DỰ ĐOÁN XÁC SUẤT XE BUÝT VỀ TRẠM ĐÚNG GIỜ

Lê Thị Minh Thùy

Đại Học Bách Khoa TPHCM

Ngày 8 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

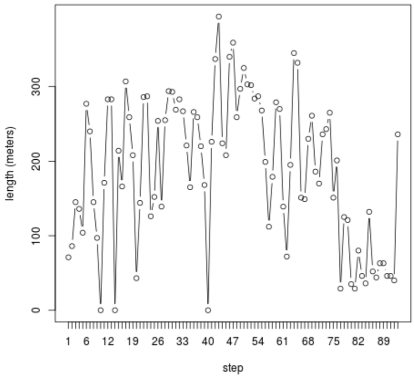
Dữ liệu sau khi được đồng bộ hóa khoảng cách thời gian hồi đáp (20 giây)

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

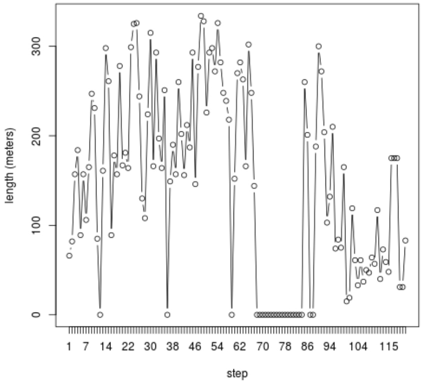
STT	Khoảng cách thời gian hồi đáp (giây)	Khoảng cách bước đi (mét)
10	20	0
11	20	171
12	20	283
13	20	283
14	20	0
15	20	214
16	20	166
17	20	307

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

Trực quan hóa giá trị các bước di chuyển trong 80% thời gian di chuyển



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
- Tần số và tần suất
- 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 - Tần số và tần suất

Xét một biến ngẫu nhiên X, nhận các giá trị $x_1, x_2, ..., x_{m-1}, x_m$ với $x_1 < x_2 < ... < x_{m-1} < x_m$
Giả sử rằng ta đã thực hiện n quan sát khác nhau trên X.
Khi đó tần số của $x_i(i = 1, ..., m)$ là số lượng quan sát nhận giá trị x_i
Tần suất f được tính

$$f_i \geq 0 \quad \text{và} \sum_{i=1,2,...,m} f_i = n$$

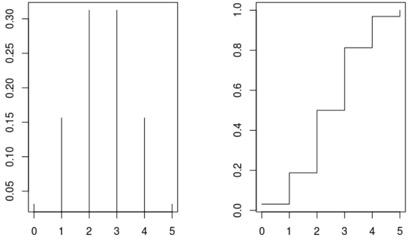
Tần suất p được tính

$$p_i = \frac{f_i}{n}(i = 1, ..., m)$$

Tần suất tích lũy là tổng các tần suất của các giá trị nhỏ hơn hoặc bằng x_i, được tính bởi

$$P_i = \sum_{j=1}^{i} p_j$$

Cơ sở lý thuyết - Tần số và tần suất



Ví dụ về tần suất và tần suất tích lũy

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:

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

Cơ sở lý thuyết - Biến ngẫu nhiên

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 $\mathrm{A} = \mathrm{X} \leq \mathrm{x}$ hay $\{X \leq x\}$. Vậy xác suất $\mathbb{P}[A]{=}\mathbb{P}[X \leq x]$ luôn tồn tại.

Cơ sở lý thuyết - Biến ngẫu nhiên rời rạc

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	...	x_{m-1}	x_m
$p_k := p(x_k) = \mathbb{P}[X = x_k]$	p_0	p_1	...	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.

Cơ sở lý thuyết - Biến ngẫu nhiên liên tục

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 trên 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

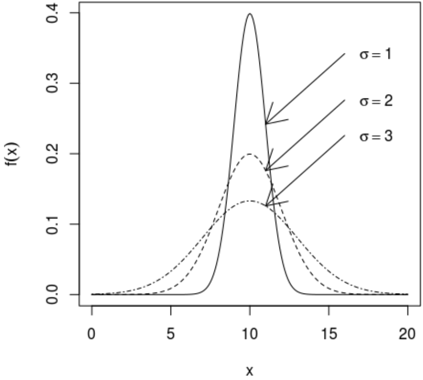
- Rời rạc hóa thuộc tính dạng con số thành dạng phân loại => Dễ tranh cãi
- Sử dụng hàm mật độ xác suất cho biến liên tục

Cơ sở lý thuyết - Phân phối chuẩn (Gauss)

Biến ngẫu nhiên X có hàm mật độ là hàm

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

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



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 $\mathrm{P_Y}(0)=\mathrm{P_Y}(1)=\frac{1}{2}$

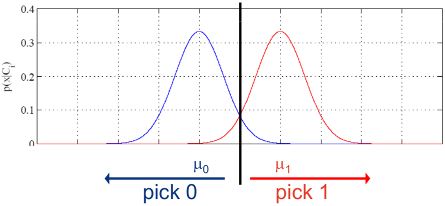
Công thức hàm phân phối Gaussian $\mathrm{G}(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} * e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

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.

