Métodos não-paramétricos

De modo geral, podemos dizer que métodos não paramétricos removem a necessidade de assumirmos um tipo particular de função para fazer algum processo de inferência. Um exemplo desse tipo de abordagem já utilizado por nós é o bootstrap e o histograma.

Kernel density estimation

Histogramas, como já vimos, permitem estimar a forma da distribuição de probabilidade de um dado, usando, apenas o próprio dado. De modo mais geral, podemos pensar que histogramas são um caso particular do procedimento chamado Kernel density estimation (KDE). Para notar essa similaridade, considere um conjunto de dados em d dimensões $\{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n\}$. Assuma que a região do dado possa ser dividida em N hipercubos de volume h^d. Considere, ainda, uma função

$$K(\bar{u}) = \begin{cases} 1 & \text{se } |u_j| < 1/2 \quad \forall j = 1, 2, \dots, d \\ 0 & \text{caso contrário.} \end{cases}$$

Desse modo, o número de pontos $\bar{x}_i$ contidos no hipercubo centrado em $\bar{x}_k$ é

$$C_{\bar{x}_k} = \sum_{i=1}^{n} K\left(\frac{\bar{x}_k - \bar{x}_i}{h}\right)$$

Note que a quantidade de $K\left(\frac{\bar{x}_k - \bar{x}_i}{h}\right)$ é 1 se $\bar{x}_i$

estiver dentro do cubo. Uma vez que estamos interessado na densidade de probabilidade, podemos escrever

$$\hat{P}(\bar{x}_k) = \xi \sum_{i=1}^{n} K\left(\frac{\bar{x}_k - \bar{x}_i}{h}\right)$$

sendo ξ uma contante de normalização. Para determinar essa constante, notamos que:

$$\int \hat{P}(\bar{x})\, d^d\bar{x} = 1 \Rightarrow \sum_{k=1}^{N} \hat{P}(\bar{x}_k)\, h^d = 1$$

$$\sum_{k=1}^{N} \xi \sum_{i=1}^{n} K\left(\frac{\bar{x}_k - \bar{x}_i}{h}\right) h^d = 1$$

$$\xi h^d \sum_{k=1}^{n} C_{\bar{x}_k} = 1$$

$$\xi h^d n = 1$$

$$\xi = \frac{1}{n h^d}$$

Sendo assim, ficamos com

$$\hat{P}(\bar{x}_k) = \frac{1}{n h^d} \sum_{i=1}^{n} K\left(\frac{\bar{x}_k - \bar{x}_i}{h}\right)$$

Nessa expressão, h é o chamado bandwidth ou tamanho do bin.

Entre outros resultados, é possível mostrar que

$$V(\hat{p}(\bar{x})) \lesssim \frac{1}{n h^d}$$

e que risco

$$R(p, \hat{p}) = \int \mathbb{E}\left[(p(x) - \hat{p}(x))^2\right] d^d\bar{x} \lesssim h^{2d} + \frac{1}{n h^d}$$

Além disso,

$$\sup_{p} R(p,\hat{p}) \leq C_0 \left(\frac{1}{n}\right)^{\frac{2}{d+2}}$$

<u>Exemplo notebook</u>

O procedimento anterior, que também é chamado de Parzen-Rosenblatt window aproach, tem vários problemas, entre eles:

-> é naturalmente descontínuo

-> pondera igualmente todos os $\bar{x}_i$, independentemente da distância ao centro do cubo.

Por essas razões, é comum substituir a função $K(u)$ por funções mais suaves, os chamados <u>smooth kernel functions</u>. Essas funções são usualmente radialmente simétricas com as seguintes propriedades:

$$\int K(\bar{x})\, d^d\bar{x} = 1$$

$$\int x\, K(\bar{x})\, d^d\bar{x} = 0$$

$$0 < \int x^2\, K(\bar{x})\, d^d\bar{x} < \infty$$

Um exemplo típico é o kernel gaussiano

$$K(\bar{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\bar{x}^2}{2}\right)$$

