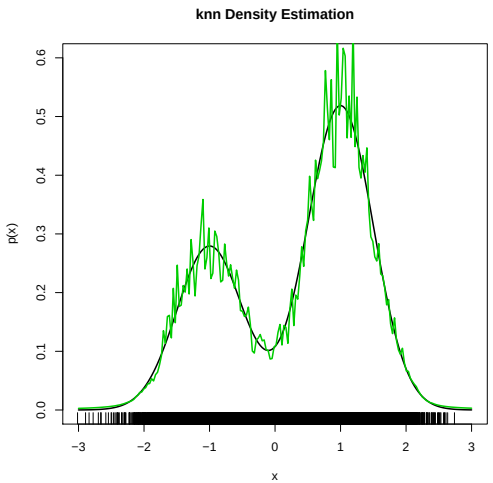
$n = 10000$



$k = 1$



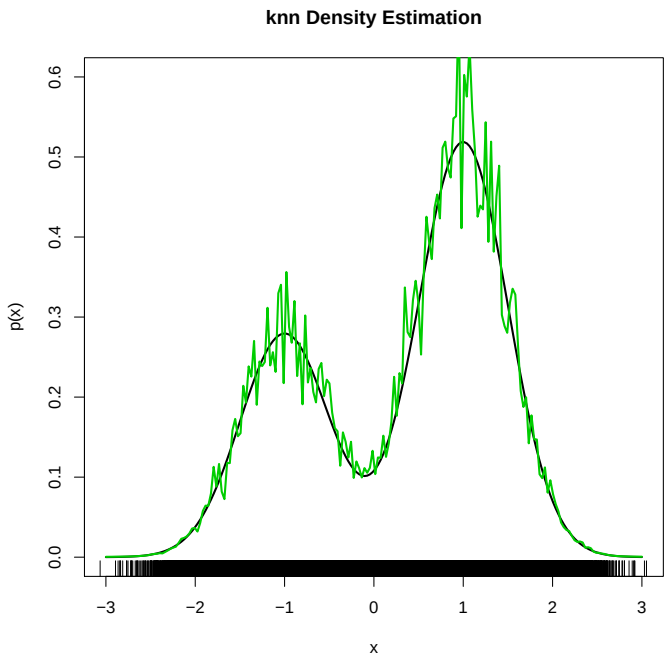
$k = 10$

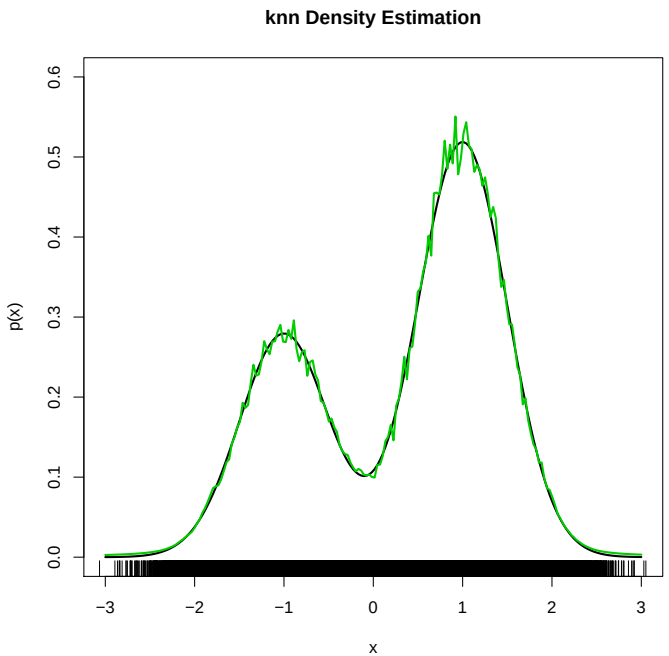


$k = 50$

k-近傍法の k と推定結果

$n = 100000$



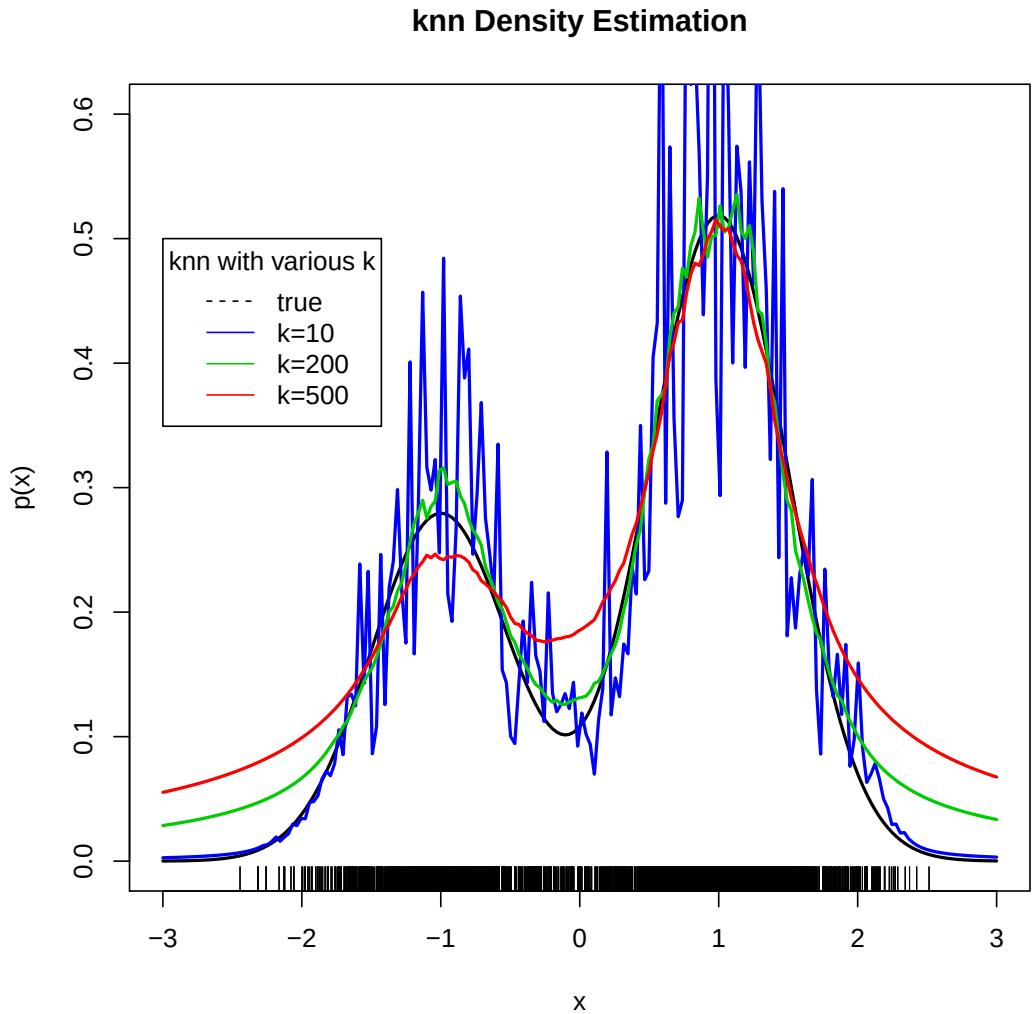


$k = 50$ $k = 500$

サンプルサイズが大きくなるにつれ，k も大きくしたほうがよい．

構成

k-近傍密度推定と k の関係



カーネル密度推定のカーネルと同様に，k は適切に選ぶ必要がある．
クロスバリデーションを使えばよい．

構成

平均二乗誤差

$k \geq 1$ に対する，k-近傍密度推定量を

$$\hat{p}(x) = \frac{k}{nV_d \|x - X_{(k)}\|^d}$$

とする．

- 推定量 $\hat{p}$ の点 x における**平均二乗誤差 (Mean Squared Error)** :

$$\mathrm{MSE}(\hat{p}(x), k) := \mathrm{E}[(\hat{p}(x) - p(x))^2].$$

ただし，期待値 $\mathrm{E}[\cdot]$ はサンプル $\{X_i\}_{i=1}^n$ の出方についてとる．

Q: k に対して平均二乗誤差はどのように振る舞うか？

平均二乗誤差の漸近的振る舞い

$n \to \infty$ に伴い，$k \to \infty$ とする．その時，

$$\mathrm{MSE}(\hat{p}(x), k) = \frac{p^2(x)}{k} + \frac{c^2(x)}{p^{4/d}(x)} \left(\frac{k}{n}\right)^{4/d} + o\left(\frac{1}{k} + \left(\frac{k}{n}\right)^{4/d}\right),$$

ただし，

$$c(x) = \frac{1}{2(d+2)\pi} \Gamma^{2/d}\left(\frac{d+2}{2}\right) \mathrm{Tr}[\nabla \nabla^\top f(x)].$$

証明は省略．例えば Biau et al. (2011) を参照．
これより最適な k は

$$k^* = \left\lceil \frac{f^{2+4/d}(x) d}{4c^2(x)} n^{4/(d+4)} \right\rceil \propto n^{4/(d+4)}$$

