ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA

LÊ THỊ MINH THÙY

DỰ ĐOÁN XÁC SUẤT XE BUÝT VỀ TRẠM ĐÚNG GIỜ

Chuyên ngành: Khoa học máy tính
Mã số: 60.48.01

LUẬN VĂN THẠC SĨ

TP.Hồ Chí Minh, Ngày 14 tháng 7 năm 2017

Công trình được hoàn thành tại:
Trường Đại Học Bách Khoa - ĐHQG - TPHCM

Công trình được hoàn thành tại: **Trường Đại Học Bách Khoa - ĐHQG - TPHCM**
Cán bộ hướng dẫn khoa học: TS. Huỳnh Tường Nguyên
(Ghi rõ họ, tên, học hàm, học vị và chữ ký)

Cán bộ chấm nhận xét 1:
(Ghi rõ họ, tên, học hàm, học vị và chữ ký)

Cán bộ chấm nhận xét 2:
(Ghi rõ họ, tên, học hàm, học vị và chữ ký)

Luận văn thạc sĩ được bảo vệ tại Trường Đại học Bách Khoa, ĐHQG Tp. HCM
Ngày 14 tháng 7 năm 2017.

Thành phần đánh giá hội đồng luận văn thạc sĩ bao gồm:
1. (Chủ tịch)
2. (Thư ký)
3. (Phản biện 1)
4. (Phản biện 2)
5. (Uỷ viên)

Xác nhận của Chủ tịch Hội đồng đánh giá luận văn và Trưởng khoa quản lý chuyên ngành sau khi luận văn đã được sửa chữa (nếu có).

CHỦ TỊCH HỘI ĐỒNG	TRƯỞNG KHOA KH&KT Máy Tính
(Họ tên và chữ ký)	(Họ tên và chữ ký)

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

NHIỆM VỤ LUẬN VĂN THẠC SĨ

Họ tên học viên: Lê Thị Minh Thùy — MSHV: 13070269

Ngày, tháng, năm sinh: 22/01/1986 — Nơi sinh: Đồng Nai

Ngành: Khoa học máy tính — Mã số: 60.48.01

I. TÊN ĐỀ TÀI: Dự đoán xác suất xe buýt về trạm đúng giờ

II. NHIỆM VỤ VÀ NỘI DUNG:

1. Tìm thuật toán đơn giản, phù hợp để giải bài toán
2. Hiện thực để chứng minh thuật toán đã chọn lập trình được

III. NGÀY GIAO NHIỆM VỤ: ngày 20 tháng 6 năm 2016

IV. NGÀY HOÀN THÀNH NHIỆM VỤ: Ngày 14 tháng 7 năm 2017

V. CÁN BỘ HƯỚNG DẪN: TS. Huỳnh Tường Nguyên

Tp. HCM, Ngày 14 tháng 7 năm 2017.

CÁN BỘ HƯỚNG DẪN
(Họ tên và chữ ký)

TRƯỞNG KHOA KH&KT Máy Tính
(Họ tên và chữ ký)

LỜI NÓI ĐẦU

Để hoàn thành được Luận Văn Thạc Sĩ ngày hôm nay, tôi phải cảm ơn rất nhiều người nhưng trong khuôn khổ giới hạn không được phép trình bày hết, cho nên tôi xin được gửi lời cảm ơn ngắn gọn đến tất cả mọi người. Tôi chân thành cảm ơn quá trình 4 năm học tập chương trình Thạc Sĩ Khoa Học Máy Tính tại trường Đại Học Bách Khoa, có cơ hội trao đổi kinh nghiệm quý báu với thầy cô và các bạn cùng khóa.

Trong chương Cơ sở lý thuyết của Luận Văn Thạc Sĩ của mình, tôi có sao chép một phần nội dung trong sách "Thống kê Công nghiệp hiện đại với ứng dụng viết trên R, MINITAB và JMP" của thầy giáo Nguyễn Văn Minh Mẫn, cuốn sách thuộc bản quyền tiếng Việt của Viện Nghiên cứu cao cấp về Toán. Tôi xin phép không trả tiền bản quyền sử dụng, do Luận Văn Thạc Sĩ của tôi chỉ sử dụng cho ngày bảo vệ hội đồng Thạc Sĩ và sau đó bỏ đi. Tôi cam đoan nếu sử dụng chỉ một phần nội dung cuốn sách này cho mục đích thương mại hay bất kỳ hoạt động khoa học nào như đăng báo hay thuyết trình bất cứ đâu, tôi sẽ trả tiền bản quyền.

Ngày 14 tháng 7 năm 2017
Lê Thị Minh Thùy

TÓM TẮT LUẬN VĂN

Dùng xác suất Bayes là lý thuyết nền tảng để dự đoán xác suất xe buýt về trạm đúng giờ với dữ liệu mẫu ngẫu nhiên và số lượng giới hạn.

ABSTRACT

Bayes' theorem is used to predict the probability of a bus on the specific route arriving the destination on time with the random and limited quantity of sample data.

LỜI CAM ĐOAN

Tôi cam đoan rằng, ngoại trừ các kết quả tham khảo từ các công trình khác như đã ghi rõ trong luận văn, các công việc trình bày trong luận văn này do chính tôi thực hiện và không nội dung nào của luận văn này đã được nộp để lấy một bằng cấp ở trường này hoặc trường khác.

Ngày 14 tháng 7 năm 2017
Lê Thị Minh Thùy

Mục lục

Chương 1

GIỚI THIỆU ĐỀ TÀI

1.1 Giới thiệu đề tài

Theo nghiệp vụ xe buýt, các bác tài xế chỉ có khoảng thời gian cố định cộng thêm linh động trễ thêm vài phút để hoàn tất một lộ trình đã vạch sẵn. Khi lái xe lâu năm trên một lộ trình giống nhau, các bác tài xế khi đi được một phần đoạn đường, họ sẽ ước lượng quãng thời gian còn lại có hoàn thành kịp tiến độ cho quãng đường còn lại hay không.
Mục đích Luận Văn: dùng Thống Kê định lượng kinh nghiệm nhấm chừng này.

1.2 Động cơ

Hiện nay, có nhiều ứng dụng hay nghiên cứu khoa học khai thác thông tin GPS của các chuyến xe buýt di chuyển trên địa bàn TP. Hồ Chí Minh để rút trích thông tin có ích với mong muốn với tri thức mới tìm được trong dữ liệu thô có thể giúp ích và cải thiện dịch vụ phục vụ của phương tiện giao thông công cộng này.

1.3 Mục tiêu

Đề bài Luận Văn cụ thể: Dùng Thống Kê định lượng kinh nghiệm di chuyển trên một lộ trình quen thuộc, dự đoán xe buýt về trạm đích đúng giờ hay trễ giờ.
Tác giả không đặt mục tiêu tham vọng giải bài toán này có thể đem ra ứng dụng thực tế được vì bản thân tác giả không phải là nhà thống kê, không có nhiều kinh nghiệm làm việc với kích thước mẫu rất nhỏ có thể đưa ra dự đoán đúng trên kích thước quần thể. Nhưng thông qua Luận Văn, tác giả chứng minh được sử dụng kiến thức Thống Kê cơ bản (không chứng minh công thức Toán học vì tác giả không phải là nhà Toán học), giải được bài toán trên.
Ngoài ra, tác giả muốn nhấn mạnh về kết quả dự đoán chỉ mang tính xác suất. Ví dụ, sau khi đi được 2/3 chặng, bác tài xế nhận được kết quả xác suất 99% xe về trạm đúng giờ. Nhưng gặp sự cố tại 1/3 chặng đường cuối, tình trạng kẹt xe nặng, xe về trạm đích không đúng giờ, trái ngược hoàn toàn với con số xác suất đúng giờ "rất đẹp" 99%. Bác tài xế phải chấp nhận xác suất 1% về trạm trễ giờ vẫn có khả năng xảy ra lớn do bác tài xế không thể điều khiển tình trạng xe chạy tại 1/3 chặng đường cuối. Nếu không đặt nặng việc dự báo bắt buộc phải chính xác, các con số xác suất được thông báo vẫn có ý nghĩa chừng mực nếu bác tài xế biết rằng các con số xác suất này sử dụng những chuyến di chuyển trong quá khứ để dự báo tương lai cho chuyến đi hiện tại.
Luận Văn nhấn mạnh việc hiểu được định nghĩa xác suất, hiểu được các sự kiện xảy ra trên thực tế không bao giờ được dự đoán chắc chắn 100% xảy ra, mà chỉ dự đoán khả năng có thể xảy ra. Đến đây, tác giả mong đợi người đọc hiểu được ý nghĩa ứng dụng thực tế của Luận Văn.
Tác giả nhắc lại đề tài Luận Văn: Dùng Thống Kê định lượng kinh nghiệm di chuyển trên một lộ trình quen thuộc, dự đoán xác suất xe buýt về trạm đích đúng giờ.

1.4 Phương pháp nghiên cứu

Phương pháp nghiên cứu là thu thập số lượng dữ liệu mẫu di chuyển giới hạn, quan sát các bước di chuyển được ghi nhận cách nhau 20 giây trên 2/3 lộ trình đi từ BX Củ Chi - BX An Sương, trực quan hóa dữ liệu này, cộng thêm kiến thức lập luận thực tế ví dụ như nếu chuyến đi đó có nhiều bước di chuyển dài trong từng khoảng 20 giây thì chuyến đi đó có khả năng về trạm đúng giờ và chấp nhận xác suất, cho biết khả năng xảy ra, chấp nhận có yếu tố ngẫu nhiên xảy ra ảnh hưởng đến kết quả dự đoán. Quan sát dữ liệu cho ta khởi đầu cách tiếp cận, lấy cảm hứng kiến thức xác suất và thống kê để tìm ra phương pháp giải bài toán.

1.5 Một số kết quả thu được

Tác giả nói qua sơ lược phương pháp giải bài toán, có thể lúc này người đọc chưa hiểu hết.
Bước 1: Tuyến xe buýt 74 ngẫu nhiên được chọn và chọn lộ trình cố định BX Củ Chi - BX An Sương. Không sử dụng sức mạnh tính toán hiệu năng cao trên kích thước dữ liệu siêu lớn. Chọn mẫu dữ liệu có số lượng giới hạn và ngẫu nhiên.
Bước 2: Làm sạch dữ liệu

1. Loại bỏ những chuyến đi về trạm đích trễ một cách bất thường so với thông thường. Nghĩa là thông thường chỉ mất 45 phút để hoàn thành lộ trình, cần loại bỏ những chuyến về trạm đích quá trễ, ví dụ sau 1 tiếng, vì tìm hiểu chúng chẳng có ích gì cả, chúng xảy ra vì hiện tượng hiếm xảy nó và chúng chẳng đại diện cho số đông. Thống kê tìm những đặc điểm đúng cho số đông.

2. Do giới hạn về kỹ thuật thu nhận dữ liệu GPS sớm hơn hay trễ hơn, chuẩn hóa những bước di chuyển trong từng khoảng 20 giây.

