DỰ ĐOÁN XÁC SUẤT XE BUÝT VỀ TRẠM ĐÚNG GIỜ

Lê Thị Minh Thùy

Đại Học Bách Khoa TPHCM

Ngày 5 tháng 7 năm 2017

Nội dung

1. Giới thiệu đề tài
2. Dữ liệu
3. Phương pháp nghiên cứu
4. Cơ sở lý thuyết
5. Kết luận

Giới thiệu đề tài

Dự đoán xác suất xe buýt tuyến 72
lộ trình xuất phát từ BX Củ Chi về trạm đích BX An Sương
đúng giờ (hạn mức 45 phút)

Dữ liệu thô

Ví dụ 10 dữ liệu thô:

	Vĩ độ	Kinh độ	Thời điểm xuất hiện
1	10.844095	106.613688333333	2016-09-02 07:25:43
2	10.84298	106.614991666667	2016-09-02 07:26:02
3	10.8424316666667	106.615195	2016-09-02 07:26:22
4	10.8426816666667	106.615596666667	2016-09-02 07:26:42
5	10.84309	106.615203333333	2016-09-02 07:27:02
6	10.846395	106.61304	2016-09-02 07:30:48
7	10.84664	106.612861666667	2016-09-02 07:31:01
8	10.8475833333333	106.612253333333	2016-09-02 07:31:21
9	10.8488916666667	106.611426666667	2016-09-02 07:31:41
10	10.84932	106.61117	2016-09-02 07:31:53

Dữ liệu làm việc

Dữ liệu sau khi được đồng bộ hóa khoảng cách thời gian hồi đáp (20 giây)

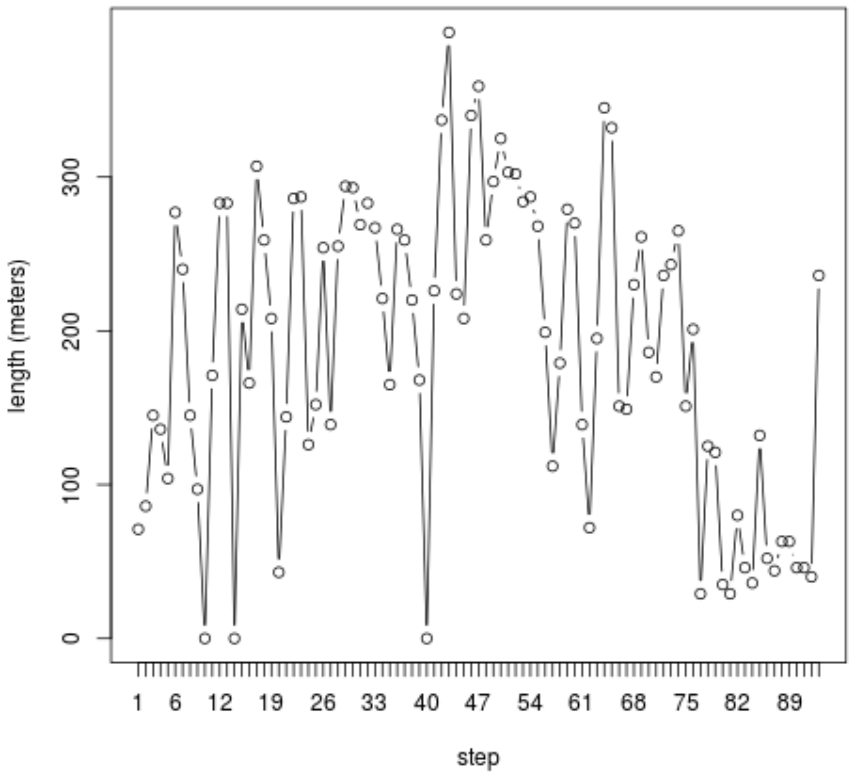
STT	Khoảng cách thời gian hồi đáp (giây)	Khoảng cách bước đi (mét)
1	20	71
2	20	86
3	20	145
4	20	136
5	20	104
6	20	277
7	20	240
8	20	145

STT	Khoảng cách thời gian hồi đáp (giây)	Khoảng cách bước đi (mét)
10	20	0
11	20	171
12	20	283
13	20	283
14	20	0
15	20	214
16	20	166
17	20	307

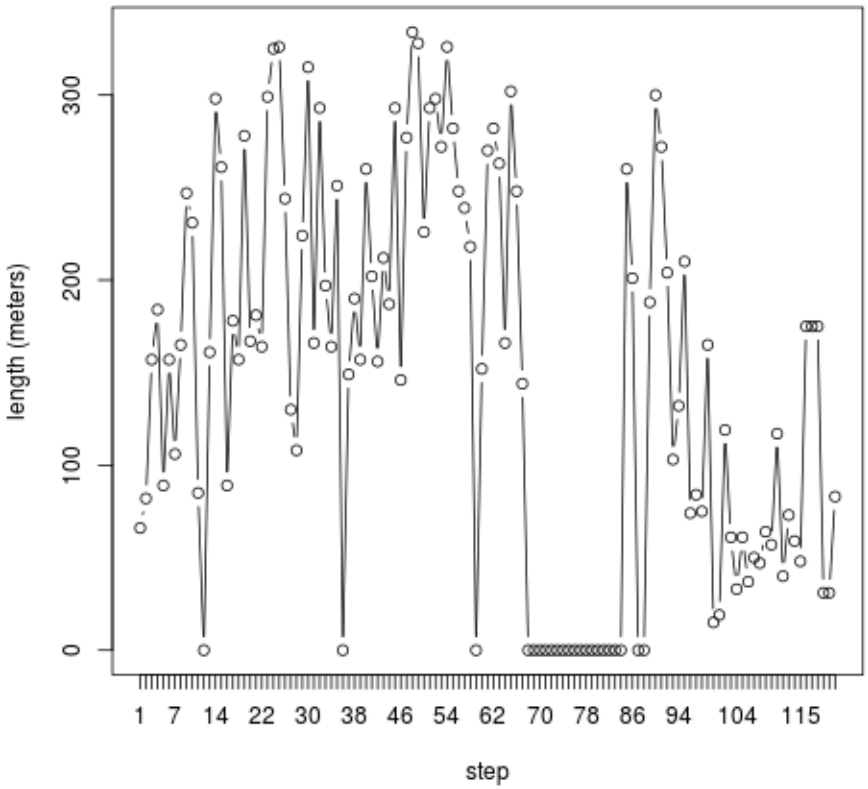
STT	Khoảng cách thời gian hồi đáp (giây)	Khoảng cách bước đi (mét)
85	20	132
86	20	52
87	20	44
88	20	63
89	20	63
90	20	46
91	20	46
92	21	40