で，このとき

$$\mathrm{MSE}(\hat{p}(x), k^*) = O(n^{-\frac{4}{d+4}}).$$

カーネル密度推定と同じオーダー (二回微分可能な pdf 推定の最適オーダー)．

次元の呪い

カーネル密度推定と同様に

$$\mathrm{MSE}(\hat{p}(x), k^*) = O\left(n^{-\frac{4}{d+4}}\right)$$

は，d が大きくなると収束レートが遅くなる．
「次元の呪い」はノンパラメトリック推定において必ず現れる．

構成

k-近傍法による判別

- 密度推定はそこまで安定していない．
- カーネル密度推定を用いたほうが良さそう．
- k-近傍法は，むしろ判別でよく用いられている．
- 機械学習などで用いられる最も初等的な判別器の一つ．
- へたに凝った方法を用いるより精度が出たりする．

k-近傍法による判別

サンプル：$\{(X_i, Y_i)\}_{i=1}^n \in \mathbb{R}^d \times \{1, \ldots, Q\}$
ただし，$Y_i \in \{1, 2, \ldots, Q\}$．つまり Q 個のカテゴリへの分類．

やりたいこと：サンプルから新しい点 x を Q 個のカテゴリへ判別したい．
（例：手書き文字認識）

$n_1, n_2, \ldots, n_Q$: カテゴリ q のサンプルの数とする $(\sum_{q=1}^Q n_q = n)$．

1. 点 x に最も近い k 個のサンプルを取ってくる．それを $\{(X_{(1)}, Y_{(1)}), (X_{(2)}, Y_{(2)}), \ldots, (X_{(k)}, Y_{(k)})\}$ とする．
2. この k 個の内，最も数が多かったカテゴリへ x を判別: k_q を k 個のサンプルのうち，カテゴリ q であるものの数として $(\sum_{q=1}^L k_q = k)$，

$$\hat{q}(x) = \mathrm{argmax}_{q=1,\ldots,Q} k_q.$$

判別の解釈

直観：近くに沢山あるカテゴリがもっとも確からしいだろう．

点 x がカテゴリ q である確率を $P(q|x)$ とする．ベイズの定理より，

$$P(q|x) = \frac{p(x|q)P(q)}{\sum_{q=1}^{Q} p(x|q)P(q)} = \frac{p(x|q)P(q)}{p(x)}$$

である．ただし，

- $p(x|q)$ はカテゴリ q に属するサンプルの確率密度関数．
- $p(x)$ は x の周辺確率密度関数．
- $P(q)$ はカテゴリ q が出る確率．