Bước 3 Làm việc với dữ liệu, trong một chuyến đi, cắt bỏ 1/3 đoạn đường sau, chỉ giữ lại 2/3 đoạn đường trước, dán nhãn "đúng giờ" hay "trễ giờ" nhờ thời gian hoàn thành chuyến đi phân phối chuẩn. Trực quan những bước di chuyển của mỗi chuyến đi.
Quan sát và nhận thấy trên thực tế không tồn tại hai chuyến xe cùng lộ trình có những bước di chuyển cách nhau 20 giây giống nhau hoàn toàn cho nên đừng đi tìm một công thức chuẩn cho một chuyến xe về trạm đúng giờ là vô ích, mà hãy quan sát những bước nhảy trong khoảng thời gian đều 20 giây mà nhận ra được nếu chuyến đi đó có nhiều bước di chuyển dài trong từng khoảng 20 giây thì chuyến đi đó có khả năng về trạm đúng giờ.
Phân chia dữ liệu mẫu thành 2 phần: phần để học có kèm theo kết quả về trạm đích, phần để kiểm tra đã loại bỏ nhãn đúng giờ hay trễ giờ tại trạm đích.
Bước 4 Phân loại những bước di chuyển: bước đi rất nhỏ 0-15 (km/h), bước đi nhỏ 15-30 (km/h), bước đi trung bình 30-45 (km/h), bước đi xa 45-60 (km/h), bước đi rất xa trên 60 km/h.
Bước 5 Không đếm số lần xảy ra những bước chuyển rời rạc mà sử dụng Kernel Density Estimation dự đoán hàm mật độ xác suất, ký hiệu pdf (probability density function) của biến ngẫu nhiên.
Bước 6 Sử dụng công thức xác suất Bayes nền tảng $\mathbb{P}(B|A) = \frac{\mathbb{P}(B)\cdot\mathbb{P}(A|B)}{\mathbb{P}(A)}$ để giải bài toán dán nhãn.
Công thức xác suất Bayes phù hợp để dự đoán vì khi ta quan sát biến cố A xảy ra và có xác suất có điều kiện $\mathbb{P}(A|B)$ được biết trước, lý thuyết Bayes cho ta cách tính để đánh giá xác suất một sự kiện B chưa quan sát được xảy ra sau khi biến cố A đã xảy ra.
Tác giả nhắc lại đề tài Luận Văn: Dùng Thống Kê định lượng kinh nghiệm di chuyển trên một lộ trình quen thuộc, dự đoán xác suất xe buýt về trạm đích đúng giờ.

Chương 2

CƠ SỞ LÝ THUYẾT

2.1 Lý thuyết xác suất cơ bản

Tổng thể và mẫu

Kham khảo trang 19, 24, 25 chương 2 từ sách [1]

- Một **tổng thể** (còn được gọi là quần thể) thống kê là một tập các phần tử có một thuộc tính chung nhất định.
  Ví dụ, tập hợp tất cả các chuyến xe buýt từ BX Củ Chi - BX An Sương của tuyến 74 vào ngày 17 tháng 10 năm 2016 là một tổng thể hữu hạn và có thực.
  Một ví dụ khác, tập tất cả các chuyến xe buýt từ BX Củ Chi - BX An Sương của tuyến 74 có thể về trạm trễ trong điều kiện: mưa nhiều, đường ngập, hẹp, rào chắn lô cốt, quá tải xe máy xuyên suốt cả ngày và va chạm xe máy thường xuyên. Tổng thể này là vô hạn và giả định.

- Một **mẫu** là một tập hợp các phần tử trong một tổng thể nhất định. Một mẫu thường được chọn ra từ một tổng thể với mục tiêu quan sát các đặc tính của tổng thể ấy và đưa ra các quyết định thống kê có liên quan đến các đặc trưng tương ứng.
  Chẳng hạn, xét vài trăm triệu chuyến xe từ BX Củ Chi - BX An Sương của tuyến 74, công ty quản lý muốn tìm ra được đặc trưng tình trạng di chuyển 2/3 chặng đường như thế nào để cảnh báo sớm tình trạng đến đích trễ cho các bác tài xế. Nếu không sử dụng sức mạnh tính toán của điện toán đám mây, chỉ có giới hạn sức tính toán của con người, những nhà thống kê sẽ rút ra đặc trưng của vài trăm triệu dữ liệu bằng cách làm việc trên mẫu ngẫu nhiên rút ra từ vài trăm triệu dữ liệu trên. Những thủ tục lấy mẫu như vậy để đưa ra các quyết định thống kê được gọi là phương pháp lấy mẫu chấp nhận. Ngoài ra, Toán Thống Kê cung cấp phương pháp ước lượng bằng cách sử dụng các mẫu chọn ra từ các tổng thể hữu hạn, bao gồm cả việc lấy mẫu ngẫu nhiên có hoàn lại và lấy mẫu ngẫu nhiên không hoàn lại.

Như vậy, trong Toán Thống Kê, tồn tại các phương pháp như phương pháp thí nghiệm thống kê, phương pháp lấy mẫu chấp nhận, phương pháp kiểm định giả thuyết,.. để từ mẫu ngẫu nhiên đủ rút ra đặc trưng của tổng thể với xác suất xảy ra cao. Nhưng để thực hiện được điều này, nó vượt quá kiến thức kỹ sư máy tính. Cho nên Luận Văn này không đặt tham vọng, giải trên mẫu có thể kết luận trên tổng thể với xác suất xảy ra cao.
Ngoài ra, tôi có đưa thêm một số giả định trong khi xem xét mẫu được lấy ngẫu nhiên trong Luận Văn:

- Tôi tập trung vào đo lường chuyến đi trong một khoảng thời gian cụ thể, tháng 9 năm 2016, các mẫu của tôi không rải rác qua các tháng, các năm.

- Chúng ta chỉ dự đoán được một phần trong nhiều hiện tượng mà ta gặp phải. Xem xét tất cả các chuyến xe trên mọi điều kiện không thể kiểm soát được trên thực tế là quá tốn kém và không thực tế. Cho nên các yếu tố bên ngoài như thời điểm xuất phát, thời tiết, điều kiện mặt đường,... được coi là các yếu tố độc lập để giảm sự phức tạp giải bài toán.

Biến cố và không gian mẫu: Diễn tả hình thức của thí nghiệm

Kham khảo trang 52, 55, 59, 60, 61, 63 chương 3 từ sách [1]
Ta thấy cùng một tuyến đường xe buýt 74 cố định, từ BX Củ Chi - BX An Sương, các chuyến xe trong cùng tháng 9 có thời gian hoàn thành khác nhau, không biết trước được một cách chắc chắn. Nguyên nhân là do những yếu tố bên ngoài như mặt đường, thời tiết, giờ cao điểm, thấp điểm,..đều có ảnh hưởng đến kết quả. Toán Thống Kê giúp ta có phương pháp làm việc trên dữ liệu có tính ngẫu nhiên như vậy.
Không gian mẫu là tập hợp của tất cả các kết cục có thể của một thí nghiệm cụ thể. Chẳng hạn, 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 $\mathcal{S}$.
Một cách để mô tả một phân phối các giá trị mẫu, đặc biệt hữu ích trong các mẫu lớn, là xây dựng một phân phối tần số (tần suất) của các giá trị mẫu. Ta phân biệt giữa hai loại tần suất, cụ thể là tần suất của các biến rời rạc và các biến liên tục. Một biến ngẫu nhiên X được gọi là rời rạc nếu nó chỉ nhận hữu hạn hoặc vô hạn đếm các giá trị khác nhau. Ví dụ, số thẻ máy tính bị lỗi trong một hàng sản xuất là một biến ngẫu nhiên rời rạc. Một biến ngẫu nhiên được gọi là liên tục nếu về mặt lý thuyết, nó có thể nhận tất cả các giá trị có thể trong một khoảng số thực cho trước. Ví dụ, điện áp đầu ra của một nguồn điện là một biến ngẫu nhiên liên tục.

Tần suất của biến ngẫu nhiên rời rạc

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. Ta kí hiệu tần số của x_i bởi f_i. Dễ thấy rằng

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

Tập các cặp có thứ tự (x_i, f_i) tạo nên phân bố tần số hay phân phối tần số của X. Một khái niệm rất có ích là tần suất, được định nghĩa bởi

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

Hiển nhiên $p_i \geq 0$ và $\sum_{i=1,2,...,m} p_i = 1$.

Giá trị	Tần số	Tần suất
x_1	f_1	$p_1 = \frac{f_1}{n}$
x_2	f_2	$p_2 = \frac{f_2}{n}$
.		.
.		.
.		.
x_m	f_m	$p_m = \frac{f_m}{n}$
Tổng	n	1

Ngoài tần số và tần suất, tần số tích lũy của một biến cũng là một khái niệm hữu ích. Tần số tích lũy của x_i là tổng của các tần số của các giá trị nhỏ hơn và bằng x_i, được tính bằng $F_i = \sum_{j=1,2,...,i} f_i$.
Còn tần suất tích lũy là tổng các tần suất của các giá trị nhỏ hơn hoặc bằng x_i, được tính bởi

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

Giá trị	Tần suất p	Tần suất tích lũy P
x_1	p_1	$P_1 = p_1$
x_2	p_2	$P_2 = p_1 + p_2$
.		.
.		.
.		.
x_m	p_m	$P_m = p_1 + p_2 + ... + p_m = 1$
Tổng	1	

Tần suất của biến ngẫu nhiên liên tục

Với trường hợp biến ngẫu nhiên liên tục, ta phân hoạch toàn bộ tập giá trị của biến quan sát thành m khoảng con. Nói chung, nếu tập giá trị của X nằm giữa L và H, ta xác định các số

$$b_0, b_1, b_2, ..., b_{m-1}, b_m$$

sao cho L = $b_0 \leq b_1 \leq ... \leq b_{m-2} \leq b_m$ = H
Các trị $b_0, b_1, b_2, ..., b_{m-1}, b_m$ được gọi là các cận của m khoảng con
Sau đó ta phân loại các giá trị X vào khoảng (b_{i-1}, b_i) nếu ($b_{i-1} < X \leq b_i (i = 1, ..., m)$). Nếu X = b_0, ta gán nó vào khoảng con đầu tiên.

Khoảng con	Điểm giữa	Tần số	Tần suất	Tần số tích lũy	Tần suất tích lũy
$b_0 - b_1$	$\bar{b_1}$	f_1	$p_1 = \frac{f_1}{n}$	$F_1 = f_1$	$P_1 = p_1$
$b_1 - b_2$	$\bar{b_2}$	f_2	$p_2 = \frac{f_2}{n}$	$F_2 = f_1 + f_2$	$P_2 = p_1 + p_2$
.		.			
.		.			
.		.			
$b_{m-1} - b_m$	$\bar{b_m}$	f_m	$p_m = \frac{f_m}{n}$	$F_m = f_1 + ... + f_m$	$P_2 = p_1 + ... + p_m = 1$
Tổng		n	1		

Xác suất của biến cố

Thông thường chúng ta gộp tất cả các biến cố vào một tập $\mathcal{Q} := \{A : A \subset \mathcal{S}$ là một biến cố$\}$, gọi là tập các biến cố.
Xét một hàm $\mathbb{P} : \mathcal{Q} \to \mathbb{R}$ xác định trên $\mathcal{Q}$, gán cho mỗi biến cố A một số thực, ký hiệu A $\mapsto \mathbb{P}$[A], $\mathbb{P}$[A] (hay $\mathbb{P}$(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:

- **A1** Xác suất là không âm, $\mathbb{P}$(A) ≥ 0
- **A2** Không gian mẫu $\mathcal{S}$ có xác suất 1, $\mathbb{P}(\mathcal{S})$=1
- **A3** Xác suất của các biến cố rời nhau. Khi có hai biến cố A, B mà A $\cap$ B = $\emptyset$ thì

$$\mathbb{P}(A \cup B) = \mathbb{P}(A \text{ hay } B) = \mathbb{P}(A) + \mathbb{P}(B)$$

Tổng quát hơn, nếu ta có $E_1, E_2, ..., E_n$ (n ≥ 1) là các biến cố rời nhau từng đôi một thì

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

Xác suất có điều kiện và sự độc lập của các biến cố

Khi các biến cố khác nhau có liên quan, việc thực hiện một biến cố có thể cung cấp cho ta thông tin liên quan để cải thiện, nâng cao khả năng đánh giá của ta về các biến cố khác. Biến cố B đã xảy ra, tức là $\mathbb{P}[B] > 0$, xác suất biến cố A cũng xảy ra là gì? Suy nghĩ một cách tích cực, ta thấy nên thu

hẹp không gian mẫu S tới không gian trong đó B đã xảy ra (nhằm mục đích so sánh giữa A ∩ B và B)
Xác suất có điều kiện của biến cố A cho biết biến cố B đã xảy ra, $\mathbb{P}[B] > 0$ là