Dữ liệu trực quan hóa

Trực quan hóa giá trị các bước di chuyển trong 80% quãng đường đầu



Chuyến xe 1 - đúng giờ



Chuyến xe 2 - trễ giờ

Phương pháp nghiên cứu

- số lượng dữ liệu mẫu giới hạn
- các bước di chuyển cách nhau 20 giây
- 80% lộ trình từ BX Củ Chi - BX An Sương
- trực quan hóa dữ liệu
- nếu có nhiều bước di chuyển dài thì xác suất về trạm đúng giờ cao
- chấp nhận có yếu tố ngẫu nhiên ảnh hưởng đến kết quả dự đoán

Cơ sở lý thuyết

- Thống Kê
- Không gian mẫu và biến cố
- Xác suất của biến cố
- Xác suất có điều kiện và sự độc lập của các biến cố
- Biến ngẫu nhiên
- Biến ngẫu nhiên rời rạc
- Biến ngẫu nhiên liên tục
- Xác suất Bayes
- Vấn đề thực tế khi áp dụng xác suất Bayes
- Phân phối chuẩn Gauss
- Phân loại Gaussian Bayes
- Thuật toán Kernel Density Estimation

Unofficial definition of statistics

Statistics is the science of problem-solving in the presence of **variability**.

Explain the word "variability"[1]:

- There are many situations that we encounter in science (or more generally in life) in which the outcome is uncertain.
- If the same measurement were repeated, then the answer would likely change

[1] http://www.stat.uci.edu/what-is-statistics/

Cơ sở lý thuyết - Không gian mẫu và biến cố

Định nghĩa không gian mẫu

Không gian mẫu là tập hợp của tất cả các kết cục có thể của một thí nghiệm cụ thể.

Định nghĩa biến cố

Các biến cố sơ cấp hay điểm mẫu là những phần tử của không gian mẫu.

Ví dụ:
Thí nghiệm tung một đồng xu, kết quả ngẫu nhiên là ngửa (Head, $\mathcal{H}$) hoặc sấp (Tail, $\mathcal{T}$), cho ta không gian mẫu $\mathcal{S} = \{\mathcal{H}, \mathcal{T}\}$
Các biến cố sơ cấp hay điểm mẫu là những phần tử của $\mathcal{S}$

Cơ sở lý thuyết - Xác suất của biến cố

Tập các biến cố $\mathcal{Q} := \{A : A \subset \mathcal{S}$ là một biến cố$\}$
Xét một hàm $\mathbb{P} : \mathcal{Q} \to \mathbb{R}$
$\mathbb{P}$(A) là khả năng hoặc cơ hội mà biến cố A xảy ra.
$\mathbb{P}$ được gọi là hàm xác suất khi thỏa mãn những tiên đề cơ bản sau đây:

- $\mathbb{P}(A) \geq 0$
- $\mathbb{P}(\mathcal{S})=1$
- Nếu ta có $E_1, E_2, ..., E_n$ (n $\geq$ 1) là các biến cố rời nhau từng đôi một thì

$$\mathbb{P}\left[\bigcup_{i=1}^{n} E_i\right] = \sum_{i=1}^{n} \mathbb{P}[E_i]$$

Cơ sở lý thuyết - Xác suất có điều kiện và sự độc lập của các biến cố

Nếu biến cố A xảy ra phụ thuộc vào biến cố B đã xảy ra $\mathbb{P}[B] > 0$ thì xác suất đồng thời của hai biến cố A và B:

$$\mathbb{P}(AB) = \mathbb{P}(A \cap B) = \mathbb{P}(B) \cdot \mathbb{P}(A|B)$$

Nếu biến cố A và B độc lập, sự xuất hiện của A không có liên quan đến sự xuất hiện của B thì xác suất đồng thời của hai biến cố A và B:

$$\mathbb{P}(AB) = \mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

Định nghĩa biến ngẫu nhiên

Một biến ngẫu nhiên là một hàm giá trị thực $X(\omega)$ (hay X) xác định trên một không gian mẫu $\mathcal{S}$, sao cho các biến cố $\{\omega \in \mathcal{S} : X(\omega) \leq x\}$ có thể được gán các xác suất, với mọi $-\infty < x < \infty$. Ta ghi $X : \mathcal{S} \to \mathbb{R}$. Thật vậy, với bất kỳ $x \in \mathbb{R}$, tập tiền ảnh $\{\omega \in \mathcal{S} : X(\omega) \leq x\} \subseteq \mathcal{S}$ rõ ràng là một biến cố, và được ký hiệu là A = X ≤ x hay $\{X \leq x\}$. Vậy xác suất $\mathbb{P}[A]$=$\mathbb{P}[X \leq x]$ luôn tồn tại.