- $p(x)$ は k-近傍法で推定可能.
- $P(q)$ もサンプルから推定可能 (多項分布の最尤推定):

$$\hat{P}(q) = \frac{n_q}{n}.$$

- $p(x|q)$ も k-近傍法で推定可能 :

$$\hat{p}(x|q) = \frac{k_q}{n_q V_d \|X_{(k)} - x\|^d}$$

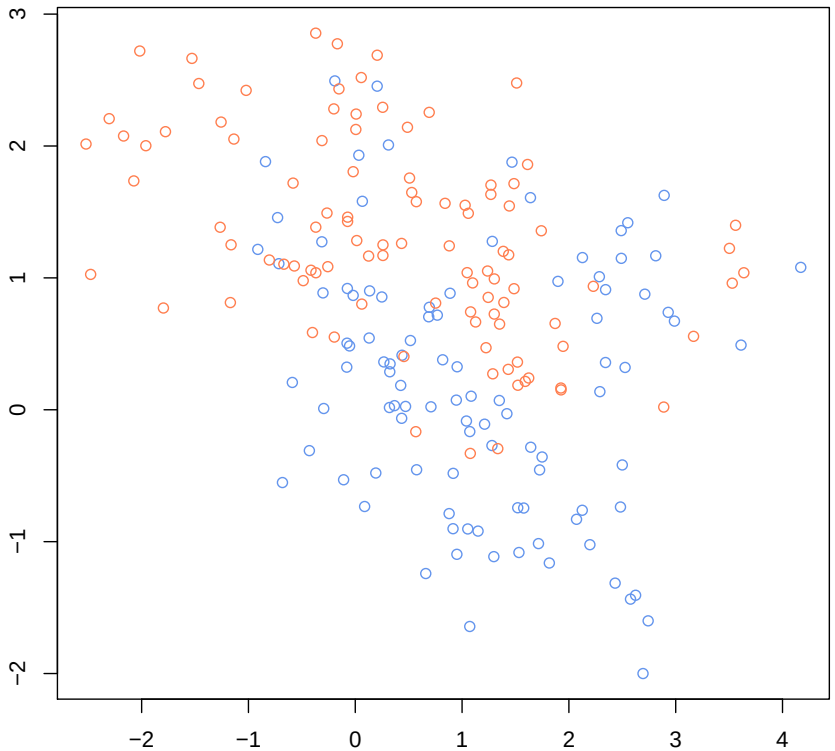
(半径 $\|X_{(k)} - x\|$ の球に k_q 個入っている).

以上をベイズの定理に代入すると,

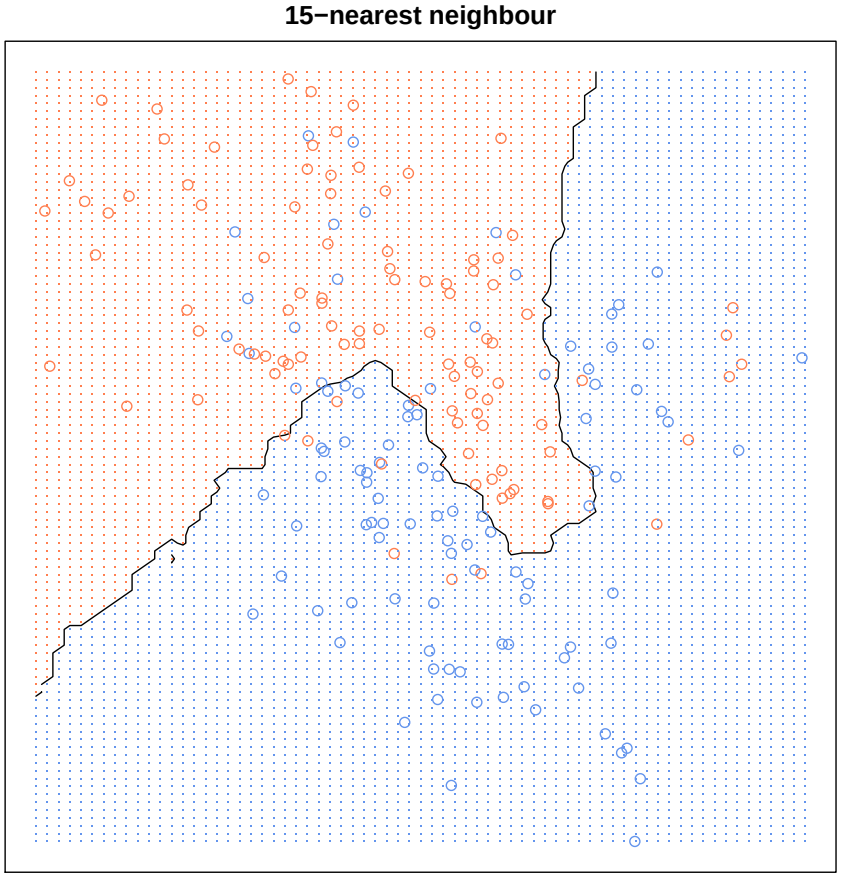
$$\begin{aligned}
\hat{P}(q|x) &= \frac{\hat{p}(x|q)\hat{P}(q)}{\hat{p}(x)} \\
&= \frac{\frac{k_q}{n_q V_d \|X_{(k)} - x\|^d} \times \frac{n_q}{n}}{\frac{k}{n V_d \|X_{(k)} - x\|^d}} \\
&= \frac{k_q}{k}.
\end{aligned}$$

を得る. これの最大値が $\hat{q}(x)$ である.

k-近傍法による判別の結果



k-近傍法による判別の結果



判別誤差のクロスバリデーション

Cross Validation

判別誤差を最小にする k 選びたい.

J-fold クロスバリデーション

1. サンプルを J 個に等分割する.
2. J 分割の j 番目のグループを $\mathcal{I}_j$ とおく: $\{(X_i, Y_i)\}_{i \in \mathcal{I}_j}$ $(j = 1, \ldots, J)$.
3. $\hat{f}_{(j)} : \mathbb{R}^d \to \{1, 2, \ldots, Q\}$ を j 番目のグループ $\mathcal{I}_j$ を 除いて 推定された判別器とする.
4. 判別誤差の推定量 $\widehat{E}(k)$ を計算:

$$\widehat{E}(k) = \sum_{j=1}^{J} \left(\frac{1}{|\mathcal{I}_j|} \sum_{i \in \mathcal{I}_j} \mathbf{1}[\hat{f}_{(j)}(X_i) \neq Y_i] \right).$$

5. $\widehat{E}(k)$ を最小にする k を採用.