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/

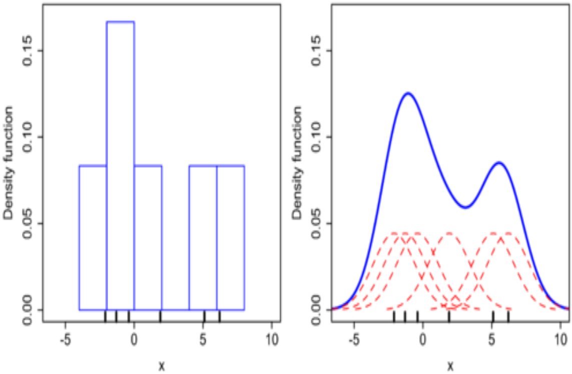
Constructing a Kernel Density Estimate: Step by Step

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

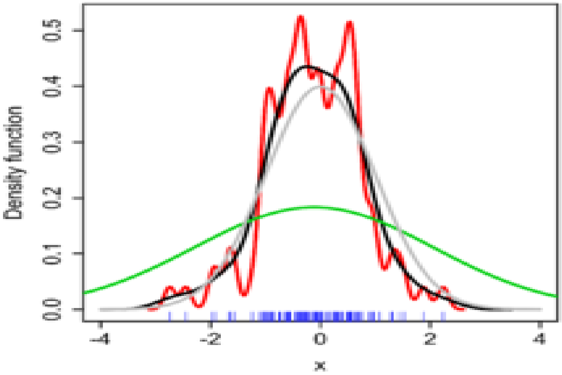


[3]https://en.wikipedia.org/wiki/Kernel_density_estimation

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

Bài toán

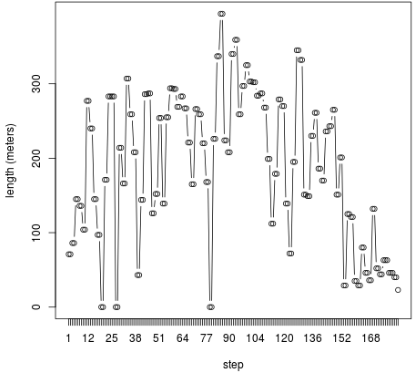
Dự đoán xác suất xe buýt tuyến 72 lộ trình xuất phát từ BX Củ Chi về trạm đích BX An Sương đúng giờ (hạn mức 45 phút)

Dữ liệu thô là tọa độ vĩ độ, kinh độ và thời điểm xuất hiện tại vị trí đó của xe buýt
Ví dụ 7 dữ liệu thô:

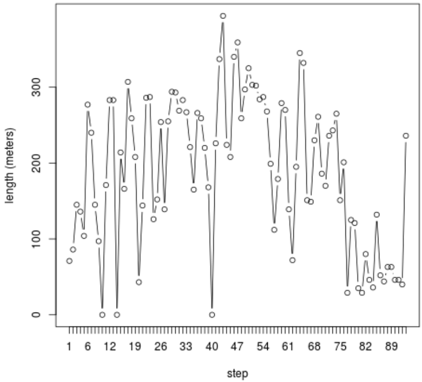
	Vĩ độ	Kinh độ	Thời điểm xuất hiện
1	10.844095	106.613688333333	2016-09-02 07:25:43
2	10.84298	106.614991666667	2016-09-02 07:26:02
3	10.8424316666667	106.615195	2016-09-02 07:26:22
4	10.8426816666667	106.615966666667	2016-09-02 07:26:42
5	10.84309	106.615203333333	2016-09-02 07:27:02
6	10.846395	106.61304	2016-09-02 07:30:48
7	10.84664	106.612861666667	2016-09-02 07:31:01