Định nghĩa biến ngẫu nhiên rời rạc

X(.) là biến có một phạm vi $\mathcal{S}_X = X(\mathcal{S})$ là tập giá trị rời rạc (hữu hạn hoặc vô hạn đếm được, nghĩa là có lượng số không quá lượng số tập tự nhiên $\mathbb{N}$)

Bảng phân phối xác suất của X được cho bởi

X	x_0	x_1	$\dots$	x_{m-1}	x_m
$p_k := p(x_k) = \mathbb{P}[X = x_k]$	p_0	p_1	$\dots$	p_{m-1}	p_m

Cơ sở lý thuyết - Biến ngẫu nhiên rời rạc

Tập giá trị

$$\mathcal{S}_X = \{x_0, x_1, x_2, ..., x_{m-1}, x_m\}, m \in \mathbb{N}$$

Hàm mật độ xác suất

$$p(x) = \mathbb{P}[X = x] = \mathbb{P}[\{\omega : X(\omega) = x\}], x \in \mathcal{S}_X$$

với $p(x) \geq 0$ và $\sum_{x \in \mathcal{S}_X} p(x) = 1$

Hàm tích lũy xác suất

$$F_X(x) = \mathbb{P}[X \leq x] = \mathbb{P}[\omega \in S : X(\omega) \leq x] = \sum_{x_k \leq x} p(x_k) = \sum_{x_k \leq x} p_k, x \in \mathbb{R}$$

Kỳ vọng (hay trung bình)

$$\mu = \mathbb{E}[X] = \sum_{x_k \in \mathcal{S}_X} x_k p_k$$

Phương sai

$$\sigma^2 = \mathbb{E}[(X - \mu)^2] = \sum [x_k - \mu]^2 p_k$$

Định nghĩa biến ngẫu nhiên liên tục

X(.) là liên tục khi nó có phạm vi bao gồm khoảng con (hay toàn bộ) tập số thực, nghĩa là $\mathcal{S}_X \in \mathbb{R}$

Tập giá trị X nhận vô hạn giá trị không đếm được, $\mathcal{S}_X \subset \mathbb{R}$

Hàm mật độ xác suất

Tính chất của hàm mật độ xác suất f gồm:

- f(u) $\geq$ 0, $\forall$u.
- $\int_{-\infty}^{\infty} f(u)du = 1 = \mathrm{F}(-\infty)$
- $$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b)$$
$$= \int_a^b f(u)du = F(b) - F(a)$$
- Đạo hàm $f(x) = \frac{dF(x)}{dx}$ có thể không tồn tại ở một số hữu hạn giá trị x, trong khoảng hữu hạn bất kỳ.

Hàm tích lũy xác suất

Tồn tại một hàm số không âm $f(u)$ thỏa

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(u)d(u), \infty < x < -\infty$$

F(x) gọi là hàm phân phối (tích lũy) xác suất (c.d.f) của X, f(u) là hàm mật độ của X

Tính chất của hàm phân phối xác suất F gồm:

- F liên tục, và
- $\lim_{x->-\infty}$ F(x)=0, $\lim_{x->\infty}$F(x)=1
- F không giảm, nghĩa là nếu $x_1 < x_2$ thì $F(x_1) \leq F(x_2)$, và
- Quan hệ với f: hàm phân phối xác suất F(x) có đạo hàm $\frac{dF(x)}{dx} = f(x)$

Cơ sở lý thuyết - Xác suất Bayes

Định nghĩa

Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B)\cdot\mathbb{P}(A|B)}{\mathbb{P}(A)}$$

Cơ sở lý thuyết - Vấn đề thực tế khi áp dụng xác suất Bayes

Biến hoặc giá trị biến rtrên thực tế rất hiếm khi được phân loại, trong khi giải thuật xác suất Bayes $\mathbb{P}(B|A) = \frac{\mathbb{P}(B)\cdot\mathbb{P}(A|B)}{\mathbb{P}(A)}$ làm việc với biến và giá trị biến được phân loại.
Trong Luận Văn này:

- Biến là những bước di chuyển
- Giá trị biến là số lần lặp lại những bước di chuyển đó

Cơ sở lý thuyết - Vấn đề thực tế khi áp dụng xác suất Bayes

Có hai hướng giải quyết đối với thuộc tính ở dạng con số

- Rời rạc hóa thuộc tính dạng con số thành dạng phân loại => Dễ tranh cãi
- Sử dụng hàm mật độ xác suất cho biến liên tục

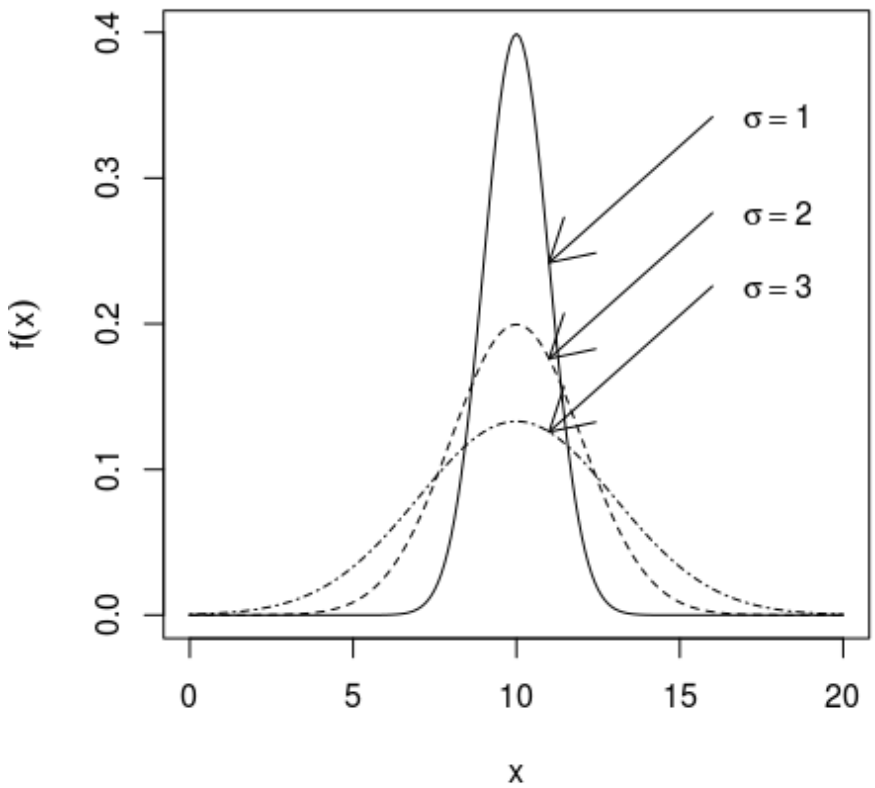
Cơ sở lý thuyết - Phân phối chuẩn (Gauss)