データのスケーリング

d 次元データを用いる場合，座標ごとのスケールが大きく違う場合，推定結果が悪くなる場合がある．
講義で紹介した k-近傍法は等方的にユークリッドノルムによる球を用いているためである．

解決法

- 座標ごとの分散を揃える (座標ごと正規化する):

$$X_{i,j} \leftarrow X_{i,j} / \mathrm{std}(X_{:,j}).$$

- 座標ごとに幅が異なる楕円や長方形を使う．その際，$V_d \|x - X_{(k)}\|^d$ の代わりに楕円や長方形の体積を用いる．

構成

k-近傍法を行える R 関数

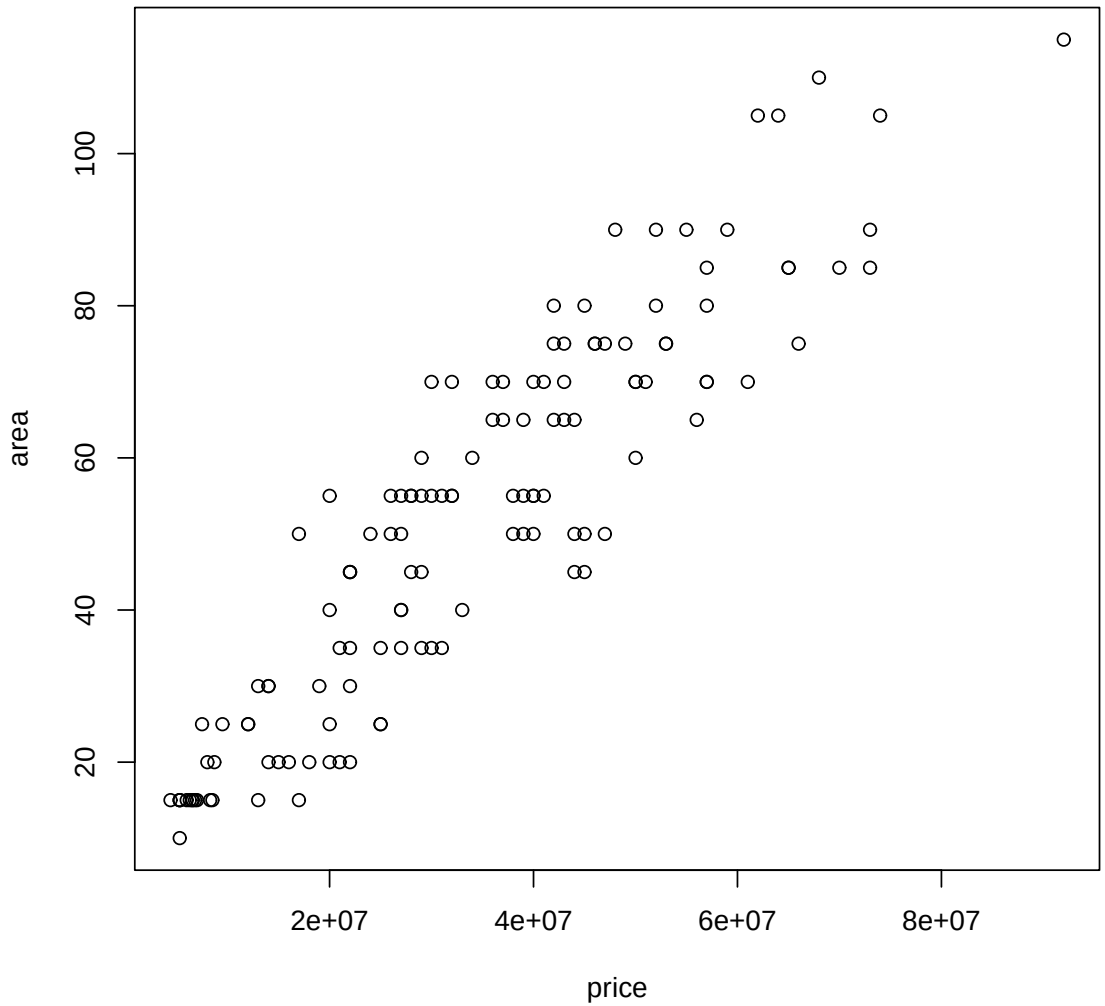
使い方はスクリプトを参照

k-近傍密度推定方のパッケージは見つからなかった．スクリプトに関数を定義してある．

- knn(train, test, cl, k = 1, l = 0, prob = FALSE, use.all = TRUE): FNN パッケージに入っている．
  - l や use.all はタイを処理するための引数．
  - 同パッケージに入っている knn.dist などを用いれば KD 木（等）を用いた k-近傍の高速計算が可能．
- kknn: kknn パッケージに入っている．重み付き k-近傍法が使える．

