2.5 ノンパラメトリック法

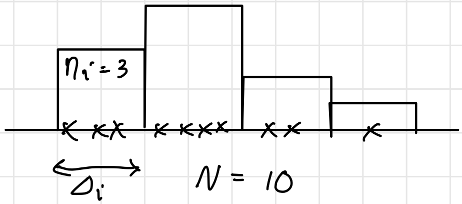
これまで、データの生成分布にガウス分布等を仮定する、パラメトリックな方法扱ってきた。ここでは、分布の形状について僅かな仮定しかおかないノンパラメトリックなアプローチを示す。

・ヒストグラム密度推定法
・カーネル密度推定法
・最近傍法

ヒストグラム密度推定法

x を幅Δ_i で区切って、各区間ごとに確立密度を次のように定める。

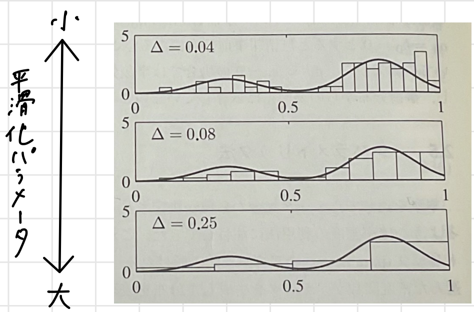
$$p_i = \frac{n_i}{N \Delta_i}$$



これは $\int p(x)\,dx = 1$ をみたす。

$$\because \int p(x)\,dx = \sum_i p_i \Delta_i = \sum_i \frac{n_i}{N\Delta_i} \Delta_i = \sum_i \frac{n_i}{N} = 1$$

Δ_i の大きさが全て同一にΔのとき、これは平滑化パラメータとなる。



2.5.1 カーネル密度推定法

ヒストグラム密度推定法には、二つの主要な問題がある。(i). 確率密度関数が連続とならない (ii). 次元が増えると区間の総数が指数的に増加する。この問題を解決している二つの方法（カーネル密度推定法、K近傍法）について学ぶ。

まず、「カーネル」がどのようなものであるかについて考える。

確率密度を $p(x)$ としたときに、小さな領域 R に割り当てられた確率は、

$$P = \int_R p(x)\,dx$$

データが N コあるとき、R に K コのデータが入る確率は、

$$\mathrm{Bin}(K \mid N, P) = \frac{N!}{K!(N-K)!} P^K (1-P)^{N-K}$$

N が大きいとき、分散 $\frac{P(1-P)}{N} \to 0$ より、K は期待値で近似できる。

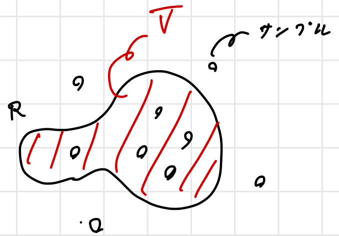
$$K \simeq NP$$

一方で、R が十分小さく、領域内で一定とみなせるならば、

$$P \simeq p(x) \underbrace{V}_{R\text{の体積}}$$

これより、$p(x)$ について解くと、

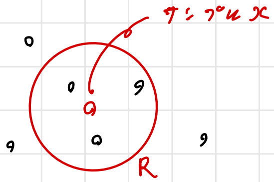
$$p(x) = \frac{K}{NV}$$



方針

K近傍法：Kを固定し、Vを推定する。

K=3のとき

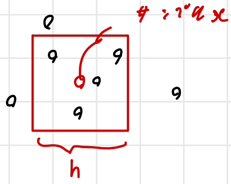


データが3つ含まれるまでRを広げていき、Vを求める。

$$p(x) = \frac{K}{NV}$$

カーネル推定法：Vを固定し、Kを推定する。

$V = h^D$ （D次元立方体の体積）とすると、



立方体の中に含まれるサンプル数Kを求める。

$$p(x) = \frac{K}{NV}$$

カーネル法を解くために、超立方体R内に含まれるサンプル数Kを求める。
次の関数 k() を使う。

$$k(u) = \begin{cases} 1 & |u_i| \leq \frac{1}{2} \\ 0 & それ以外 \end{cases}$$

原点を中心とした、一辺の長さ1の立方体

$k\left(\frac{x - x_n}{h}\right)$ は x を中心とした、一辺 h の立方体に x_n が含まれるときのみ 1 となる。よって K は次のようになる。

$$K = \sum_{n=1}^{N} k\left(\frac{x - x_n}{h}\right)$$

よって x での推定密度は次のようになる。

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^D} k\left(\frac{x - x_n}{h}\right)$$

ここで、関数k()はカーネル関数と呼ばれる。今回のような非連続なカーネル関数の代わりに、連続なカーネル関数を用いることで、p(x) も連続となる。例として、次のようなガウスカーネルを考える。

$$k(u) = \frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{\|u\|^2}{2}\right)$$

次に、K、Vを次のように置くことで、p(x) を求めることができる。

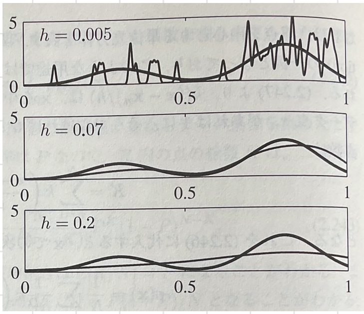
$$K = \sum_{n=1}^{N} k\left(\frac{x - x_n}{h}\right) \qquad V = \frac{1}{h^D}$$