Biến ngẫu nhiên X có hàm mật độ là hàm

$$f(x) = n(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} * e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty, \mu \in \mathbb{R}, \sigma^2 > 0$$

Ký hiệu $X \sim \mathbf{N}(\mu, \sigma)$

Cơ sở lý thuyết - Phân phối chuẩn (Gauss)



Hàm mật độ của $\mathbf{N}(\mu, \sigma)$ với $\mu = 10, \sigma = 1, 2, 3$

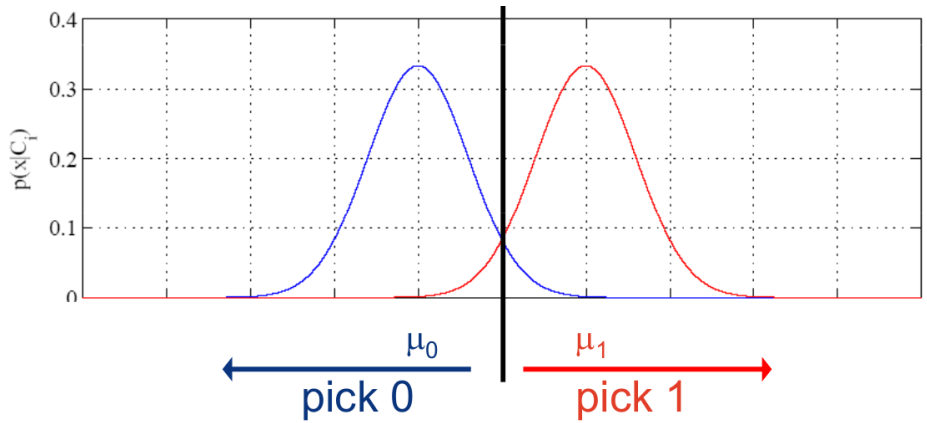
Cơ sở lý thuyết – Phân loại Gaussian Bayes

Gọi X là biến đầu vào
Gọi Y là biến phân loại lớp 0 hoặc 1, có xác suất $\mathrm{P_Y}(0)=\mathrm{P_Y}(1)=\frac{1}{2}$

Công thức hàm phân phối Gaussian $\mathrm{G}(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} * e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

với $\mu = \frac{\sum_1^n x_i}{n}$ và $\sigma^2 = \frac{\sum_1^n (x_i-\mu)^2}{n}$

Biến X có hàm phân phối Gaussian khác nhau theo mỗi phân loại, nghĩa là
$\mathrm{P_{X|Y}}(x|0) = \mathrm{G}(x,\mu_0,\sigma_0)$
$\mathrm{P_{X|Y}}(x|1) = \mathrm{G}(x,\mu_1,\sigma_1)$



Nếu x < $\frac{\mu_1+\mu_2}{2}$ thì phân loại x vào 0, nếu x > $\frac{\mu_1+\mu_2}{2}$ thì phân loại x vào 1.

Cơ sở lý thuyết – Thuật toán Kernel Density Estimation[2]

What is Kernel Density Estimation?

In statistics, kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a continuous random variable based on a finite data sample.

- A kernel is a special type of probability density function (PDF) with the added property that it must be even.
- Non-parametric models differ from parametric models in that the model structure is not specified a priori but is instead determined from data. The term non-parametric is not meant to imply that such models completely lack parameters but that the number and nature of the parameters are flexible and not fixed in advance.

[2] https://chemicalstatistician.wordpress.com/2013/06/09/exploratory-data-analysis-kernel-density-estimation-in-r-on-ozone-pollution-data-in-new-york-and-ozonopolis/