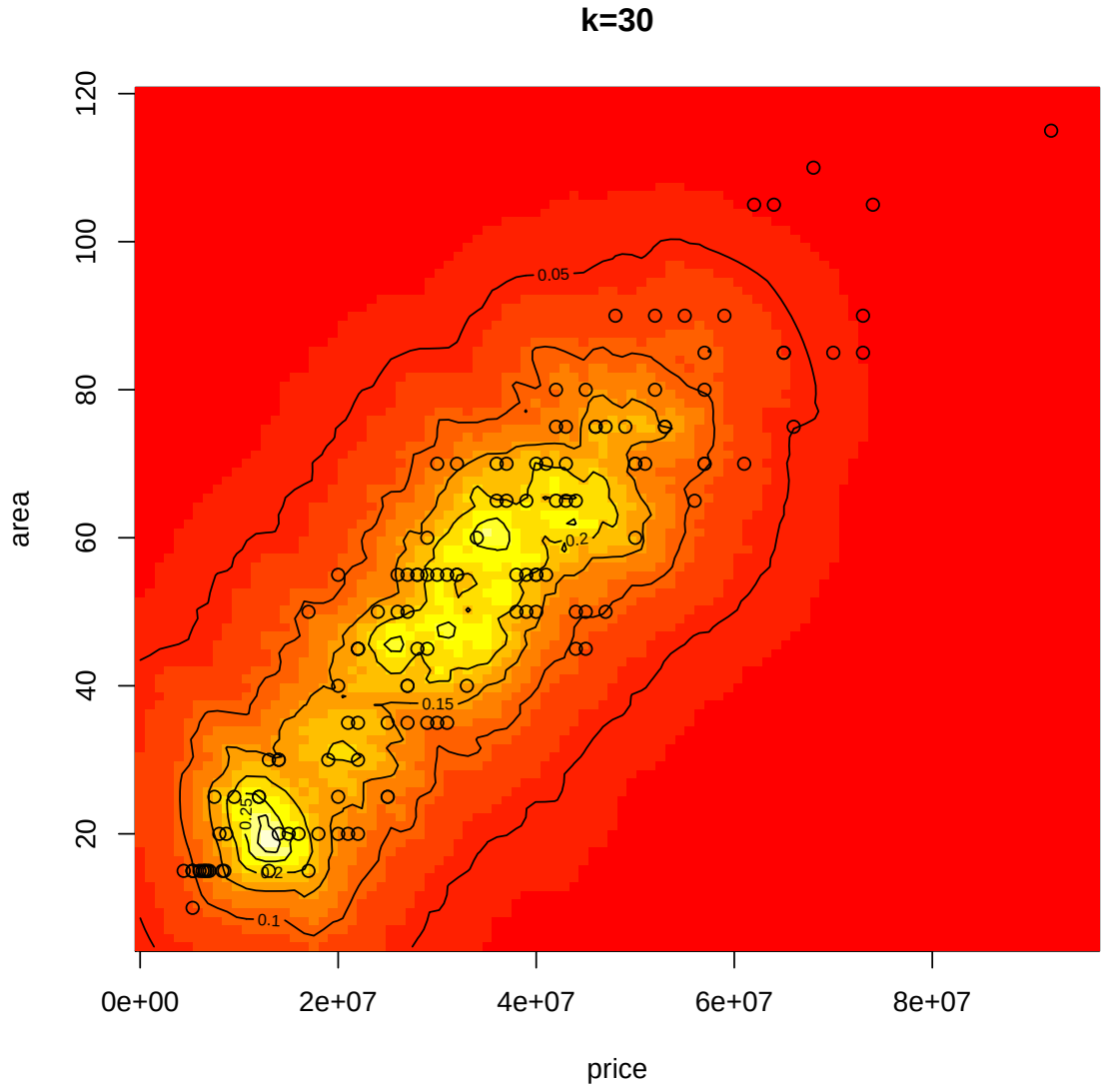
世田谷区中古マンション価格データ

$k = 30$ を採用．不安定．



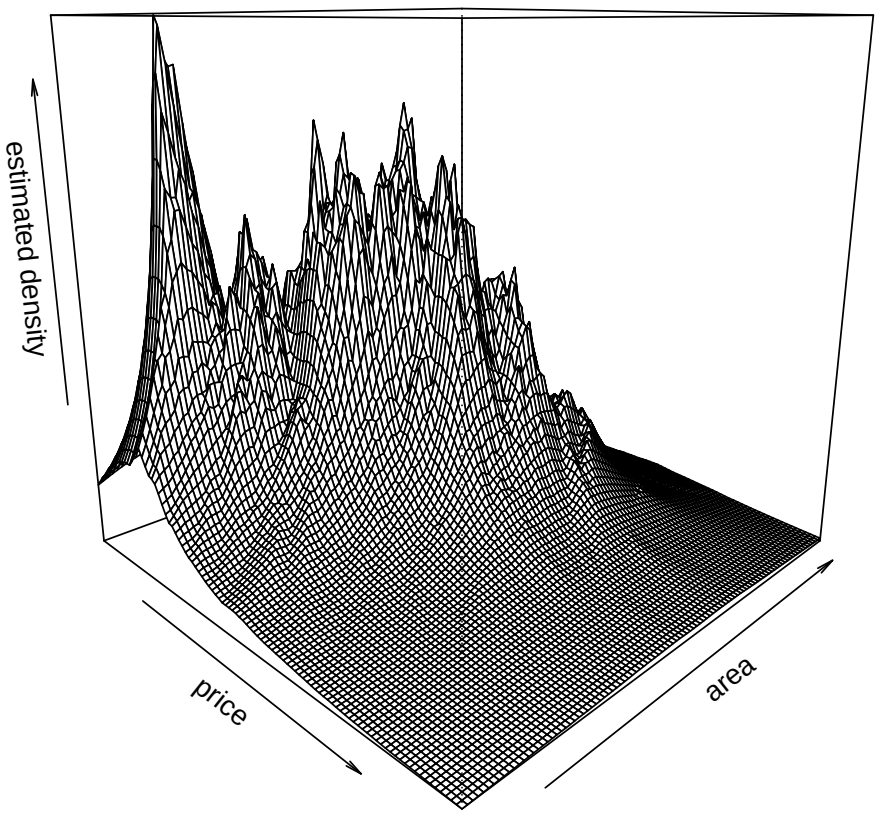
世田谷区中古マンション価格データ

$k = 30$ を採用．不安定．

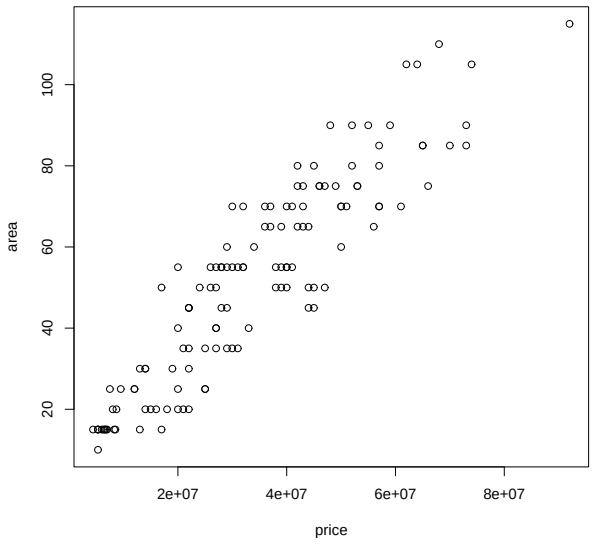


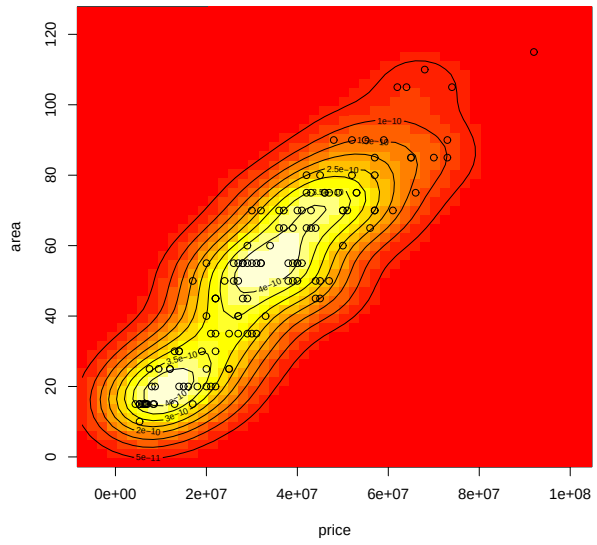
世田谷区中古マンション価格データ