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

Theo đó xác suất đồng thời của hai biến cố A và B là

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

Ví dụ: Thí nghiệm đo chiều dài của một thanh thép. Không gian mẫu là $\mathcal{S}$ = (19.5, 20.5)[cm]. Hàm xác suất gán bất kỳ một khoảng tập con của S một xác suất bằng chiều dài của nó. Cho A = (19.5,20.1), nghĩa là $\mathcal{P}(A)$ = 20.1-19.6=0.6 và B = (19.8,20.5), nghĩa là $\mathcal{P}(B)$ = 20.5-19.8=0.7. Khoảng tập con chung giữa A∩B=(19.8,20.1), nghĩa là $\mathcal{P}(A \cap B)$ = 20.1-19.8=0.3.
Cho một độ dài và biết thêm điều kiện độ dài này thuộc khoảng B, và chúng ta phải đoán xem nó có thuộc về A không? Ta tính xác suất có điều kiện

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{0.3}{0.7} = 0.4286$$

Mặt khác, nếu thông tin độ dài thuộc về B không biết trước, thì với độ dài của câu hỏi trên, xác suất nó thuộc về A bằng $\mathbb{P}(A)$=0.6. Vậy có một sự khác biệt giữa xác suất có điều kiện và không điều kiện. Điểu này cho thấy hai biến cố A và B là phụ thuộc.
Hai biến cố A và B được gọi là **độc lập** nếu

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

Nói khác đi, biến cố A và B độc lập nếu sự xuất hiện của A không có liên quan theo bất kỳ cách nào đến sự xuất hiện của B: $\mathbb{P}(A|B) = \mathbb{P}(A)$ và $\mathbb{P}(B|A) = \mathbb{P}(B)$
Nếu A và B là các biến cố độc lập thì

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

hay tương đương với

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

nghĩa là xác suất hai biến cố độc lập đồng thời xảy ra bằng tích các xác suất riêng lẻ.
Tổng quát, các biến cố $A_1, A_2, ..., A_n$ là n biến cố **độc lập lẫn nhau** thì

$$\mathbb{P}\left[\bigcap_{i=1}^{n} A_j\right] = \prod_{i=1}^{n} \mathbb{P}[A_i]$$

Biến ngẫu nhiên

Kham khảo trang 62, 63 chương 3 từ sách [1]
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.
Kham khảo ví dụ trang 20 chương 2 từ [1]
Xét một thí nghiệm trong đó ta tung một đồng xu một lần. Giả sử đồng xu là cân đối và đồng chất để khả năng xuất hiện một trong hai mặt là như nhau. Hơn nữa, giả định rằng hai mặt của đồng xu được gán nhãn bởi "0" và "1". Nói chung, chúng ta không thể dự đoán chắc mặt nào sẽ hiện ra. Nếu mặt "0" xuất hiện thì chúng ta gán cho một biến X giá trị 0; nếu mặt "1" xuất hiện, ta gán cho X giá trị 1. Vì những giá trị mà X sẽ nhận được trong một chuỗi các thử nghiệm như vậy là không thể dự đoán được một cách chắc chắn, nên chúng ta gọi là X một biến ngẫu nhiên. Một ví dụ cụ thể cho chuỗi ngẫu nhiên gồm các giá trị 0, 1 được tạo ra bằng cách này là như sau: 0,1,1,0,1,0,1,1,1,1,1,1,0,1,1,1,1

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 $\mathbb{N}$)
Trường hợp hữu hạn phần tử thì ta thường ghi tập giá trị

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

Trường hợp X(.) có phạm vi vô hạn đếm được thì ta ghi

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

tập này có cùng lượng số với tập $\mathbb{N}$.
Đối với một biến ngẫu nhiên X rời rạc, ta có các khái niệm sau
Hàm mật độ xác suất là

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

$p(x)$ là xác suất mà X nhận một giá trị cụ thể x $\in \mathcal{S}_X$. Ta phải có

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

Phân phối xác suất của một biến ngẫu nhiên mô tả cách các xác suất được phân phối trên các giá trị của biến ấy. Tập giá trị $\mathcal{S}_X = \{x_0, x_1, x_2, ..., x_{m-1}, x_m\}$(còn được gọi không gian mẫu của X là hữu hạn), cho ta **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

Hàm tích lũy xác suất của X được tính bằng cách lấy tổng các xác suất của các giá trị $x_k \in \mathcal{S}_X$ mà $x_k \leq x$, đó là

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

Tham số đặc trưng của biến rời rạc X với hàm mật độ $p_k = p(x_k)$, tập giá trị $\mathcal{S}_X$:

- Kỳ vọng (hay trung bình) μ và phương sai V[X] = σ^2 của X cho bởi

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

$$\sigma^2 = \mathbb{E}[(X - \mu)^2] = \sum_{x_k \in \mathcal{S}_X} [x_k - \mu]^2 p_k$$

- Độ lệch chuẩn của biến X là $\sigma_X = \sigma = \sqrt{Var(X)}$

Biến ngẫu nhiên liên tục

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 $\mathcal{S}_X \in \mathbb{R}$. Một cách toán học thì biến X liên tục thì

- X nhận vô hạn giá trị không đếm được, $\mathcal{S}_X \subset \mathbb{R}$ và
- 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 $\mathrm{F}(x_1) \leq \mathrm{F}(x_2)$, và
– Quan hệ với f: hàm phân phối xác suất $\mathrm{F}(x)$ có đạo hàm $\frac{dF(x)}{dx} = f(x)$

Tính chất của hàm mật độ xác suất f gồm:

- f(u) ≥ 0, $\forall$u, hay đường biểu diễn f nằm trên phần dương. Hơn thế $\int_{-\infty}^{\infty} f(u)du = 1 = \mathrm{F}(x)$
- Xác suất của biến cố "a $\leq$ X $\leq$ b" là diện tích giới hạn bởi hàm mật độ, trục hoành và 2 đường thẳng u = a và u = b:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X \leq b) = \int_a^b f(u)du = F(b) - F(a)$$

  hay là

$$\mathbb{P}(X \geq b) = \int_b^{\infty} f(u)du = 1 - F(b)$$

- 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.

Công thức Bayes và ứng dụng của nó

Công thức Bayes cho ta cách thức cơ bản để đánh giá bằng chứng trong các dữ liệu liên quan đến các thông số chưa biết, hoặc một số biến cố không quan sát được. Giả sử rằng một thí nghiệm ngẫu nhiên cho kết quả trong một biến cố A (hoặc phần bù của nó), ngoài ra các kết quả này phụ thuộc vào một biến cố B mà ta không trực tiếp quan sát được, nhưng xác suất có điều kiện $\mathbb{P}$(A|B) là được biết trước.
Bayes nói biến cố quan sát được A có liên quan đến biến cố B không quan sát được thông qua các xác suất có điều kiện. Để cân nhắc bằng chứng cho thấy A có ảnh hưởng trên B, diễn đạt bởi $\mathbb{P}(B|A)$ đầu tiên chúng ta giả định một xác suất $\mathbb{P}$(B) mà được gọi là xác suất tiên nghiệm. Xác suất tiên nghiệm $\mathbb{P}(B)$ thể hiện mức độ tin tưởng của chúng ta vào sự xảy ra của biến cố B. Ta luôn có điều kiện sau, cho phép tính xác suất hậu nghiệm $\mathbb{P}(B|A)$, là xác suất B xảy ra sau khi quan sát A (nên có mẫu số $\mathbb{P}$(A))

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

Tổng quát, giả sử {B_1,...,B_m} là một phân vùng của không gian mẫu. Các biến cố B_1,...,B_m không trực tiếp quan sát hay kiểm chứng được, nhưng các xác suất có điều kiện $\mathbb{P}(A|B_i)$ là được biết trước. Giả định các xác suất tiên nghiệm là $\mathbb{P}(B_i)$, thể hiện mức độ tin tưởng của ta vào sự xảy ra của biến cố B_i. Sau khi quan sát A ta chuyển đổi các xác suất tiên nghiệm của B_i thành các xác suất hậu nghiệm $\mathbb{P}(B_i|A)$, theo trên

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

Vì {B_1,...,B_m}là một phân vùng của $\mathcal{S}$, $\mathcal{S} = \bigcup_{j=1}^{m} \mathrm{B}_j$, nên A = A$\mathcal{S}$ = $\bigcup_{j=1}^{m}$ AB_j, vậy

$$\mathbb{P}(A) = \sum_{j=1}^{m} \mathbb{P}(AB_j) = \sum_{j=1}^{m} \mathbb{P}(B_j) \cdot \mathbb{P}(A|B_j)$$

Công thức Bayes tổng quát là:

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(B_i) \cdot \mathbb{P}(A|B_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B_i) \cdot \mathbb{P}(A|B_i)}{\sum_{j=1}^{m} \mathbb{P}(B_j) \cdot \mathbb{P}(A|B_j)}$$

Công thức đánh giá kết quả phân loại dựa vào xác suất Bayes

Nhìn vào ma trận dưới đây

Actual class \ Predict class	C_1	¬ C_1
C_1	True positive (TP)	False positive (FN)
¬ C_1	False negative (FP)	True negative (TN)

Tưởng tượng một kiểm tra sự xuất hiện bệnh trên mỗi người. Nếu kết quả kiểm tra là người đó có bệnh thì phân loại là positive, ngược lại là negative. Tuy nhiên, kết quả kiểm tra của mỗi người vẫn có thể không đúng với thực tế bệnh có xuất hiện trên người đó, cho nên mới có phân loại kết quả kiểm tra như sau:

- True positive (viết tắt TP): Người đó có bệnh và kết quả kiểm tra là có bệnh, phân loại đúng
- False positive (viết tắt FP): Người đó không có bệnh nhưng kết quả kiểm tra là có bệnh, phân loại sai
- True negative (viết tắt TN): Người đó không có bệnh và kết quả kiểm tra cũng là không có bệnh, phân loại đúng
- False negative (viết tắt FN): Người đó có bệnh nhưng kết quả kiểm tra là không có bệnh, phân loại sai

Ta có một số công thức đánh giá kết quả phân loại dựa vào xác suất Bayes
Gọi P = TP + FN và N = FP + TN
Độ chính xác = $\frac{TP+TN}{P+N}$
Lỗi sai = $\frac{FP+FN}{P+N}$ = 1 - Độ chính xác

Ví dụ phân loại dựa vào xác suất Bayes

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Bảng 2.1: Bảng dữ liệu

Gọi x_i tương ứng là một dòng dữ liệu ở bảng 2.1
Cho dòng dữ liệu
x_{15} = (age = youth, income = medium, student = yes, credit_rating = fair)
Câu hỏi: x_{15} được phân loại vào *buys_computer = yes* hay *buys_computer = no*?

age	buys_computer	
	yes	no
youth	2	3
middle_aged	4	0
senior	3	2

income	buys_computer	
	yes	no
low	3	1
middle	4	2
high	2	2

credit_rating	buys_computer	
	yes	no
fair	6	2
excellent	3	3

student	buys_computer	
	yes	no
yes	6	1
no	3	4

Bảng 2.2: Bảng thông tin tóm tắt (từ bảng 2.1)

Từ bảng 2.2, ta tính thêm thông tin xác suất

$$\begin{aligned}\mathbb{P}(\text{buys\_computer} = \text{yes}) &= \frac{\sum_{\mathcal{S}=\{\text{buys\_computer = yes}\}} x_i}{\sum_{\mathcal{S}=\{\text{buys\_computer = yes, buys\_computer = no }\}} x_i} \\ &= \frac{9}{14} = 0.643 \\ \mathbb{P}(\text{buys\_computer} = \text{no}) &= \frac{\sum_{\mathcal{S}=\{\text{buys\_computer = no}\}} x_i}{\sum_{\mathcal{S}=\{\text{buys\_computer = yes, buys\_computer = no }\}} x_i} \\ &= \frac{5}{14} = 0.357\end{aligned}$$