Cơ sở lý thuyết – Thuật toán Kernel Density Estimation

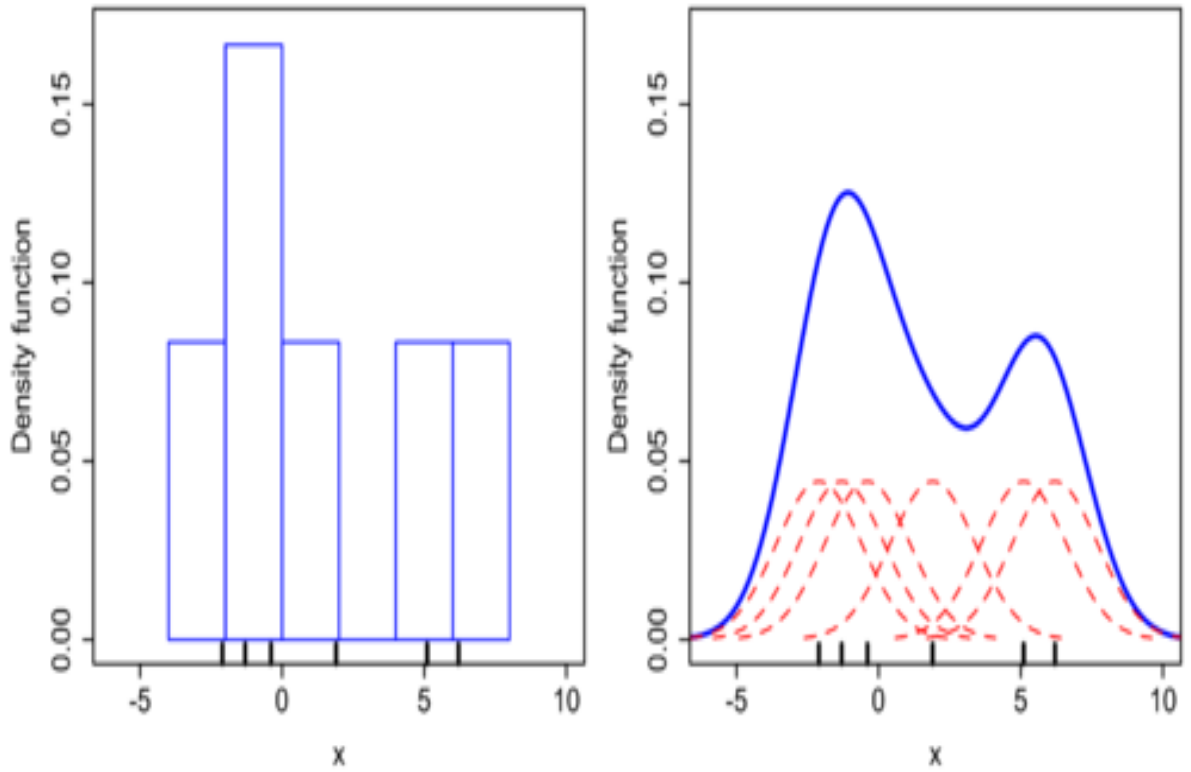
Constructing a Kernel Density Estimate: Step by Step

1. Choose a kernel, the common ones are normal (Gaussian), uniform (rectangular) and triangular.
2. At each datum, x_i, build the scaled kernel function $K_h = \frac{1}{h}K[\frac{(x-x_1)}{h}]$ where $K()$ is your chosen kernel function. The paramter h is called the bandwidth, the window width, or the smoothing paramter.
3. Add all of the individual scaled kernel functions and divide by n; this places a probability of $\frac{1}{n}$ to each x_i. It also ensures that the kernel density estimate integrates to 1 over its support set.

$$\hat{f}(x_i) = \hat{p}_{KDE}(x_i) = \frac{1}{n}\sum_{i=1}^{n} K_h = \frac{1}{n}\frac{1}{h}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

Cơ sở lý thuyết – Thuật toán Kernel Density Estimation[3]

Thuật toán Kernel Density Estimation



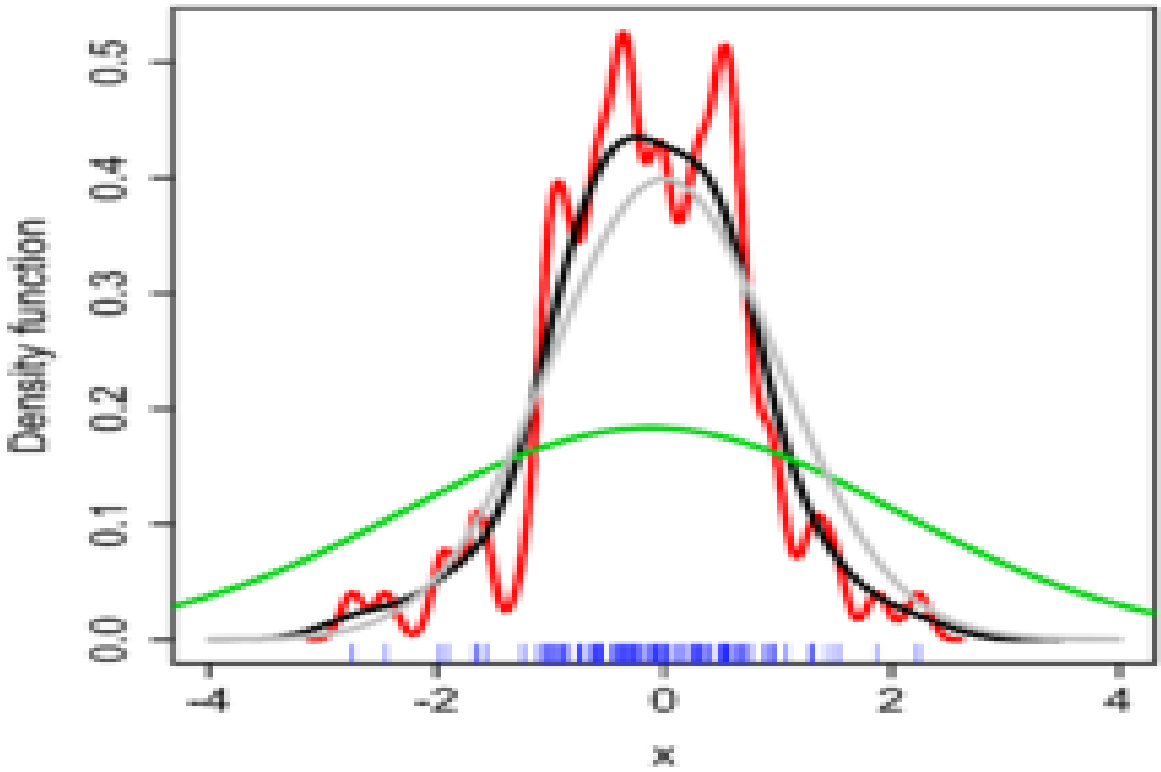
[3] https://en.wikipedia.org/wiki/Kernel_density_estimation

Cơ sở lý thuyết – Thuật toán Kernel Density Estimation

Choosing the Bandwidth

It turns out that the choosing the bandwidth is the most difficult step in creating a good kernel density estiamte that captures the underlying distribution of the variable.