合ってる??

$$P(x) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{D/2}}\exp\left(-\frac{\|x-x_n\|^2}{2h^2}\right)$$

平滑化パラメータ

小



大

カーネル関数は、次の条件を満たす任意の関数を選択することができる。

$$\begin{cases} k(u) \geq 0 \\ \int k(u)\,du = 1 \end{cases}$$

このような密度推定法は、訓練段階では計算を必要としないが、推論時にはデータの個数に比例した計算が必要となるという欠点がある。

2.5.2 最近傍法

カーネル法には、カーネル幅を決めるパラメータが一定であるため、密度の高い領域では平滑化されすぎ、密度が低いところではノイズが多くなりやすい性質がある。最近傍法による推定では、密度に応じてhを変化させることでこの問題を解決している。

Kの値を固定し、xを中心とした小球が、K個のサンプルを含むまで、小球の半径を大きくしていく。この時の小球の体積をVとして、p(x) を求める。この方法をK近傍法と呼ぶ。

K近傍法は、クラス分類問題に拡張することができる。クラスに関わらず、K個のサンプルを含むような小球の体積Vを求める。
クラスkの総サンプル数をN_k、小球内のクラスkのサンプル数をK_kとすると、次が成り立つ。

$$P(x \mid C_k) = \frac{K_k}{N_k V}$$ クラスkの確率密度

$$p(x) = \frac{K}{N V}$$ クラス全体の確率密度

$$P(C_k) = \frac{N_k}{N}$$ クラスの事前分布

これらを組み合わせると、xが得られた時に、それがクラスkである確率は、次のようになる。

$$P(C_k \mid x) = \frac{P(x \mid C_k)\, p(C_k)}{P(x)}$$

$$= \frac{N V}{K} \cdot \frac{K_k}{N_k V} \cdot \frac{N_k}{N}$$

$$= \underbrace{\frac{K_k}{K}}$$

小球内のサンプルのうち、クラスkのサンプルの割合

K近傍法もカーネル密度推定法も、データ全体を保持しなくてはならない。データ集合が大きいと膨大な計算量が必要になるが、木構造などによるデータの保持によって効率化できる。

例：3近傍法を適用したいとする。

$$X = [\overbrace{4, 8, 9, 2, 7, 6}^{N=6}]$$

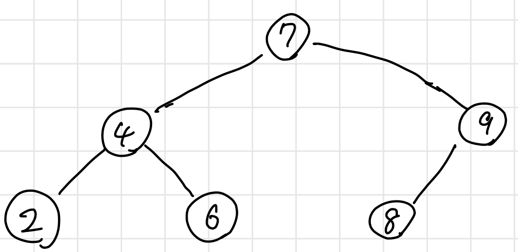
の中から、$x = 5$ に近い3つのサンプルを得たい。

<u>ナイーブな方法</u>

X内の全ての要素と x=5 を比較する。
-> 計算量：N

<u>木構造を使う方法</u>

二分探索木を使う例を挙げる。
二分探索木とは、任意のノードについて、右にぶら下がる木はそのノードの値よりも小さく、左にぶら下がる木はそのノードの値よりも小さくなるような木構造



高さは $\log_2 N$

x=5 に近い3つのサンプル(K=3)を探すときの計算量を考える。
このために、x=5 より大きく最も近いサンプル3つと、x=5 より小さく最も小さい3つをそれぞれ持ってきて、6つの中から比較する手法をとるとする。

このとき、1個のサンプルを取るために必要な計算量は logN であるため(*)、全体の計算量は 2K logN + 2K = O(K logN) となる。

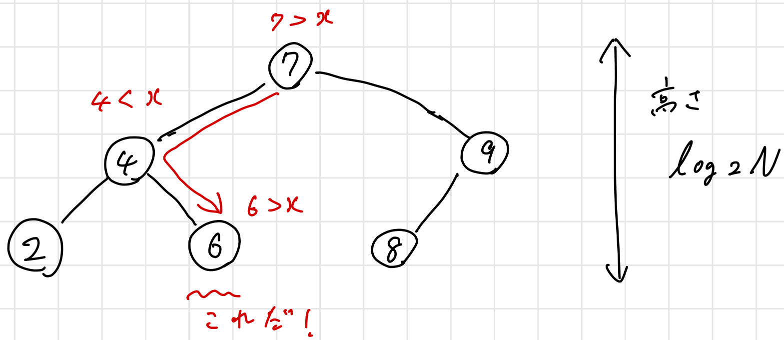
サンプルを2K個とってくる。

1回とってくるのにlog Nの計算

2Kのうち、x=5に近いK個を選ぶ。

(*) について、

x=5 より大きい、最も近いサンプルを探す。



計算量 = 比較の回数 = 木の高さ = $\log_2 N$

次は、x=6 として 6 より大きな最も近いサンプルを探す
という操作を繰り返す。

参考：

Kaggle GMの人