Kết Luận

Chỉ giữ lại các bước di chuyển trong 80% thời gian đầu



Toàn bộ di chuyển chuyến xe 1



80% di chuyển chuyến xe 1

Dữ liệu làm việc là 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

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

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

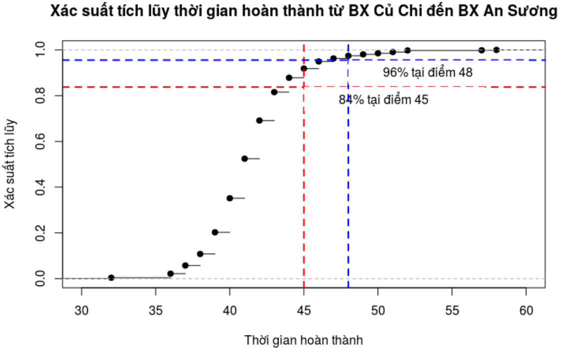
Kết Luận

Bài toán này sẽ có 5 biến:

- X_1: bước di chuyển 0-15 km/h
- X_2: bước di chuyển 15-30 km/h
- X_3: bước di chuyển 30-45 km/h
- X_4: bước di chuyển 45-60 km/h
- X_5: bước di chuyển 60 km/h

Kết Luận

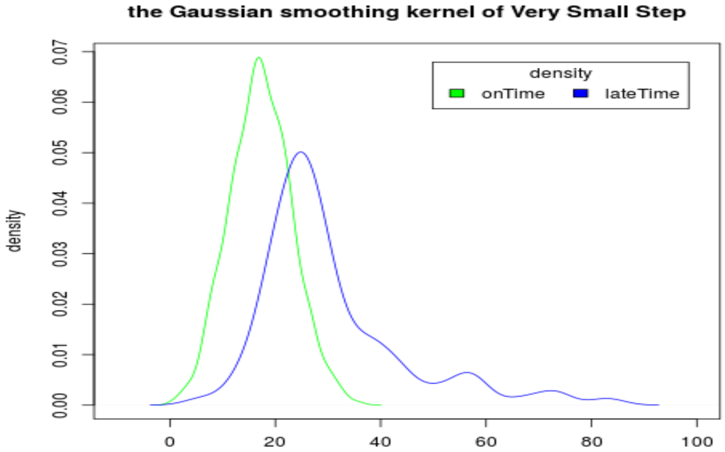
Ta chọn thời gian hoàn thành trước đúng 45 phút là về đích đúng giờ
Như vậy ta có $\mathrm{P}(\text{về đích đúng giờ}) = 0.84$ và $\mathrm{P}(\text{về đích trễ giờ}) = 0.16$



Kết Luận

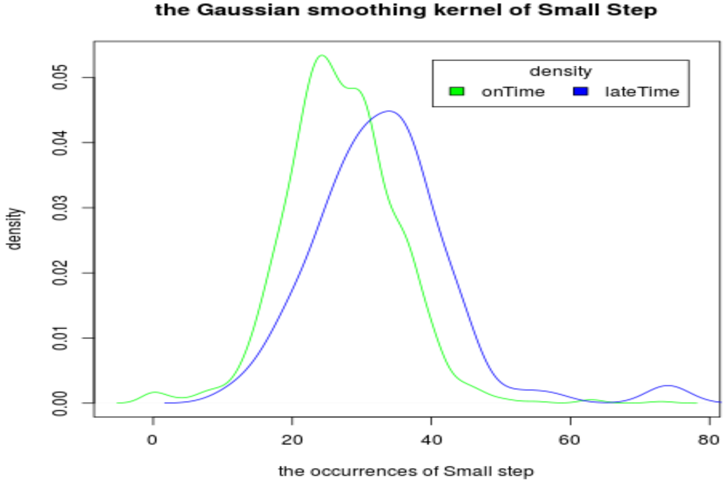
Dùng giải thuật Kernel Density Estimation để vẽ hàm mật độ xác suất của các biến dưới đây:

Hàm mật độ xác suất của biến X_1 bước đi 0-15 km/h



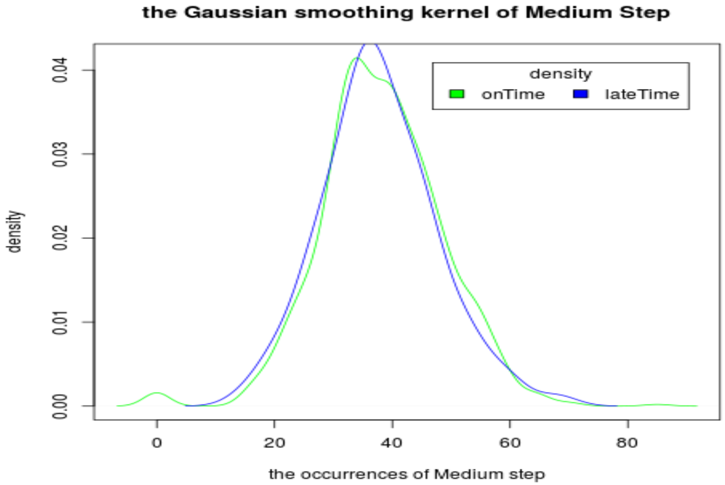
Kết Luận

Hàm mật độ xác suất của biến X_2 bước đi 15-30 km/h



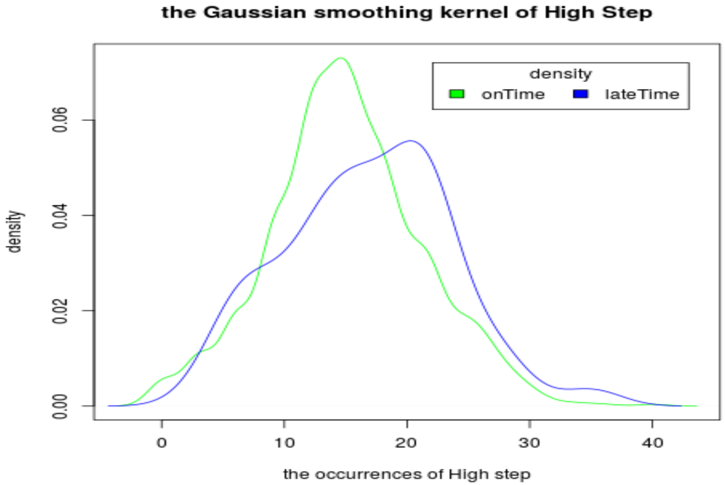
Kết Luận

Hàm mật độ xác suất của biến X_3 bước đi 30-45 km/h



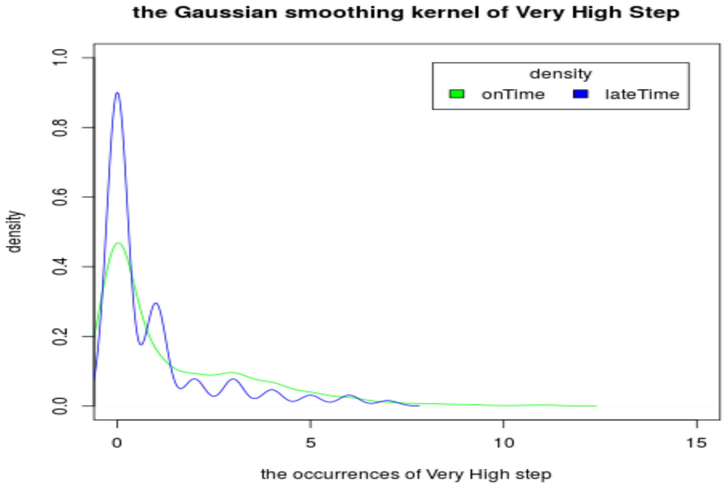
Kết Luận

Hàm mật độ xác suất của biến X_4 bước đi 45-60 km/h



Kết Luận

Hàm mật độ xác suất của biến X_5 trên 60 km/h



Kết Luận

Tìm xác suất của điểm mới

Nếu sử dụng nhân Gaussian trong thuật toán Kernel Density Estimation, ta có công thức tính xác suất của điểm mới

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

Kết luận

Áp dụng Gaussian Bayes để dự đoán xác suất xe buýt về trạm đúng giờ
Cho dữ liệu kiểm tra x có giá trị x_1, x_2, x_3, x_4, x_5

$$\mathrm{P}(onTime|x) = \frac{\mathrm{P}(x|onTime)\mathrm{P}(onTime)}{\mathrm{P}(x|onTime)\mathrm{P}(onTime) + \mathrm{P}(x|lateTime)\mathrm{P}(lateTime)}$$

Ta có $\mathrm{P}(onTime) = 0.84$ và $\mathrm{P}(lateTime) = 0.16$
Do các biến cố X_1, X_2, X_3, X_4, X_5 là các biến cố độc lập lẫn nhau cho nên nghĩa là $\mathrm{P}(x|onTime) =$
$\mathrm{P}(x_1|onTime)\mathrm{P}(x_2|onTime)\mathrm{P}(x_3|onTime)\mathrm{P}(x_4|onTime)\mathrm{P}(x_5|onTime)$
và $\mathrm{P}(x|lateTime) =$
$\mathrm{P}(x_1|lateTime)\mathrm{P}(x_2|lateTime)\mathrm{P}(x_3|lateTime)\mathrm{P}(x_4|lateTime)\mathrm{P}(x_5|lateTime)$

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.
- *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
Q&A