Thuật toán Kernel Density Estimation



Kết Luận

Định nghĩa đồ thị phụ thuộc thời gian

Một đồ thị phụ thuộc thời gian được kí hiệu $G_T(V, E, W, F)$, viết tắt G_T, trong đó V $= \{v_i\}$ là tập hợp đỉnh, E $\subseteq$ V x V là tập hợp cạnh, W và F là hàm giá trị chi phí không âm. Đối với mỗi cạnh $(v_i, v_j) \in$ E, có hai hàm: hàm chi phí thời gian $\omega_{i,j}(t) \in$ W và hàm chi phí giá cả $f_{i,j}(t) \in$ F trong đó t là biến thời gian. Hàm chi phí thời gian $\omega_{i,j}(t)$ định nghĩa mất bao lâu để di chuyển từ v_i đến v_j nếu xuất phát v_i tại thời điểm t. Hàm chi phí $f_{i,j}(t)$ định nghĩa mất bao nhiêu chi phí giá cả để đi từ v_i đến v_j nếu xuất phát v_i tại thời điểm t.

Lý thuyết áp dụng giải bài toán

Định nghĩa tính chất FIFO

Cho đồ thị G_T có chi phí phụ thuộc vào thời gian, G_T là đồ thị có tính chất FIFO khi và chỉ khi nếu hàm thời gian đến đích $arrive_{i,j}(t)$ cho mỗi cạnh (v_i, v_j) đều có tính chất FIFO, nghĩa là $arrive_{i,j}(t_1) < arrive_{i,j}(t_2)$ với $t_1 < t_2$.

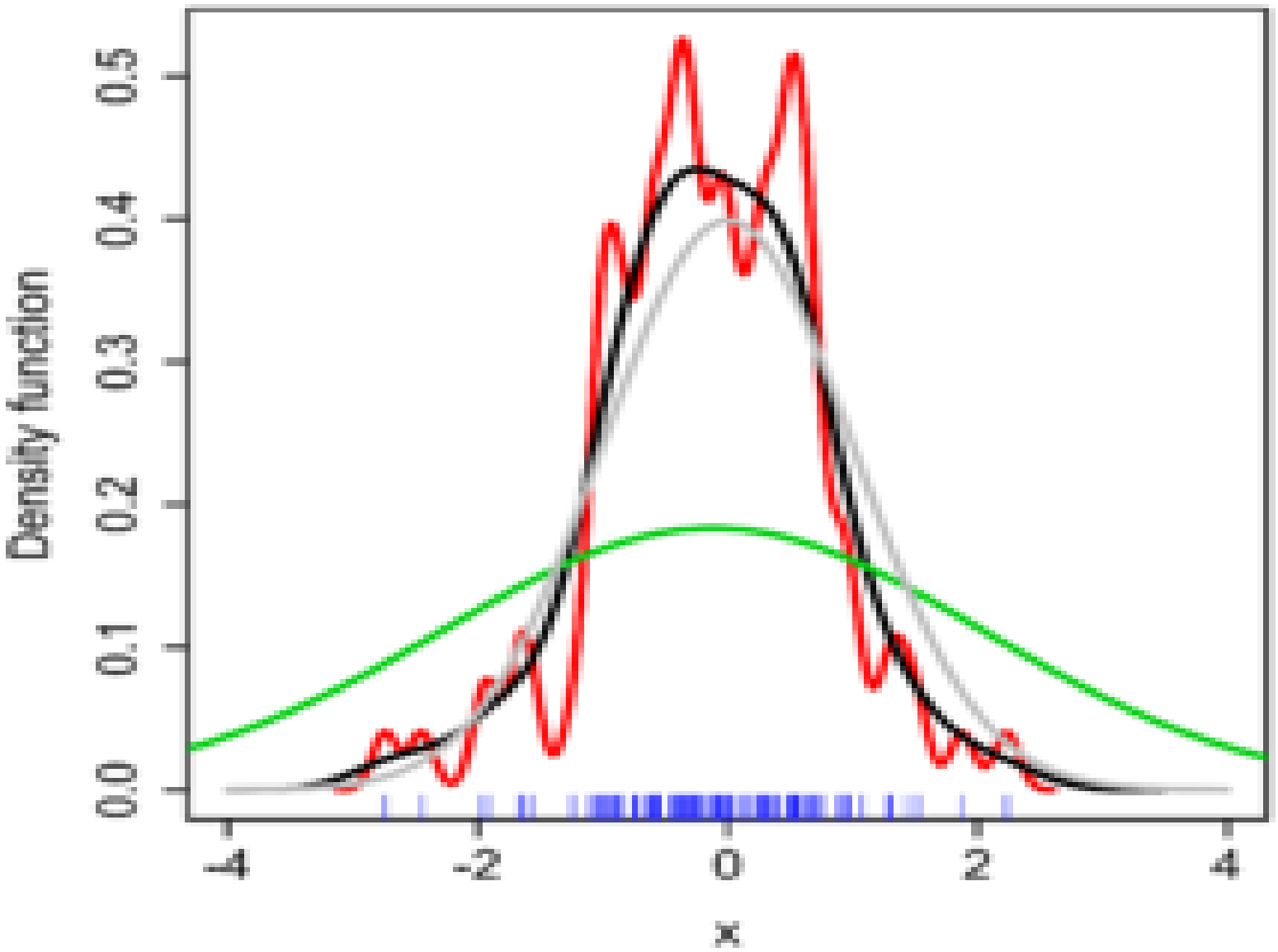
Định nghĩa tính chất Triangle Condition

Cho đồ thị G_T có chi phí phụ thuộc vào thời gian, G_T là đồ thị có tính chất FIFO thì đồ thị cũng có tính chất Triangle Condition cho mỗi cạnh (v_i, v_j) với mọi thời điểm xuất phát t, nghĩa là $arrive_{i,j}(t) < arrive_{i,j}(t + c_k)$ với c_k là tổng chi phí thời gian trì hoãn di chuyển đến đỉnh v_j khi qua đỉnh trung gian v_k.