Sử dụng xác suất có điều kiện,

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

Ta tính

$$\mathbb{P}(\text{age} = \text{youth} \mid \text{buys\_computer} = \text{yes}) = \frac{\sum_{\mathcal{S}=\{\text{age=youth}\}\cap\mathcal{S}=\{\text{buys\_computer} = \text{yes}\}} x_i}{\sum_{\mathcal{S}=\{\text{buys\_computer} = \text{yes }\}} x_i}$$
$$= \frac{2}{9} = 0.222$$
$$\mathbb{P}(\text{age} = \text{youth} \mid \text{buys\_computer} = \text{no}) = \frac{\sum_{\mathcal{S}=\{\text{age=youth}\}\cap\mathcal{S}=\{\text{buys\_computer} = \text{no}\}} x_i}{\sum_{\mathcal{S}=\{\text{buys\_computer} = \text{no }\}} x_i}$$
$$= \frac{3}{5} = 0.600$$
$$\mathbb{P}(\text{income} = \text{medium} \mid \text{buys\_computer} = \text{yes}) = \frac{\sum_{\mathcal{S}=\{\text{income} = \text{medium}\}\cap\mathcal{S}=\{\text{buys\_computer} = \text{yes}\}} x_i}{\sum_{\mathcal{S}=\{\text{buys\_computer} = \text{yes}\}} x_i}$$
$$= \frac{4}{9} = 0.444$$
$$\mathbb{P}(\text{income} = \text{medium} \mid \text{buys\_computer} = \text{no}) = \frac{\sum_{\mathcal{S}=\{\text{income} = \text{medium}\}\cap\mathcal{S}=\{\text{buys\_computer} = \text{no}\}} x_i}{\sum_{\mathcal{S}=\{\text{buys\_computer} = \text{no }\}} x_i}$$
$$= \frac{2}{5} = 0.400$$
$$\mathbb{P}(\text{student} = \text{yes} \mid \text{buys\_computer} = \text{yes}) = \frac{\sum_{\mathcal{S}=\{\text{student} = \text{yes}\}\cap\mathcal{S}=\{\text{buys\_computer} = \text{yes}\}} x_i}{\sum_{\mathcal{S}=\{\text{buys\_computer} = \text{yes}\}} x_i}$$
$$= \frac{6}{9} = 0.667$$
$$\mathbb{P}(\text{student} = \text{yes} \mid \text{buys\_computer} = \text{no}) = \frac{\sum_{\mathcal{S}=\{\text{student} = \text{yes}\}\cap\mathcal{S}=\{\text{buys\_computer} = \text{no}\}} x_i}{\sum_{\mathcal{S}=\{\text{buys\_computer} = \text{no }\}} x_i}$$
$$= \frac{1}{5} = 0.200$$
$$\mathbb{P}(\text{credit\_rating} = \text{fair} \mid \text{buys\_computer} = \text{yes}) = \frac{\sum_{\mathcal{S}=\{\text{credit\_rating} = \text{fair}\}\cap\mathcal{S}=\{\text{buys\_computer} = \text{yes}\}} x_i}{\sum_{\mathcal{S}=\{\text{buys\_computer} = \text{yes }\}} x_i}$$
$$= \frac{6}{9} = 0.667$$
$$\mathbb{P}(\text{credit\_rating} = \text{fair} \mid \text{buys\_computer} = \text{no}) = \frac{\sum_{\mathcal{S}=\{\text{credit\_rating} = \text{fair}\}\cap\mathcal{S}=\{\text{buys\_computer} = \text{no}\}} x_i}{\sum_{\mathcal{S}=\{\text{buys\_computer} = \text{no }\}} x_i}$$
$$= \frac{2}{5} = 0.400$$

Sử dụng các công thứ sau

$$\mathbb{P}(AB) = \mathbb{P}(A \cap B) = \mathbb{P}(B) \cdot \mathbb{P}(A|B)$$
$$\mathbb{P}\left[\bigcap_{i=1}^{n} A_j\right] = \prod_{i=1}^{n} \mathbb{P}(A_j)$$

Ta tính
Do các biến cố age, income, student, credit_rating là các biến cố độc lập với nhau, cho nên
$\mathbb{P}((\text{age} = \text{youth}, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit\_rating} = \text{fair})|\text{buys\_computer} = \text{yes})$
$= \mathbb{P}(\text{age} = \text{youth} \mid \text{buys\_computer} = \text{yes}) \times \mathbb{P}(\text{income} = \text{medium} \mid \text{buys\_computer} = \text{yes})$
$\times \mathbb{P}(\text{student} = \text{yes} \mid \text{buys\_computer} = \text{yes}) \times \mathbb{P}(\text{credit\_rating} = \text{fair} \mid \text{buys\_computer} = \text{yes})$
$= 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044.$
$\mathbb{P}((\text{age} = \text{youth}, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit\_rating} = \text{fair})|\text{buys\_computer} = \text{no})$
$= \mathbb{P}(\text{age} = \text{youth} \mid \text{buys\_computer} = \text{no}) \times \mathbb{P}(\text{income} = \text{medium} \mid \text{buys\_computer} = \text{no})$
$\times \mathbb{P}(\text{student} = \text{yes} \mid \text{buys\_computer} = \text{no}) \times \mathbb{P}(\text{credit\_rating} = \text{fair} \mid \text{buys\_computer} = \text{no})$
$= 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019.$
Cuối cùng, ta tính

$\mathbb{P}$((age = youth, income = medium, student = yes, credit_rating = fair, buys_computer = yes))
= $\mathbb{P}$((age = youth, income = medium, student = yes, credit_rating = fair)|buys_computer = yes)
× $\mathbb{P}$(buys_computer = yes)
= 0.044 × 0.643 = 0.028
$\mathbb{P}$((age = youth, income = medium, student = yes, credit_rating = fair, buys_computer = no))
= $\mathbb{P}$((age = youth, income = medium, student = yes, credit_rating = fair)|buys_computer = no)
× $\mathbb{P}$(buys_computer = no)
= 0.019 × 0.357 = 0.007
Ta thấy xác suất biến cố (age = youth, income = medium, student = yes, credit_rating = fair,buys_computer = yes) xảy ra cao hơn so với xác suất biến cố (age = youth, income = medium, student = yes, credit_rating = fair, buys_computer = no). Do đó x_{15} = (age = youth, income = medium, student = yes, credit_rating = fair) được phân loại vào *buys_computer = yes*

Vấn đề thực tế khi áp dụng xác suất Bayes

Kham khảo [3], khi khảo sát thực tế, biến hoặc giá trị của biến 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. Ví dụ như trong bài toán Luận Văn này, xem xét những bước di chuyển trong khoảng thời gian đều 20 giây trong 80% tổng thời gian di chuyển để phán đoán xe có về đích đúng giờ hay không. Như vậy, biến trong bài toán này chính là những bước di chuyển chưa được phân loại, ví dụ phân loại thành những bước di chuyển rất nhỏ, nhỏ, trung bình, cao, rất cao. Đồng thời, sau khi phân loại thành các bước di chuyển thì số lần lặp lại những bước di chuyển đó đều là các con số liên tục, chưa được phân loại, ví dụ phân loại số lần lặp rất nhỏ, nhỏ, trung bình, lớn, rất lớn. Như vậy, trên thực tế, biến hoặc giá trị của biến rất hiếm khi được phân loại sẵn để thực hiện giải tương tự như ví dụ trên.
Bảng số liệu sau mô tả giá trị của biến trong bài toán Luận Văn được định lượng là các con số liên tục.

Số lần thực hiện bước đi từ 0-15 (km/h)	Số lần thực hiện bước đi từ 15-30 (km/h)	Số lần thực hiện bước đi từ 30-45 (km/h)	Số lần thực hiện bước đi từ 45-60 (km/h)	Số lần thực hiện bước đi trên 60 km/h	Trạng thái đến đích
19	23	19	29	3	đúng giờ
16	14	40	21	0	đúng giờ
19	23	48	13	0	đúng giờ
14	27	43	12	0	đúng giờ
13	33	44	9	0	đúng giờ
45	26	25	25	0	trễ giờ
56	26	27	16	0	trễ giờ
29	22	41	22	1	trễ giờ
83	11	15	12	5	trễ giờ
28	32	36	21	0	trễ giờ

Có hai hướng giải quyết đối với thuộc tính ở dạng con số để cuối cùng vẫn áp dụng được giải thuật xác suất Bayes $\mathbb{P}(B|A) = \frac{\mathbb{P}(B)\cdot\mathbb{P}(A|B)}{\mathbb{P}(A)}$

1. Cách đơn giản nhất là rời rạc hóa thuộc tính dạng con số thành dạng phân loại. Tuy nhiên, sẽ có những tranh cãi khi phân loại, ví dụ với con số 29 lần xuất hiện nên phân loại là xuất hiện trung bình hay nhiều lần. Đồng thời, khi phân loại không tốt, có xảy ra hiện tượng hai dòng có cùng mẫu giống nhau nhưng kết quả cuối lại khác nhau. Ví dụ, cùng mẫu low,medium,medium,low,very low tượng trưng với dãy số 15,30,40,15,0 và 24,31,45,12,0 cho hai kết quả đúng giờ và trễ giờ

Số lần thực hiện bước đi từ 0-15 (km/h)	Số lần thực hiện bước đi từ 15-30 (km/h)	Số lần thực hiện bước đi từ 30-45 (km/h)	Số lần thực hiện bước đi từ 45-60 (km/h)	Số lần thực hiện bước đi trên 60 km/h	Trạng thái đến đích
15 (low)	30 (medium)	40 (medium)	15 (low)	0 (very low)	đúng giờ
24 (low)	31 (medium)	45 (medium)	12 (low)	0 (very low)	trễ giờ

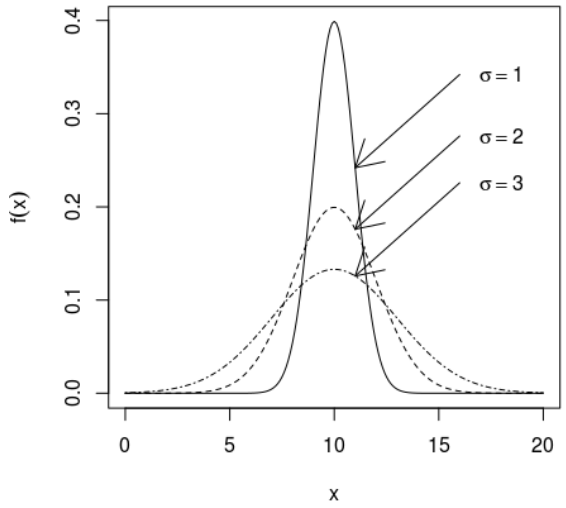
2. Sử dụng hàm mật độ xác suất (probability density function pdf) cho biến liên tục. Thông thường mọi người chọn giá trị liên tục của biến đầu vào có hàm mật độ xác suất tuân theo phân phối chuẩn hoặc phân phối Gaussian. Bạn vẫn có thể chọn phân phối khác, ví dụ như phân phối Poisson, phân phối Logarithmic,.., nếu như dữ liệu của bạn tuân theo phân phối đó.