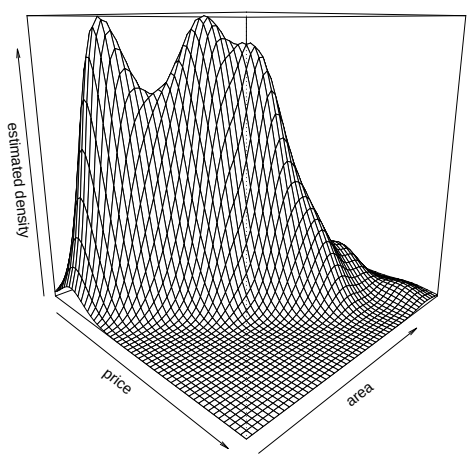
$k = 30$ を採用．不安定．



カーネル密度推定の結果 (参考)

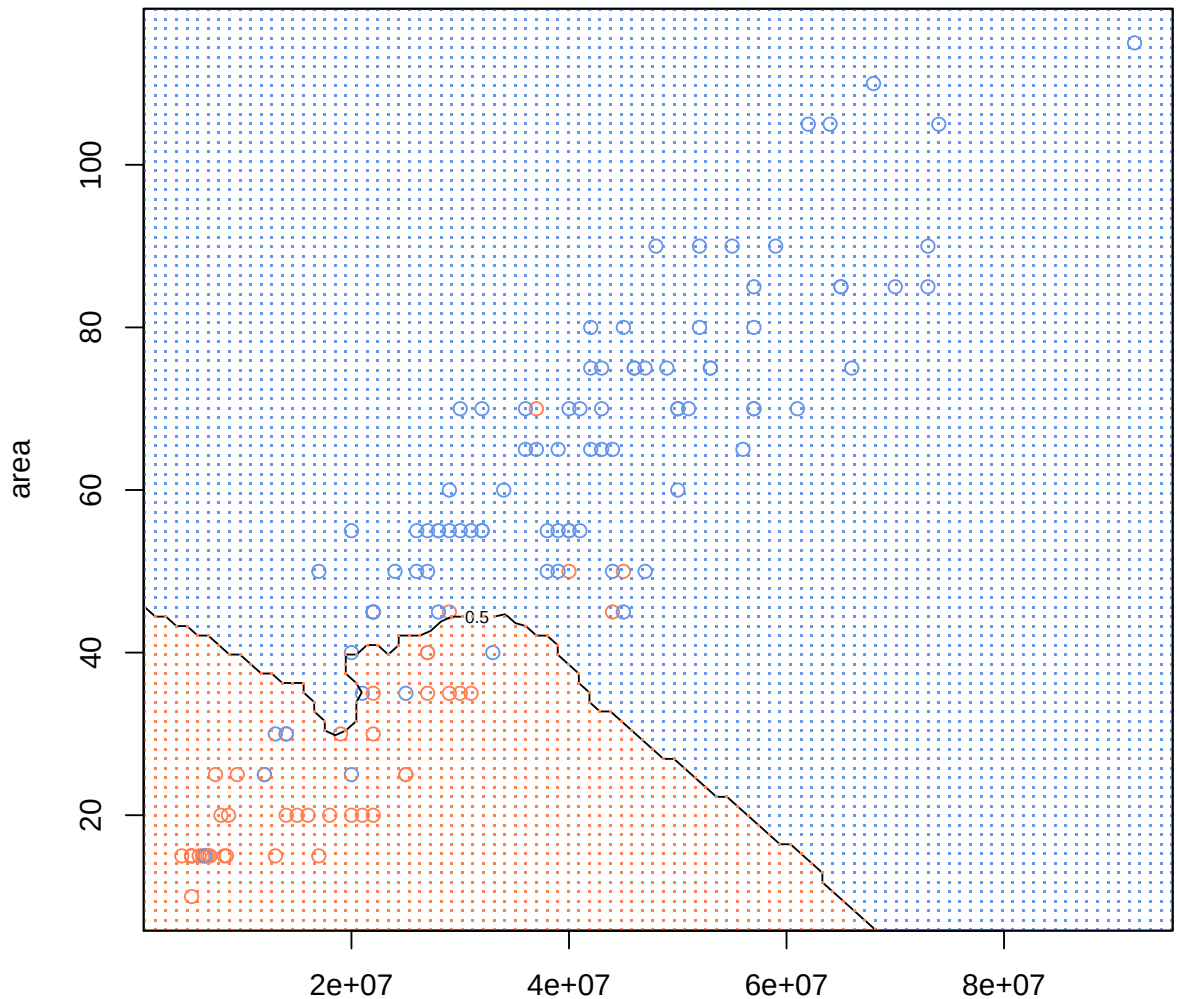






世田谷区中古マンション価格データの判別

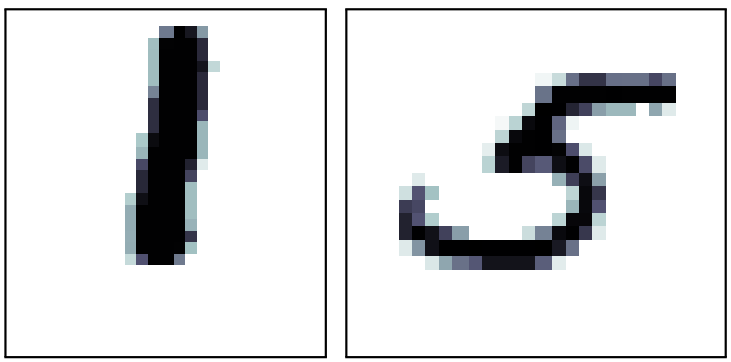
1K, 1DK, 1LDK とそれ以外（2LDK など）とを判別．$k = 10$．赤が部屋数 1 で，青がそれ以上．