Phương pháp luận tiếp cận bài toán

- Mô tả lại bài toán nghiên cứu trên đồ thị có dữ liệu ràng buộc thời gian như sau:
  Cho trước điểm xuất phát v_s, điểm đến v_e, thời gian khởi hành t_d không có chuyến bay thẳng, mục tiêu tìm con đường tối ưu p^* đưa giá rẻ nhất trong tất cả các chuyến đi vòng từ v_s đến v_e **xuất phát sớm nhất, về đích sớm nhất**.
- Sử dụng một giả sử (heuristic) giá vé chịu ảnh hưởng của khoảng cách địa lý, chọn các trạm trung gian nằm giữa điểm nguồn, điểm đích.

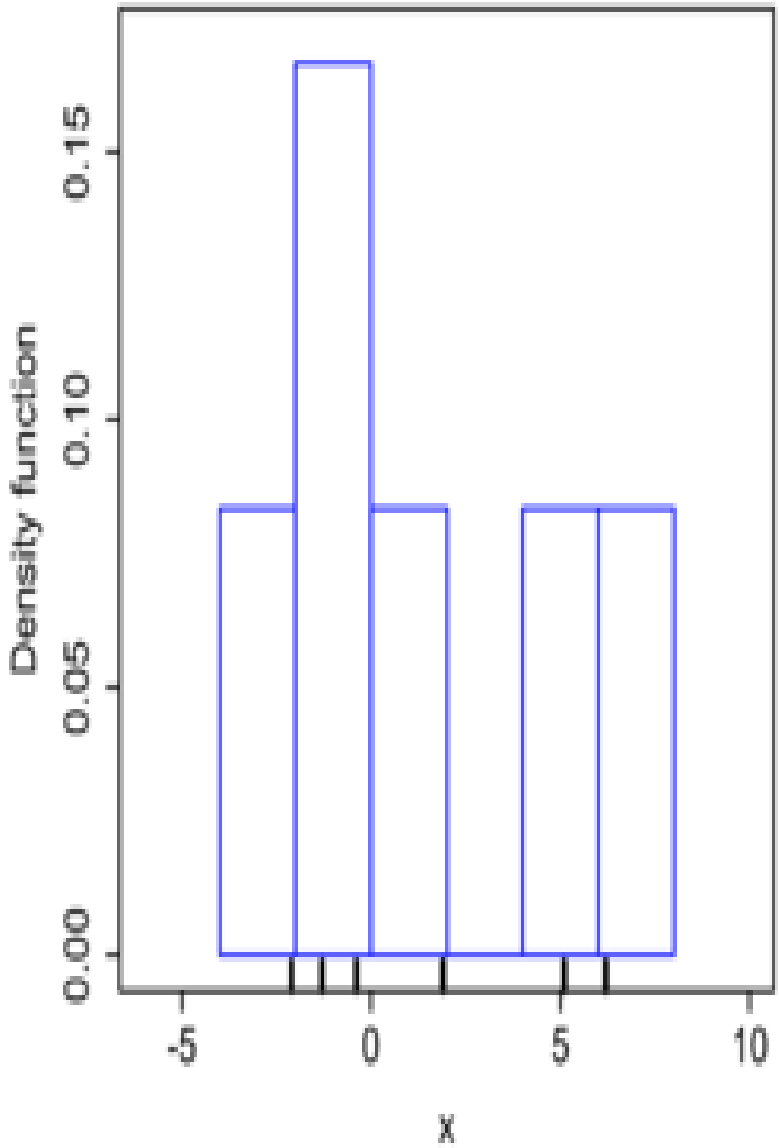
Phương pháp luận tiếp cận bài toán

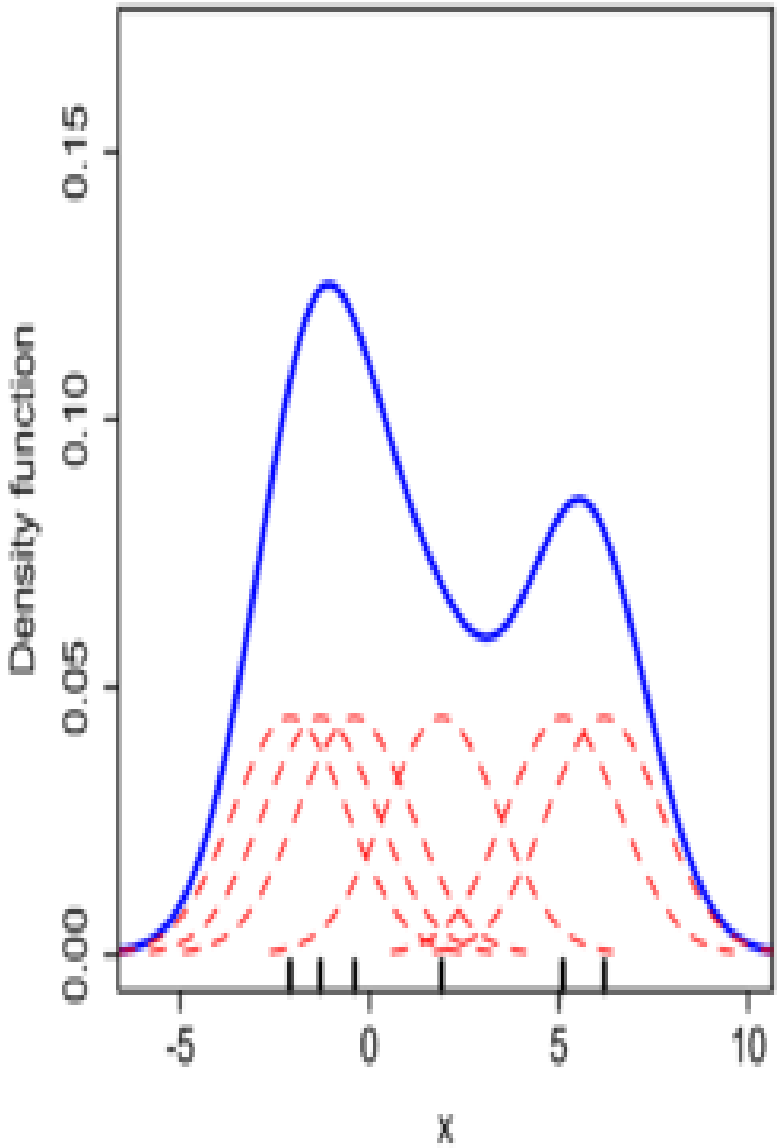


Phương pháp luận tiếp cận bài toán

- Bổ sung thêm lựa chọn cho người sử dụng
  - Đề xuất thêm chuyến bay vòng có tổng giá vé cao hơn chút xíu nhưng về đích sớm hơn rất nhiều so với chuyến bay vòng tối ưu nhất
    Hàm đánh giá lợi nhuận, cứ về sớm hơn 1 phút, tiết kiệm được 10000VND
    Gọi v^* cho tổng giá vé Ψ^* rẻ nhất, thời gian đến đích τ^*
    Gọi v_t cho tổng giá vé Ψ_t, thời gian đến đích τ_t sớm hơn τ^*
    $\Omega = \text{convertSavingTimeToBenefit}(\tau_t, \tau^*) - (\Psi_t - \Psi^*)/10000$
    với
    $\text{convertSavingTimeToBenefit}(H_1 \text{ giờ } M_1 \text{ phút}, H_2 \text{ giờ } M_2 \text{ phút}) = \lceil (H_2 * 60 + M_2 - H_1 * 60 - M_1) \rceil$
  - Nếu muốn tìm qua 2 trạm trung gian, thì bắt buộc tổng giá vé phải rẻ hơn tổng giá vé θ của chuyến bay vòng tối ưu nhất qua 1 trạm trung gian trong khoảng thời gian $[t_d, t_a]$ với t_a bằng thời gian về đích trễ hơn 7 tiếng

Phương pháp luận tiếp cận bài toán





Phương pháp đánh giá kết quả nghiên cứu

Làm thực nghiệm nhiều lần với các điểm nguồn, điểm đích, thời gian xuất phát khác nhau tại thời điểm không có chuyến bay thẳng

Kế hoạch triển khai

Kế hoạch triển khai đề tài trong 20 tuần, theo thứ tự sau