2.2 Phân phối chuẩn (Gauss)

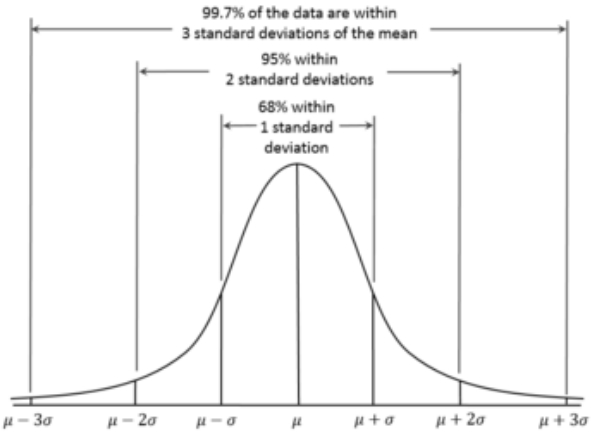
Kham khảo trang 87 đến trang 92 từ sách [1]
Phân phối Gauss (hay phân phối chuẩn) có ký hiệu là $\mathbf{N}(\mu, \sigma^2)$ chiếm một vai trò trung tâm trong lý thuyết thống kê. Hàm mật độ của $\mathbf{N}(\mu, \sigma^2)$ được cho bởi công thức

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

Biến ngẫu nhiên X có hàm mật độ là hàm Gauss $f(x)$, xem hình dưới đây, thì ta nói X có phân phối chuẩn. Ký hiệu $X \sim \mathbf{N}(\mu, \sigma)$



Hình 2.1: Hàm mật độ của $\mathbf{N}(\mu, \sigma)$ với $\mu = 10, \sigma = 1, 2, 3$



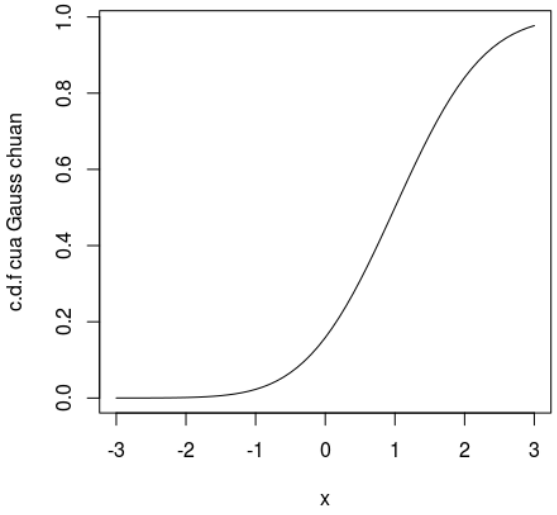
Hình 2.2: Hàm mật độ của $\mathbf{N}(\mu, \sigma)$

Đồ thị của $f(x)$ là một đường cong hình chuông đối xứng qua đường thẳng $X = \mu$. Sự phân phối của hàm mật độ được xác định bởi phương sai σ^2, theo nghĩa là hầu hết diện tích đường cong $f(x)$ (rõ hơn là 99.7% diện tích) ở giữa hai đường thẳng $\mu-3\sigma$, $\mu+3\sigma$. Khi chúng ta di chuyển xa khỏi giá trị trung bình trong cả hai hướng, đường cong chuẩn tiếp cận trục hoành.
Biến ngẫu nhiên Z tuân theo phân phối Gauss chuẩn tắc khi hàm mật độ của Z là hàm số Gauss

$$f(x) = n(x, 0, 1) = \frac{1}{\sqrt{2\pi}} * e^{-\frac{(x)^2}{2}}$$

Ký hiệu $Z \sim \mathbf{N}(0, 1)$. Giá trị kỳ vọng và phương sai của Z tương ứng là $\mathbf{E}(Z) = 0$ và $\mathbf{V}(Z) = 1$. Hàm phân phối của biến $Z \sim \mathbf{N}(0, 1)$ là

$$\Phi(x) = F(x) = \mathbb{P}[Z \leq x] = \int_{-\infty}^{x} f(t)d(t)$$



Hình 2.3: Hàm tích lũy xác suất $\Phi(x)$ của phân phối Gauss chuẩn

Hàm $\Phi(x)$ hình trên còn được gọi là hàm phân phối Gauss chuẩn tắc, cho ta diện tích khu vực trên trục x, bên dưới đường cong hàm mật độ của (Z) và bên trái của giá trị x.

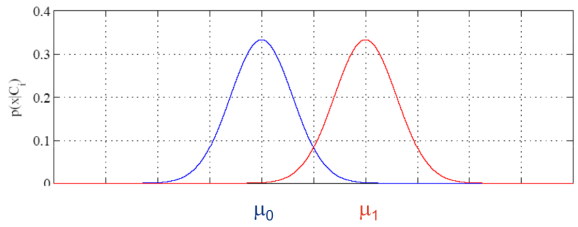
Việc tính diện tích (xác suất) phía dưới đường cong $f(x)$ của một biến $X \sim \mathbf{N}(\mu, \sigma^2)$ cho mỗi cặp tham số μ và σ là không thực tế. Ta có thể sử dụng biến chuẩn hóa - phép biến đổi Z của X, cho bởi

$$Z = \frac{X - \mu}{\sigma}$$

Khi đó $Z \sim \mathbf{N}(0, 1)$.

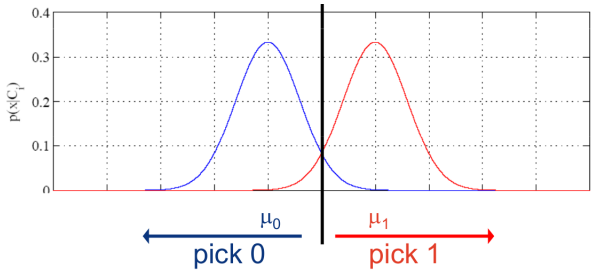
2.3 Sơ lược về phân loại Gaussian Bayes

Định nghĩa không chính thống:
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 $\mathrm{G}(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} * e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
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
$\mathrm{P_{X|Y}}(x|0) = \mathrm{G}(x, \mu_0, \sigma_0)$
$\mathrm{P_{X|Y}}(x|1) = \mathrm{G}(x, \mu_1, \sigma_1)$
Hình vẽ trực quan



Hình 2.4: Biến X có hàm mật độ phân phối Gaussian khác nhau theo mỗi phân loại

Hình vẽ trên cho ta nhận xét, 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.



Hình 2.5: 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

Ví dụ về phân loại Gaussian Bayes
Bài toán phân loại đứa trẻ (viết tắt c) hay người lớn (viết tắt a) dựa vào hai biến chiều cao[cm] (viết tắt h), cân nặng[kg] (viết tắt w). Có 4 dữ liệu adult và 12 dữ liệu child.
Ta có $\mathrm{P}(a) = \frac{4}{4+12} = 0.25$, $\mathrm{P}(c) = \frac{12}{4+12} = 0.75$
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 giả thiết, biến chiều cao và cân nặng độc lập nhau. Lập phân phối Gaussian cho từng biến thuộc từng phân loại

1. Hàm mật độ phân phối Gaussian cho phân loại người lớn

(a) Biến chiều cao có phân phối Gaussian $G(\mu_{h,a}, \sigma^2_{h,a})$ với $\mu_{h,a} = \frac{\sum_{i:y_i=a} h_i}{4}$, $\sigma^2_{h,a} = \frac{\sum_{i:y_i=a}(h_i-\mu_{h,a})^2}{4}$

(b) Biến cân nặng có phân phối Gaussian $G(\mu_{w,a}, \sigma^2_{w,a})$ với $\mu_{w,a} = \frac{\sum_{i:y_i=a} w_i}{4}$, $\sigma^2_{w,a} = \frac{\sum_{i:y_i=a}(w_i-\mu_{w,a})^2}{4}$

2. Hàm mật độ phân phối Gaussian cho phân loại đứa trẻ

(a) Biến chiều cao có phân phối Gaussian $G(\mu_{h,c}, \sigma^2_{h,c})$ với $\mu_{h,c} = \frac{\sum_{i:y_i=c} h_i}{4}$, $\sigma^2_{h,c} = \frac{\sum_{i:y_i=c}(h_i-\mu_{h,c})^2}{4}$

(b) Biến cân nặng có phân phối Gaussian $G(\mu_{w,c}, \sigma^2_{w,c})$ với $\mu_{w,c} = \frac{\sum_{i:y_i=c} w_i}{4}$, $\sigma^2_{w,c} = \frac{\sum_{i:y_i=c}(w_i-\mu_{w,c})^2}{4}$

Cho dữ liệu kiểm tra x có giá trị chiều cao x_h, giá trị cân nặng x_w. Hỏi nên phân loại x vào đứa trẻ hay người lớn?
Để trả lời câu hỏi trên, ta cần so sánh hai xác suất $P(a|x)$ và $P(c|x)$, xác suất nào lớn hơn thì phân loại x vào lớp đó.
Ta có: $P(a|x) = \frac{P(x|a)P(a)}{P(x|a)P(a)+P(x|c)P(c)}$ và $P(c|x) = \frac{P(x|c)P(c)}{P(x|a)P(a)+P(x|c)P(c)}$
Vì mẫu số giống nhau, ta chỉ cần so sánh tử số $P(x|a)P(a)$ và $P(x|c)P(c)$
Ta đã có $P(a) = 0.25$, $P(c) = 0.75$.
Ta cần tính
$P(x|a)$=$P(h_x|a)P(w_x|a)$ với $P(h_x|a) = \frac{1}{\sqrt{2\pi\sigma^2_{h,a}}} * e^{-\frac{(h_x-\mu_{h,a})^2}{2\sigma^2_{h,a}}}$, $P(w_x|a) = \frac{1}{\sqrt{2\pi\sigma^2_{w,a}}} * e^{-\frac{(w_x-\mu_{w,a})^2}{2\sigma^2_{w,a}}}$
và $P(x|c)$=$P(h_x|c)P(w_x|c)$ với $P(h_x|c) = \frac{1}{\sqrt{2\pi\sigma^2_{h,c}}} * e^{-\frac{(h_x-\mu_{h,c})^2}{2\sigma^2_{h,c}}}$, $P(w_x|c) = \frac{1}{\sqrt{2\pi\sigma^2_{w,c}}} * e^{-\frac{(w_x-\mu_{w,c})^2}{2\sigma^2_{w,c}}}$

2.4 Thuật toán Kernel Density Estimation

Định nghĩa hàm Kernel:
Hàm kernel có các thuộc tính giống hàm mật độ xác suất (pdf) với một thuộc tính được cộng thêm là hàm kernel bắt buộc phải hàm chẵn (even function), nghĩa là với mọi giá trị x và -x trong miền trục hoành của f, ta có: $f(x) = f(-x)$. Nói cách khác, hàm chẵn đối xứng qua trục tung y, ví dụ như $|x|$, x^2, $\cos(x)$. Với sự thuận tiện trong Toán học, phân phối chuẩn thường được dùng là hàm kernel.
Định nghĩa thuật toán Kernel Density Estimation (KDE):
Trong thống kê, thuật toán Kernel Density Estimation (KDE) là phương pháp không tham số (non-parametric method) dự đoán hàm mật độ xác suất pdf của biến ngẫu nhiên liên tục. KDE là phương pháp cơ bản để làm trơn dữ liệu trên mẫu dữ liệu giới hạn để dựa vào đó suy luận tổng thể.
Phương pháp không tham số:
Không tham số không có nghĩa mô hình phân phối thiếu tham số mà tham số (ví dụ như trung bình, độ lệch chuẩn) và số lượng tham số không được xác định trước mà thay đổi và được xác định bởi dữ liệu. Ngoài ra, nó cũng là phương pháp không giả định dữ liệu thuộc về phân phối cụ thể được biết trước mà được quyết định được dữ liệu.
Từng bước thực hiện thuật toán Kernel Density Estimation (KDE):
Những bước tạo thành KDE:

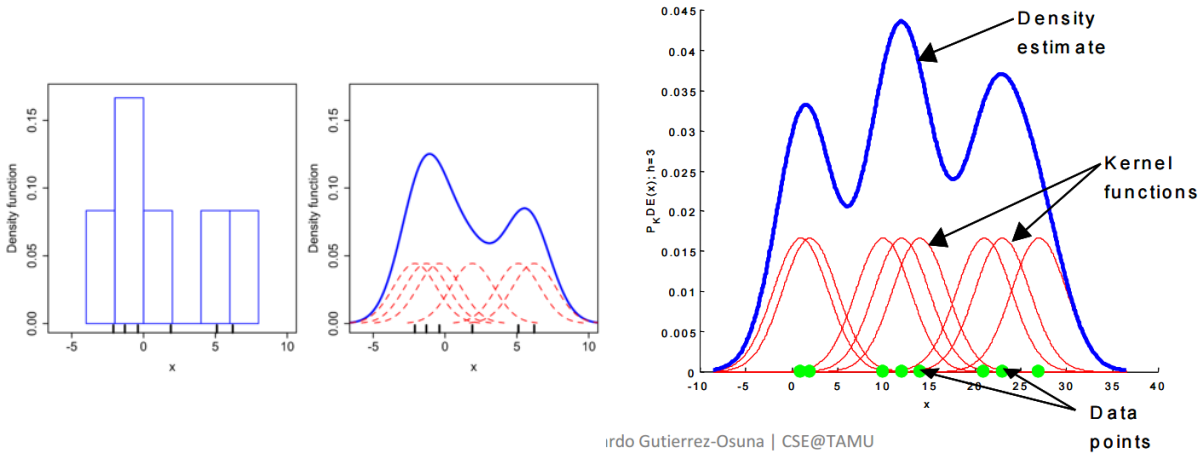
1. Chọn hàm kernel, thường chọn phân phối chuẩn (Gaussian), phân phối đều (hình chữ nhật hay hình tam giác)

2. Tại mỗi điểm dữ liệu x_i, xây dựng hàm kernel thu nhỏ (scaled kernel) $K_h = \frac{1}{h}K[\frac{(x-x_i)}{h}]$, trong đó K là hàm kernel được chọn, tham số h được gọi là bandwidth, hoặc tham số làm trơn.