<u>mostrar figura 3.27</u>

Seleção do bandwidth

Como deve estar claro, o procedimento KDE depende da escolha do bandwidth h. Valores muito grandes de h tendem a suavisar muito a estimativa de $p(\bar{x})$, enquanto valores muito pequenos produzem muitos picos. De modo geral, desejamos encontrar um valor de h que minimize

$$\mathbb{E}\left(\left(p(\bar{x}) - \hat{p}(\bar{x})\right)^2\right) = \underbrace{\mathbb{E}\left[\left(p(\bar{x}) - \hat{p}(x)\right)^2\right]}_{\text{bias}} + \underbrace{\mathbb{V}\left(\hat{p}(x)\right)}_{\text{variance}}$$

Existem várias rule-of-thumb para escolher h que são derivadas minimizando a função anterior; porém, assumindo uma forma particular para $p(\bar{x})$. Por exemplo, a regra de Scott mostra que

$$h = \sigma / n^{-1/(d+4)}$$

é a melhor escolha quando $\bar{x}$ é gaussiana com desvio padrão σ. O problema desse tipo de regra é que não conhecemos $p(\bar{x})$ na maioria dos casos típicos.

Uma outra possibilidade baseada apenas nos dados é a chamada cross-validation. A ideia é escrever o integrated squared error (ISE) como

$$ISE(\hat{p}, p) = \int \left(p(\bar{x}) - \hat{p}(\bar{x})\right)^2 d\bar{x}$$

$$= \int \hat{p}^2(\bar{x})\, d\bar{x} - 2\int p(\bar{x})\,\hat{p}(\bar{x})\, d\bar{x} + \int p^2(x)\, d\bar{x}$$

O segundo termo da expressão anterior é problemático visto que não conhecemos $p(x)$. Porém, notamos que esse termo é um valor esperado, isto é,

$$\int p(\bar{x})\hat{p}(\bar{x})d\bar{x} = \mathbb{E}(\hat{p}(\bar{x}))$$

o qual pode ser aproximado usando os dados, ou seja,

$$\mathbb{E}(\hat{p}(\bar{x})) \approx \frac{1}{n}\sum_{i=1}^{n}\hat{P}(\bar{x}_i)$$

Porém, existe ainda o problema de que $\mathbb{E}(\hat{p}(\bar{x}))$ e $\hat{p}(\bar{x})$ são estimados usando o mesmo dado. Para contornar esse problema, podemos dividir o dados em duas partes (D_1 e D_2), estimar $\hat{p}(x)$ para diferentes valores de h usando D_1 e usar D_2 para estimar $\mathbb{E}(\hat{p}(x))$. Sendo assim, ficamos com

$$ISE(\hat{p},p) \cong \int \hat{p}^2(\bar{x})d\bar{x} - \frac{2}{|D_2|}\sum_{x_i \in D_2}\hat{P}(\bar{x}) + \int p^2(\bar{x})dx$$

o qual pode ser minimizada variando o valor de h. Observe que o último termo na expressão anterior é uma constante.

Exemplo notebook

Regressão não-paramétrica

Além de estimar distribuições de probabilidade, métodos não paramétricos podem ser úteis para estimar a função geradora dos dados, isto é, a forma funcional de

$f(x)$ em

$$y_i = f(x_i) + \epsilon_i$$

sendo ϵ_i um erro aleatório.

Um exemplo desse tipo de regressão é o chamado linear smother, o qual é definido por

$$\hat{y}(x) = \sum_{i=1}^{n} l_i(x) y_i$$

sendo $l_i(x)$ uma função. O caso mais simples para essa função é assumir que valores de x_i possam ser particionados em m bins $\{B_1, B_2, \dots, B_m\}$ e

$$l_i(x) = \begin{cases} 1/K_i & \text{se } x_i \in B_i \\ 0 & \text{caso contrário} \end{cases}$$

sendo K_i o número de observações no bin B_i. Esse procedimento consiste em estimar $\hat{y}(x)$ com a média dos y_i dentro de cada bin. Esse procedimento é chamado de regressograma (em analogia ao caso dos histogramas) ou também de nearest Neighbors gregression. Nesse último caso, no lugar de dividir o dado em m bins, dado um valor de x, essa regressão retorna a média dos K primeiros vizinhos do ponto x.