- Thu thập dữ liệu trên trang web www.abay.vn: 2 tuần.
- Tìm kiếm cấu trúc lưu trữ dữ liệu các chuyến bay thẳng trên CSDL MongoDB sao cho truy vấn không gian địa lý hay tìm kiếm nhanh nhất: 1 tuần
- Tìm kiếm biểu diễn cấu trúc dữ liệu khi nạp vào bộ nhớ để sắp xếp, tìm kiếm và chọn giá trị nhỏ nhất nhanh: 1 tuần
- Xây dựng biểu đồ lớp cho ứng dụng: 2 tuần.
- Triển khai:
  - Xây dựng mã nguồn hoàn chỉnh tìm lời giải: 1 tháng
  - Xây dựng giao diện nhập, xuất lời giải trên bản đồ dùng công nghệ Google Map: 1 tuần.
  - Triển khai chạy thực nghiệm, rút ra kết luận và cải tiến nếu làm được: 9 tuần

Nội dung dự kiến của luận văn

- Giới thiệu
  - Giới thiệu bài toán
  - Dữ liệu liên quan đến bài toán
- Lý thuyết áp dụng
  - Định nghĩa và tính chất đồ thị có chi phí phụ thuộc vào thời gian
  - Cấu trúc dữ liệu lưu trữ và tìm kiếm được sử dụng trong CSDL MongoDB và nạp vào bộ nhớ
  - Những cải tiến cho bài toán chưa được đề cập tại đề cương (nếu có)
- Hướng triển khai
  - Biểu đồ lớp
  - Mã giả thuật toán
- Trình bày kết quả, đề xuất cải tiến
  - Trình bày kết quả thu được khi chạy thực nghiệm bài toán
  - Nêu những hạn chế và đề xuất (nếu có) những mong muốn cải tiến cách giải bài toán

Kết luận

Danh mục các tài liệu kham khảo

- Người dịch: Nguyễn Văn Minh Mẫn *Thống kê Công nghiệp hiện đại với ứng dụng viết trên R, MINITAB và JMP*. Nhà xuất bản Bách Khoa Hà Nội,2016, pp.19-131.
- *Naive Bayes 3: Gaussian example*, truy cập ngày 2 tháng 3 năm 2017, địa chỉ *https://www.youtube.com/watch?v=r1in0YNetG8*.
- *Exploratory Data Analysis: Kernel Density Estimation in R on Ozone Pollution Data in New York and Ozonopolis*, truy cập ngày 26 tháng 3 năm 2017, địa chỉ *https://www.r-bloggers.com/exploratory-data-analysis-kernel-density-estimation-in-r-on-ozone-pollution-data-in-new-york-and-ozonopolis/*.

Thank you for listening
Q&A