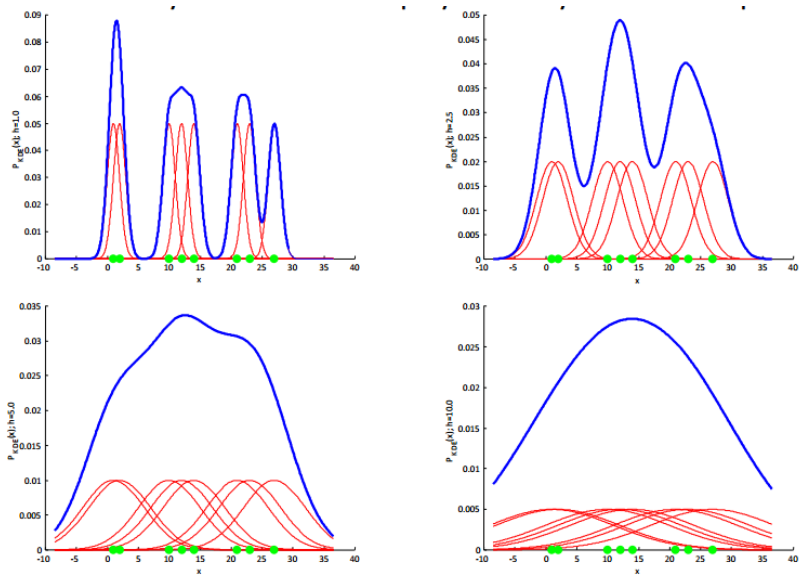
3. Cộng tất cả các hàm kernel thu nhỏ và chia cho n, điều này thể hiện xác suất $\frac{1}{n}$ đối với mỗi x_i. Đồng thời, điều này đảm bảo rằng tích phân KDE trên miền giá trị x bằng 1.

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

Chú ý bước 2, tại mỗi điểm dữ liệu, hàm kernel thu nhỏ được tạo ra sao cho mỗi điểm dữ liệu là trung tâm của hàm, điều này bảo đảm hàm kernel đối xứng qua điểm dữ liệu. Hàm pdf dự đoán bằng cách cộng các hàm kernel thu nhỏ sao đó chia cho số lượng dữ liệu để bảo đảm hai tính chất của hàm mật độ xác suất pdf $\hat{p}_{KDE}$(x), cụ thể: $\hat{p}_{KDE}$(x) > 0 và $\int_{-\infty}^{\infty} \hat{p}_{KDE}(x)dx = 1$.
Ví dụ để làm rõ điều trên, ta nhìn hình vẽ dưới đây:

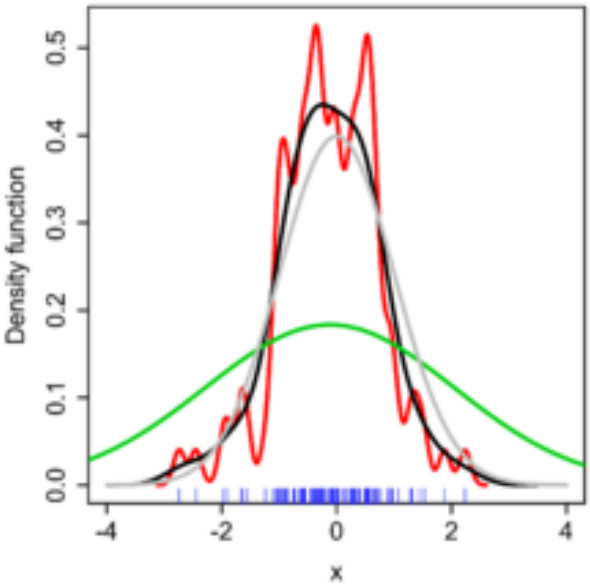


Xét tại 6 điểm dữ liệu x_1 = -2.1, x_2 = -1.3, x_3 = -0.4, x_4 = 1.9, x_5 = 5.1, x_6 = 6. Đối với histogram (hình vẽ bên trái), sẽ có 6 khoảng nhỏ, mỗi khoảng có kích thước chiều rộng 2, chiều cao tùy thuộc vào số lượng điểm dữ liệu rơi vào khoảng đó, cứ mỗi điểm rơi vào thì khoảng có có chiều cao cộng thêm 1/12. Đối với KDE (hình vẽ bên phải), ta sử dụng phân phối chuẩn với độ sai lệch (variance 2.25) trên mỗi điểm dữ liệu x_i (được thể hiện đường màu đỏ). Cộng lại các kernel này tạo ra KDE (được thể hiện đường màu xanh). KDE khá giống như histogram, ngoại trừ có thêm tính liên tục hay tính trơn nhờ sử dụng kernel thích hợp.
Chọn băng thông:
Việc chọn bandwidth là một việc khó để giải thuật KDE thể hiện tốt phân phối của biến, và sẽ không viết chi tiết ở đây. Chọn bandwidth h lớn thì làm trơn quá và che phủ cấu trúc dữ liệu, trong khi bandwidth h nhỏ thì hình dạng phân phối có nhiều gai nhọn và rất khó để thông dịch. Như hình dưới đây:



Bandwidth của kernel là tham số tự do, nhưng có ảnh hưởng lớn đến hiệu quả dự đoán. Ví dụ, lấy mẫu ngẫu nhiên (những đường vạch màu xanh da trời trên trục ngang) biết trước có phân phối chuẩn thể hiện màu xám có trung bình mẫu bằng 0 và độ sai lệch bằng 1 (biểu diễn bởi đường xám). Với

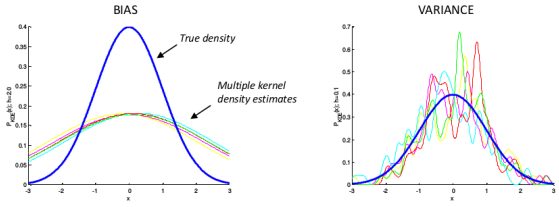
bandwidth h = 0.05 có kernel thể hiện đường cong màu đỏ, hoặc với bandwidth h = 2 có kernel thể hiện màu xanh lá đều dự đoán lệch quá xa so với kết quả biết trước, gọi đường kernel màu đỏ làm trơn quá mức (over smooth) và gọi đường kernel màu xanh lá làm trơn dưới mức (under smooth), trong khi với bandwidth h = 0.337 có kernel thể hiện đường màu đen có dự đoán gần giống với kết quả, được gọi là làm trơn tối ưu (optimal smooth).



MISE (mean integrated squared error) là tiêu chuẩn thông thường sử dụng để lựa chọn tham số bandwidth:

$$MSIE = E[\int (\hat{f}(x) - f(x))^2 dx] = E[\int (\hat{p}_{KDE}(x) - p(x))^2 dx] = \underbrace{E[\int (\hat{p}_{KDE}(x) - p(x))dx]^2}_{\text{bias}} + \underbrace{var(\hat{p}_{KDE}(x))}_{\text{variance}}$$

Bias còn được gọi là lỗi hệ thống, lỗi được gây khi tiến hành nghiên cứu ví dụ như chọn hay loại bỏ đối tượng được nghiên cứu, do lỗi thiết bị hay đo đạc, phân loại sai, hoặc thậm chí cách thức được chọn nghiên cứu không đúng hướng ngay từ đầu.
Variance được gọi là lỗi mang tính chất ngẫu nhiên, không dự đoán và không biết trước để tránh và có thể không lặp lại nếu tiếp tục thí nghiệm một lần nữa, khác với bias, lỗi sẽ lặp lại như vậy nếu vẫn tiến hành theo cách cũ. Chọn bandwidth h tốt là làm nhỏ nhất giá trị MISE.
Khi chọn bandwidth h lớn làm giảm lỗi ngẫu nhiên, nhưng làm tăng lỗi hệ thống. Trong khi bandwidth h nhỏ làm giảm lỗi hệ thống, nhưng làm tăng lỗi ngẫu nhiên. Hai hình vẽ sau thể hiện điều này



Nếu ta giả sử dữ liệu nghiên cứu tuân theo phân phối chuẩn và ta chọn Gaussian làm kernel cho giải thuật KDE, thì ta có thể biết giá trị tối ưu của h là $h^1 = 1.06\sigma N^{-\frac{1}{5}}$.

Chương 3

HIỆN THỰC VÀ THỬ NGHIỆM

3.1 Dữ liệu nghiên cứu

3.1.1 Dữ liệu thô

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

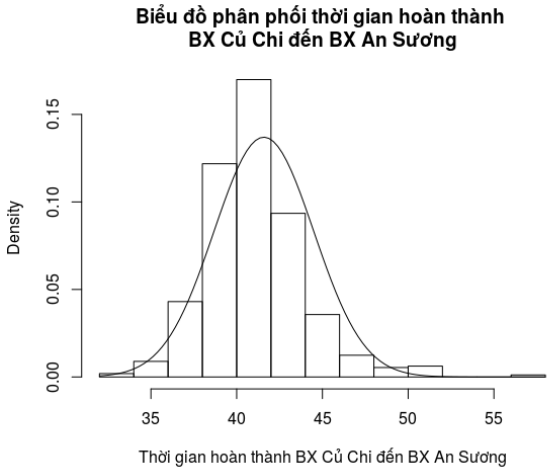
3.1.2 Tiền xử lý dữ liệu

Kiểm tra lộ trình xe buýt

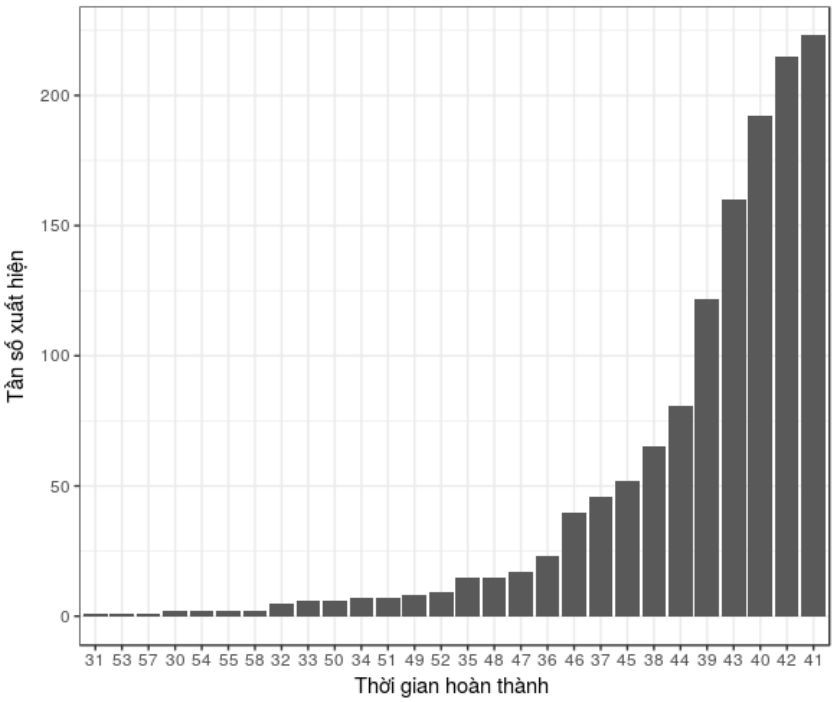
Kham khảo **PL10** tại mục Phụ Lục để lọc ra và kiểm tra dữ liệu đang khảo sát có phải là tuyến đi BX Củ Chi - BX An Sương không.