Para estudar a performance desse regressores, considere o risco

$$R(\hat{y}, y) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} \left[\hat{y}(x_i) - y(x_i) \right]^2 \right)$$

de modo que estamos interessados em obter o melhor $\hat{g}$ que minimiza o risco. Porém, novamente não sabemos quem é $y(x)$. Podemos usar o próprio dado para estimar o risco, ou seja,

$$\hat{R}(\hat{g},y) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}(x_i) - y_i\right)^2$$

na qual substituímos $y(x_i)$ pelo valor do dado y_i. Como no caso anterior, não podemos usar o próprio dado para estimar $\hat{g}(x_i)$ e o risco. Uma maneira de contornar esse problema é usar um procedimento chamado leave-one-out cross validation, no qual estimamos $\hat{g}$ usando todos os dados exceto um ponto (x_i, y_i). Por fim, usamos esse ponto para estimar o risco, o qual é costumeiramente denotado por

$$\hat{R}(\hat{g},y) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{g}_{(-i)}(x_i) - y_i\right]^2$$

e tomamos a média para todas as possibilidades de deixar de fora um ponto.

Esse procedimento de linear smoother pode ser escrito também na forma matricial usando

$$S_{ij} = l_i(x_j)$$

de modo que

$$\hat{y} = S y$$

com $\bar{y} = [y_1, y_2, \ldots, y_n]$ e $\hat{y} = [\hat{g}(x_1), \hat{g}(x_2), \ldots, \hat{g}(x_n)]$

Essa notação também permite escrever o risco como

$$\hat{R} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}(x_i)}{1 - S_{ii}} \right)^2$$

Exemplo notebook

Kernel regression

Uma maneira de generalizar o procedimento anterior é usar o chamado Nadaraya-Watson Kernel, definido por

$$\hat{y}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}$$

no qual h é o bandwidth. Desse modo, a diferença entre o KNN regression e o Kernel regression é que para esse último, temos uma média móvel contínua, enquanto para o KNN esse processo é naturalmente descontínuo.

Exemplo notebook

De modo geral, esse tipo de regressão apresenta problemas nas fronteiras do dado, sendo que esse problema se torna cada vez mais sério a medida que a dimensão do dado aumenta. (curse of dimensionality).

Curse of dimensionality (Maldição da dimensionalidade)

A expressão "maldição da dimensionalidade" refere-se vagamente a ideia de que "tudo" se torna mais complicado a medida que o número de dimesões do dado aumenta muito.

Para ter uma ideia qualitativa desse problema, considere o volume de uma esfera em n dimensões

$$V_s(n,r) = \begin{cases} \pi^{n/2} r^n / (n/2)! & n \text{ par} \\ 2^n \pi^{(n-1)/2} r^{n-1} \left(\frac{n-1}{2}\right)! \Big/ (n+1)! & n \text{ ímpar} \end{cases}$$

Considere agora uma esfera $V_s(n, 1/2)$ envolvida por um cubo de lado 1 em n dimensões. Note que o volume do cubo é 1 para todo n; porém, o volume da esfera tende a zero se $n \to \infty$. Isso significa que o volume do cubo é empurrado para longe de seu centro. Para ver isso, basta notar que a distância para o centro do cubo é $\sqrt{n}/2$ de suas arestas, enquanto para a esfera é sempre 1/2.

Uma consequência disso é que métodos baseados vizinhos mais próximos tornam-se muito difíceis de alcançar um valor pequeno para o bias explorando a localidade dos dados. Suponha, por exemplo, que desejamos localizar um ponto próximo a origem de um espaço n dimensional. Para isso, vamos usar os vizinhos desse ponto para estimar a média de seus valores. Ocorre que se a dimensão do espaço for muito grande, os primeiros vizinhos podem estar muito distantes e não serem representativos para o ponto em questão. Uma outra maneira de notar esse problema é considerar uma variável binária (como no lançamento de uma moeda). Suponha que tenhamos uma amostra de tamanho 1000. Nesse caso, podemos estimar com precisão razoável a probabilidade de cada um dos valores da variável. Suponha agora que tenhamos 10 variáveis binárias, de modo que existem $2^{10} = 1024$ possibilidades para esse conjunto. Desse modo, usando uma amostra de tamanho 1000, ao menos 24 configurações não estarão presentes na amostra e a estimativa da probabilidade de cada configuração não tem sentido. Para obter uma estatística razoável amostra deveria ser de aproximadamente 1000×1024 dados. Assim, um aumento de 10 no número de variáveis, implica um aumento de 1000 no tamanho da amostra.

Exemplo notebook