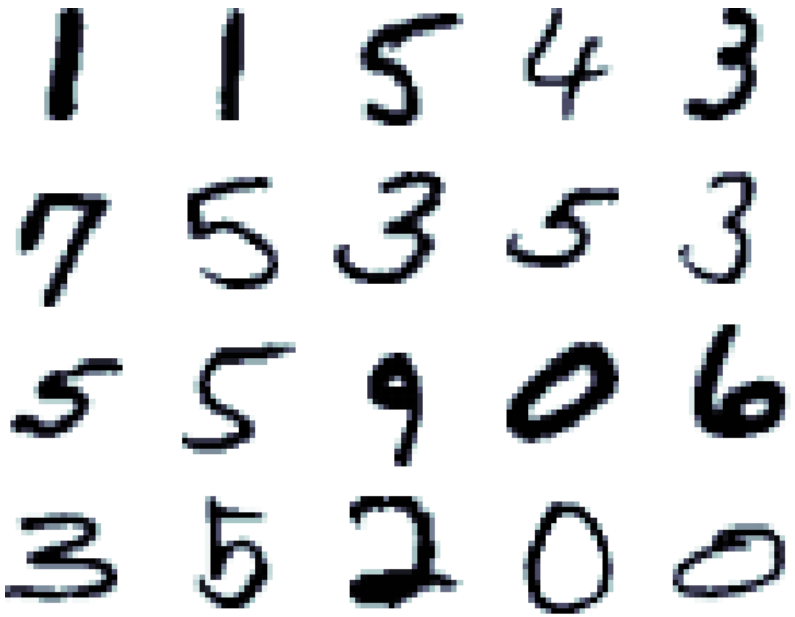


手書き文字認識

MNIST 手書き文字データ:

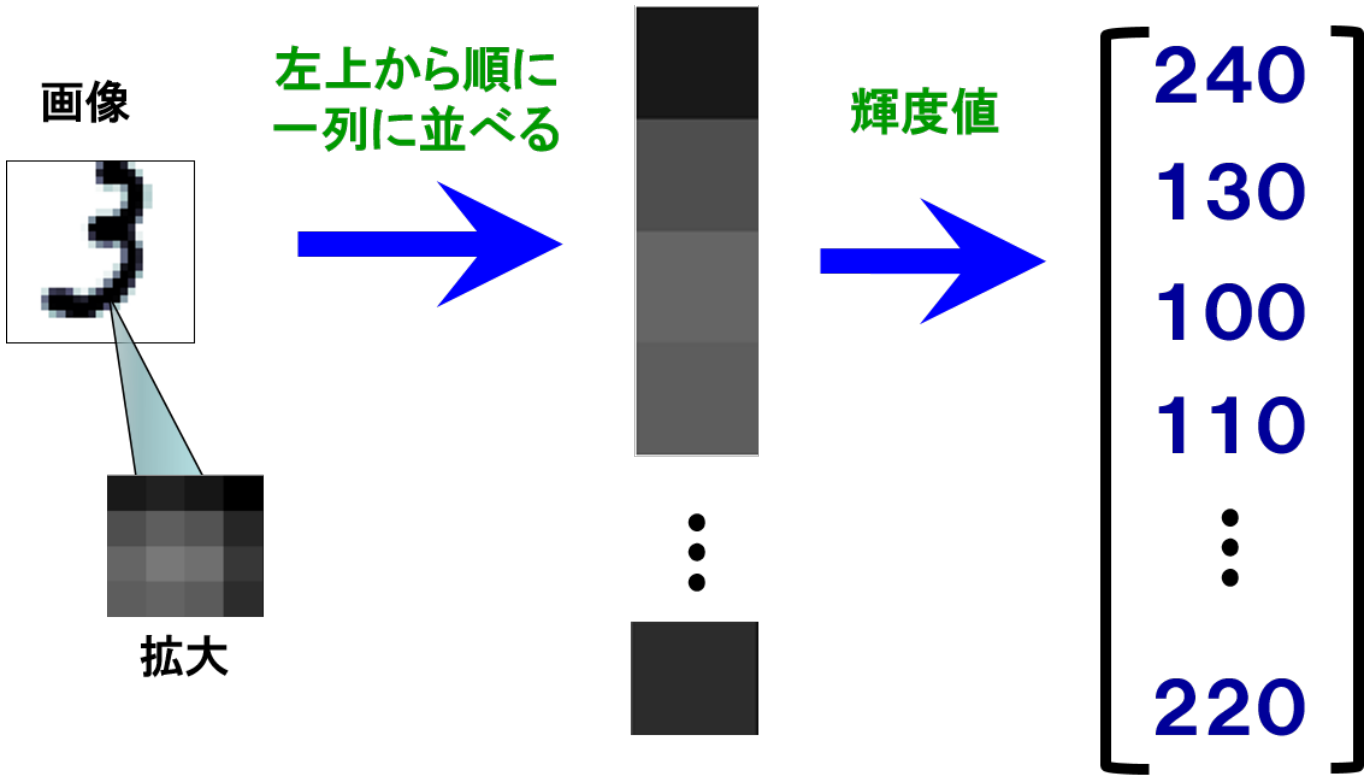
- 28×28 のグレースケール画像.
- 6000 個の訓練サンプル, 10000 個のテストサンプル.