Chọn thời gian hoàn thành một chuyến đi

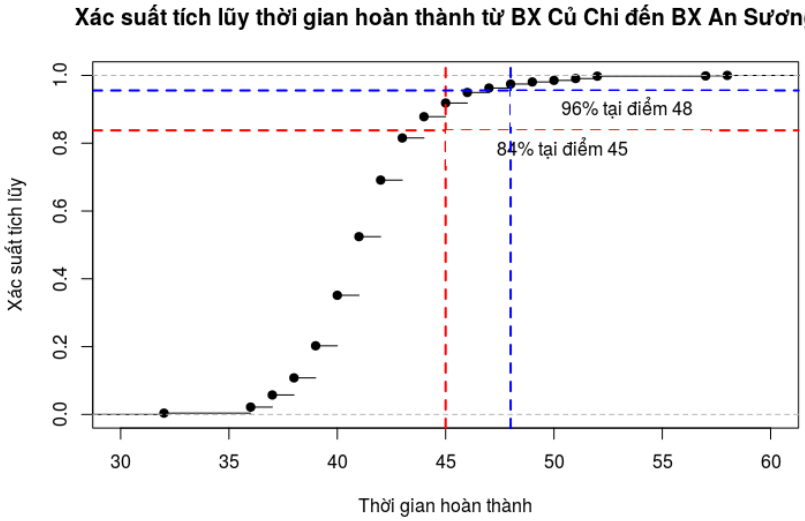
Theo quy định của quản lý xe, thời gian hoàn thành BX Củ Chi - BX An Sương là 45 phút, cho nên ta chỉ chọn những chuyến xe hoàn thành từ khoảng 30 phút đến 60 phút, những xe vượt quá thời gian hoàn thành trên 1 tiếng bị loại bỏ.
Kham khảo **PL1** tại mục Phụ Lục xuất ra hình vẽ phân phối tần suất thời gian hoàn thành lộ trình BX Củ Chi - BX An Sương



Kham khảo **PL4** tại mục Phụ Lục xuất ra hình vẽ sắp xếp thời gian hoàn thành lộ trình BX Củ Chi - BX An Sương xác suất xuất hiện tăng dần.



Kham khảo **PL3** tại mục Phụ Lục xuất ra hình vẽ cho thấy xác suất tích lũy thời gian hoàn thành lộ trình BX Củ Chi - BX An Sương



Ta chọn thời gian hoàn thành một chuyến đi từ BX Củ Chi đến BX An Sương trước đúng 45 phút là về đích đúng giờ, và từ hình vẽ trên ta thấy có 84% trong 1289 chuyến xe được khảo sát đạt được thời gian về đích là 45 phút.
Như vậy ta có P(về đích đúng giờ) = 0.84 và P(về đích trễ giờ) = 0.16

Đồng bộ hóa ghi nhận mỗi bước di chuyển

Sau đó kham khảo **PL12** tại mục Phụ Lục để tính mỗi bước di chuyển. Sau đây là ví dụ 10 bước di chuyển đầu tiên trong một chuyến đi ngẫu nhiên được chọn

Bước đầu	Bước kế tiếp	Vĩ độ đầu	Kinh độ đầu	Vĩ độ kế tiếp	Kinh độ kế tiếp	Khoảng cách (mét)	Thời gian di chuyển (giây)
0	1	10.971375	106.481906666667	10.9710116666667	106.481858333333	41	14
1	2	10.9710116666667	106.481858333333	10.9708933333333	106.482023333333	22	13
2	3	10.9708933333333	106.482023333333	10.9708166666667	106.482206666667	22	06
3	4	10.9708166666667	106.482206666667	10.9706533333333	106.482398333333	28	13
4	5	10.9706533333333	106.482398333333	10.970555	106.48252	17	11
5	6	10.970555	106.48252	10.9701433333333	106.483123333333	80	20
6	7	10.9701433333333	106.483123333333	10.9695933333333	106.48386	101	20
7	8	10.9695933333333	106.48386	10.968885	106.485015	149	20
8	9	10.968885	106.485015	10.9682416666667	106.48598	127	20
9	10	10.9682416666667	106.48598	10.9677616666667	106.487061666667	130	20
10	11	10.9677616666667	106.487061666667	10.9674233333333	106.487998333333	109	20

Khi quan sát, ta nhận thấy thông thường sau 20 giây hệ thống sẽ ghi nhận một bước chuyển mới, nhưng do trục trặc về kỹ thuật ghi nhận, vẫn thường xuyên xảy ra hiện tượng hồi đáp sớm hơn, trễ hơn so với 20 giây, cho nên ta phải làm một thao tác trung gian đồng bộ hóa 20 giây nhiều nhất có thể. Vì ta giả định những bước di chuyển hoàn toàn độc lập nhau nên ta có thể gom nhóm những bước hồi đáp không chuẩn lại thành bội số của 20 và sau đó chia đều cho 20, nếu số dư còn lại lớn hơn 15 giây thì chấp nhận hồi đáp dư đó, nếu không thì loại bỏ.
Kham khảo **PL13** tại mục Phụ lục cho thao tác đồng bộ hóa 20 giây và ví dụ sau xem kết quả đồng bộ cho một chuyến được chọn ngẫu nhiên
Dữ liệu chưa được đồng bộ hóa

STT	Khoảng thời gian hồi đáp (giây)	Khoảng cách bước đi (mét)
1	17	236
2	20	71
3	20	86
4	20	145
5	20	136
6	9	93
7	1	9
8	20	104
9	20	277
10	20	240
11	5	6
12	12	11
13	20	145
14	20	97
15	14	17
16	20	0
17	2	14
18	20	171
19	20	283
20	20	283

STT	Khoảng thời gian hồi đáp (giây)	Khoảng cách bước đi (mét)
21	21	106
22	5	3
23	20	0
24	7	17
25	20	214
26	15	115
27	10	13
28	20	166
29	20	307
30	20	259
31	20	208
32	20	43
33	20	144
34	20	286
35	20	287
36	20	126
37	20	152
38	20	254
39	18	111
40	7	9

STT	Khoảng thời gian hồi đáp (giây)	Khoảng cách bước đi (mét)
41	20	139
42	20	255
43	20	294
44	20	293
45	20	269
46	20	283
47	20	267
48	20	221
49	14	68
50	6	12
51	20	165
52	20	266
53	20	259
54	20	220
55	20	168
56	13	19
57	20	0
58	15	32
59	12	32
60	4	17

STT	Khoảng thời gian hồi đáp (giây)	Khoảng cách bước đi (mét)
61	20	226
62	20	337
63	20	394
64	20	224
65	11	39
66	5	10
67	20	208
68	20	340
69	20	359
70	20	259
71	20	297
72	20	325
73	20	303
74	20	302
75	20	284
76	20	287
77	20	268
78	20	199
79	11	17
80	6	12

STT	Khoảng thời gian hồi đáp (giây)	Khoảng cách bước đi (mét)
81	20	112
82	6	9
83	7	17
84	20	179
85	20	279
86	20	270
87	20	139
88	20	72
89	10	39
90	7	12
91	20	195
92	20	345
93	20	332
94	20	151
95	10	9
96	20	149
97	20	230
98	20	261
99	20	186
100	14	37

STT	Khoảng thời gian hồi đáp (giây)	Khoảng cách bước đi (mét)
101	19	20
102	20	170
103	20	236
104	20	243
105	20	265
106	18	51
107	7	14
108	20	151
109	20	201

Dữ liệu sau khi được đồng bộ hóa

STT	Khoảng 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
9	20	97
10	20	0
11	20	171
12	20	283
13	20	283
14	20	0
15	20	214
16	20	166
17	20	307
18	20	259
19	20	208
20	20	43

STT	Khoảng thời gian hồi đáp (giây)	Khoảng cách bước đi (mét)
21	20	144
22	20	286
23	20	287
24	20	126
25	20	152
26	20	254
27	20	139
28	20	255
29	20	294
30	20	293
31	20	269
32	20	283
33	20	267
34	20	221
35	20	165
36	20	266
37	20	259
38	20	220
39	20	168
40	20	0

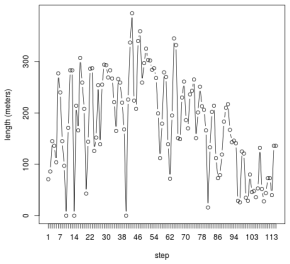
STT	Khoảng thời gian hồi đáp (giây)	Khoảng cách bước đi (mét)
41	20	226
42	20	337
43	20	394
44	20	224
45	20	208
46	20	340
47	20	359
48	20	259
49	20	297
50	20	325
51	20	303
52	20	302
53	20	284
54	20	287
55	20	268
56	20	199
57	20	112
58	20	179
59	20	279
60	20	270

STT	Khoảng thời gian hồi đáp (giây)	Khoảng cách bước đi (mét)
61	20	139
62	20	72
63	20	195
64	20	345
65	20	332
66	20	151
67	20	149
68	20	230
69	20	261
70	20	186
71	20	170
72	20	236
73	20	243
74	20	265
75	20	151
76	20	201
77	20	29
78	20	125
79	20	121
80	20	35

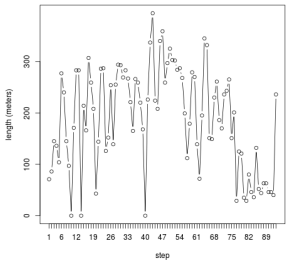
STT	Khoảng thời gian hồi đáp (giây)	Khoảng cách bước đi (mét)
81	20	29
82	20	80
83	20	46
84	20	36
85	20	132
86	20	52
87	20	44
88	20	63
89	20	63
90	20	46
91	20	46
92	21	40
93	17	236

Trực quan những bước di chuyển

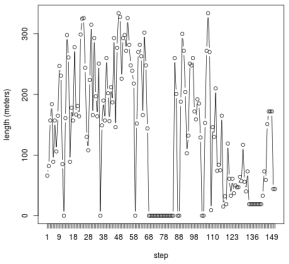
Trích ngẫu nhiên 80% thời gian di chuyển của 24 chuyến xe BX Củ Chi - BX An Sương. Trong một chuyến xe, mỗi giá trị tính bằng mét, mỗi bước di chuyển giữa các giá trị được ghi nhận sau 20 giây.
Chuyến 1: 71, 86, 145, 136, 104, 277, 240, 145, 97, 0, 171, 283, 283, 0, 214, 166, 307, 259, 208, 43, 144, 286, 287, 126, 152, 254, 139, 255, 294, 293, 269, 283, 267, 221, 165, 266, 259, 220, 168, 0, 226, 337, 394, 224, 208, 340, 359, 259, 297, 325, 303, 302, 284, 287, 268, 199, 112, 179, 279, 270, 139, 72, 195, 345, 332, 151, 149, 230, 261, 186, 170, 236, 243, 265, 151, 201, 29, 125, 121, 35, 29, 80, 46, 36, 132, 52, 44, 63, 63, 46, 46, 40, 236
Chuyến 2:...
...
Chuyến 12: 104, 145, 124, 166, 194, 205, 188, 158, 107, 121, 91, 152, 218, 93, 100, 0, 167, 167, 250, 160, 144, 223, 215, 131, 114, 241, 206, 193, 0, 0, 169, 212, 192, 200, 257, 223, 198, 166, 135, 184, 105, 90, 151, 202, 192, 168, 177, 217, 168, 243, 211, 109, 161, 272, 287, 281, 248, 257, 166, 261, 250, 152, 238, 267, 190, 111, 0, 117, 138, 209, 182, 123, 58, 151, 222, 248, 218, 110, 0, 0, 111, 211, 61, 114, 83, 109, 26, 232, 232, 232, 232, 232, 232, 232, 232, 232, 96, 37, 110, 48, 34, 121, 121, 46, 33, 88, 64, 235, 183, 125, 220
Chuyến 13:...
...
Chuyến 24: 279, 17, 52, 48, 67, 141, 131, 131, 95, 157, 155, 196, 242, 219, 50, 172, 219, 243, 148, 74, 5, 105, 194, 68, 216, 224, 231, 61, 76, 228, 295, 306, 293, 260, 236, 279, 284, 232, 174, 75, 205, 103, 98, 232, 266, 257, 188, 60, 192, 228, 256, 236, 215, 23, 10, 100, 44, 132, 255, 290, 272, 216, 237, 281, 274, 272, 294, 189, 0, 41, 84, 65, 197, 247, 221, 121, 83, 138, 252, 223, 195, 135, 17, 0, 6, 15, 44, 121, 61, 206, 220, 148, 52, 164, 208, 248, 203, 179, 11, 0, 62, 112, 252, 307, 316, 168, 104, 87, 192, 148, 131, 91, 215, 215, 143, 126, 279, 279
Kham khảo **PL2** tại mục Phụ Lục xuất ra trực quan hóa giá trị các bước di chuyển của 24 chuyến dữ liệu ngẫu nhiên, ta có