※ 講義情報ページから csv ファイルを入手可能.

データ形式



輝度値は 0 から 255 の整数値.

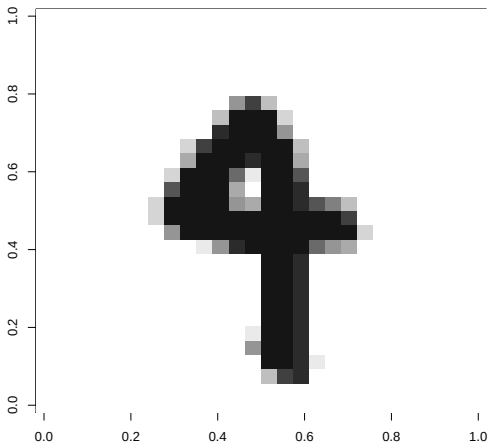
k-近傍判別

$k = 10$ で 96.65%の正答率.

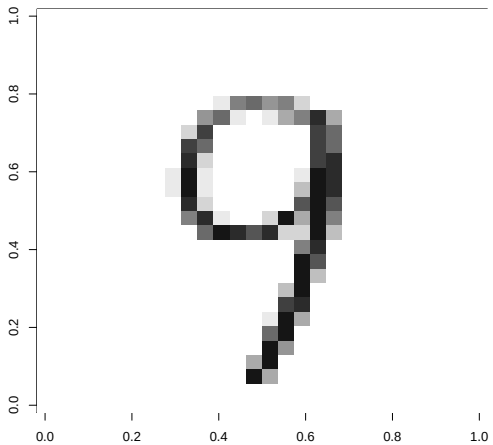
```
> res = knn(data.train_origin[,1:dimx], data.test[,1:dimx],
      data.train_origin[,dimy], k = 10, prob=FALSE)
> (clerr = mean(res!=data.test[,dimy])) #0.0335, 96.65%の確率で正解.
[1] 0.0335
> table(res,data.test[,dimy])
```

res	0	1	2	3	4	5	6	7	8	9
0	972	0	13	0	2	4	6	0	6	7
1	1	1132	12	3	11	0	4	27	4	6
2	1	2	982	3	0	0	0	4	5	3
3	0	0	2	976	0	12	0	0	11	7
4	0	0	1	1	940	1	3	2	7	10
5	2	0	0	10	0	863	2	0	9	3
6	3	1	2	1	4	6	943	0	4	1
7	1	0	17	7	1	1	0	983	7	10
8	0	0	3	6	1	1	0	0	914	2
9	0	0	0	3	23	4	0	12	7	960

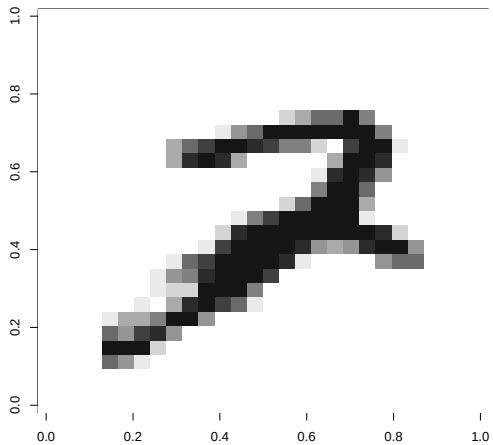
誤判別例



判別結果 9
正解 4



判別結果 4
正解 9



判別結果 7
正解 2

まとめ

k-近傍法によるデータ解析

1. 密度推定
2. 判別

k の選択は重要であった.

レポート問題 (第四回)

1. 自分の興味のあるデータで講義第八回に紹介した検定手法を当てはめてみよ．もしデータが見つからなければ「autopoll.csv」で分散分析を行ってみよ．
2. 講義第七回に説明した主成分分析を実装せよ．具体的には，データ x を受け取って第 k 番目までの主成分スコア，対応する主成分負荷 (λ_j) および主成分のリストを返す関数を作成せよ．
   - *Yk*: k 番目までの主成分スコア
   - *Lk*: k 番目までの主成分負荷 $(\lambda_1, \ldots, \lambda_k)$
   - *Vk*: k 番目までの主成分ベクトルを並べた行列
3. 上で得た自作の関数と R の組み込み関数 princomp の返り値とを比較せよ．
4. 「ken-c-kakou.csv」にあるデータを用いて主成分分析を行え（データは講義七回目の zip ファイルに入っている）．なお，このデータは総務省統計局「統計でみる都道府県のすがた．経済基盤」から抜粋した：

   `http://www.stat.go.jp/data/k-sugata/index.htm`
   `http://www.e-stat.go.jp/SG1/estat/List.do?bid=000001052235&cycode=0`

   データの読み込みは `x <- read.csv("ken-c-kakou.csv",header=TRUE)` で可能．
5. (optional) 自分が興味のあるデータで主成分分析を行え．

レポートの提出方法

- 私宛にメールにて提出.
- 件名に 必ず「データ解析第 n 回レポート」と明記し，R のソースコードと結果をまとめたレポートを送付のこと.
- 氏名と学籍番号も忘れず明記すること.
- レポートは本文に載せても良いが，pdf などの電子ファイルにレポートを出力して添付ファイルとして送付することが望ましい (これを期に tex の使い方を覚えることを推奨します).
- 出力結果をただ提示するだけでなく，必ず考察を入れること.
- 提出期限は講義最終回まで.

※相談はしても良いですが，コピペは厳禁です.

講義情報ページ

`http://www.ocw.titech.ac.jp/index.php?module=General&action=T0300&GakubuCD=100&GakkaCD=15&KougiCD=5522&Nendo=2015&Gakki=1&lang=JA&vid=03`

`http://www.is.titech.ac.jp/~s-taiji/lecture/2015/dataanalysis/dataanalysis.html`

G. Biau, F. Chazal, D. Cohen-Steiner, L. Devroye, and C. Rodríguez. A weighted k-nearest neighbor density estimate for geometric inference. *Electron. J. Statist.*, 5:204–237, 2011. doi: 10.1214/11-EJS606. URL http://dx.doi.org/10.1214/11-EJS606.