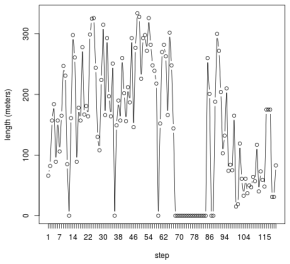
Toàn bộ di chuyển chuyến xe 1 - đúng giờ



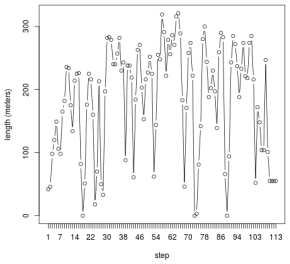
80% di chuyển chuyến xe 1 - đúng giờ



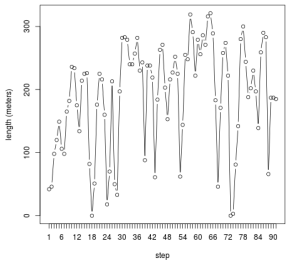
Toàn bộ di chuyển chuyến xe 2 - trễ giờ



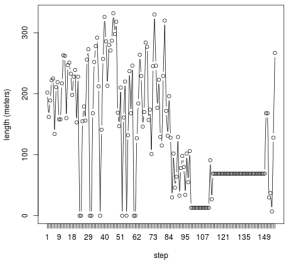
80% di chuyển chuyến xe 2 - trễ giờ



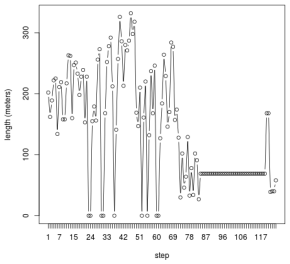
Toàn bộ di chuyển chuyến xe 3 - đúng giờ



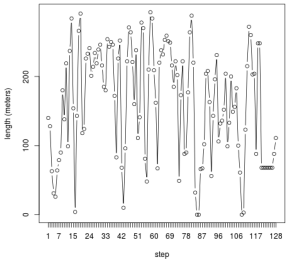
80% di chuyển chuyến xe 3 - đúng giờ



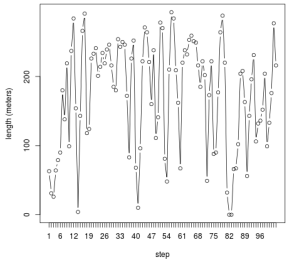
Toàn bộ di chuyển chuyến xe 4 - trễ giờ



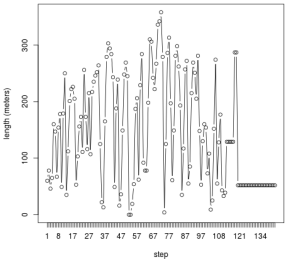
80% di chuyển chuyến xe 4 - trễ giờ



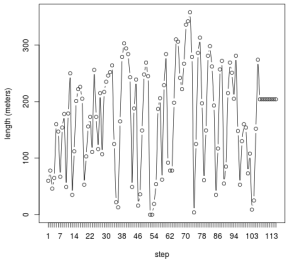
Toàn bộ di chuyển chuyến xe 5 - đúng giờ



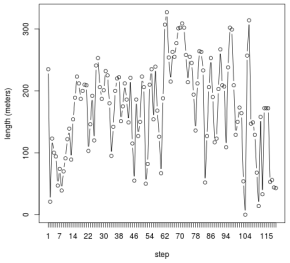
80% di chuyển chuyến xe 5 - đúng giờ



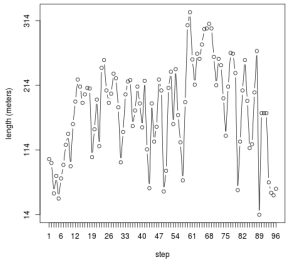
Toàn bộ di chuyển chuyến xe 6 - trễ giờ



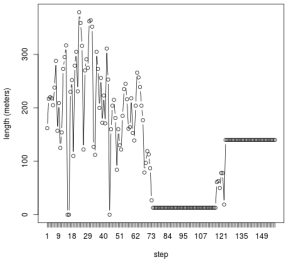
80% di chuyển chuyến xe 6 - trễ giờ



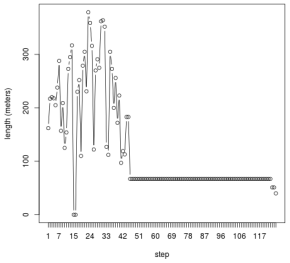
Toàn bộ di chuyển chuyến xe 7 - đúng giờ



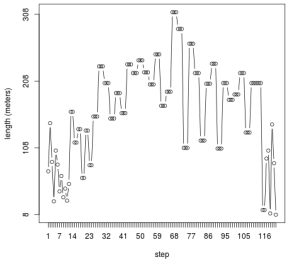
80% di chuyển chuyến xe 7 - đúng giờ



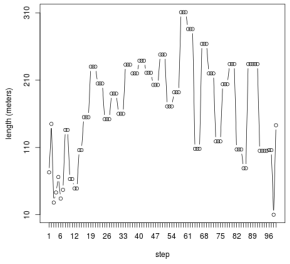
Toàn bộ di chuyển chuyến xe 8 - trễ giờ



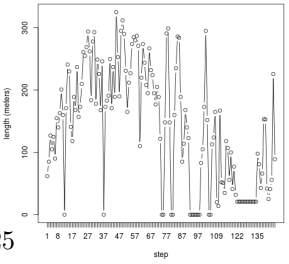
80% di chuyển chuyến xe 8 - trễ giờ



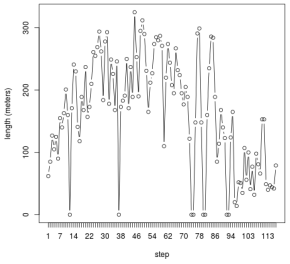
Toàn bộ di chuyển chuyến xe 9 - đúng giờ



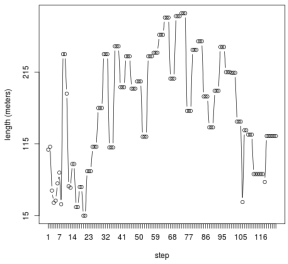
80% di chuyển chuyến xe 9 - đúng giờ



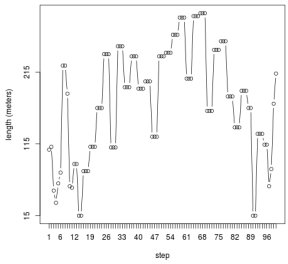
Toàn bộ di chuyển chuyến xe 10 - trễ giờ



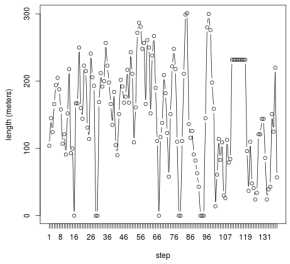
80% di chuyển chuyến xe 10 - trễ giờ



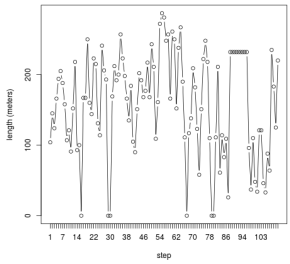
Toàn bộ di chuyển chuyến xe 11 - đúng giờ



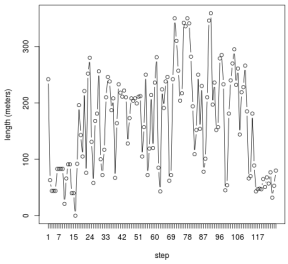
80% di chuyển chuyến xe 11 - đúng giờ



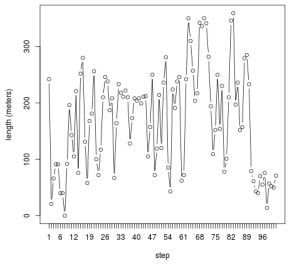
Toàn bộ di chuyển chuyến xe 12 - trễ giờ



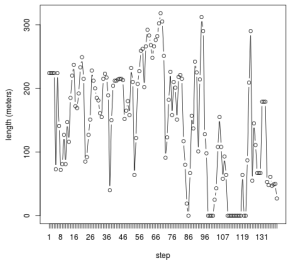
80% di chuyển chuyến xe 12 - trễ giờ



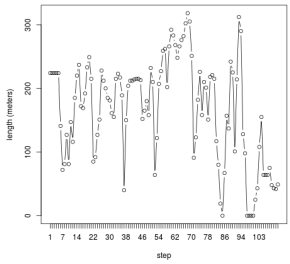
Toàn bộ di chuyển chuyến xe 13 - đúng giờ



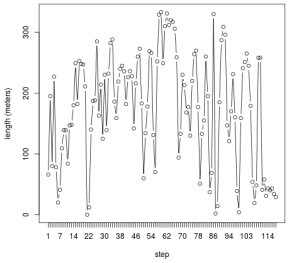
80% di chuyển chuyến xe 13 - đúng giờ



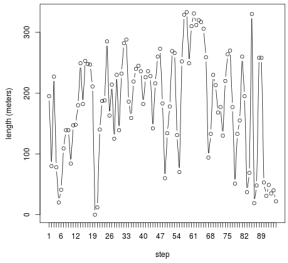
Toàn bộ di chuyển chuyến xe 14 - trễ giờ



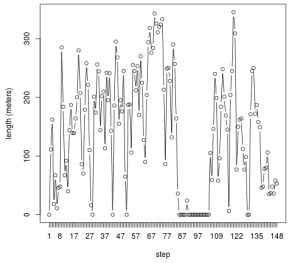
80% di chuyển chuyến xe 14 - trễ giờ



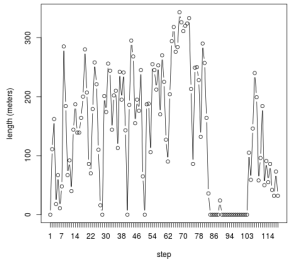
Toàn bộ di chuyển chuyến xe 15 - đúng giờ



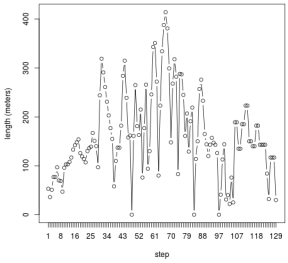
80% di chuyển chuyến xe 15 - đúng giờ



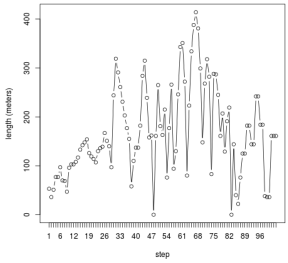
Toàn bộ di chuyển chuyến xe 16 - trễ giờ



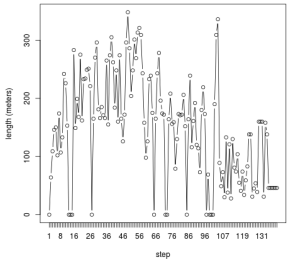
80% di chuyển chuyến xe 16 - trễ giờ



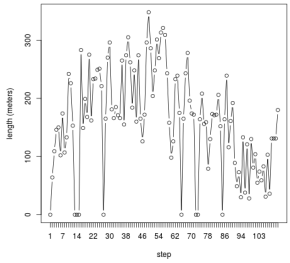
Toàn bộ di chuyển chuyến xe 17 - đúng giờ



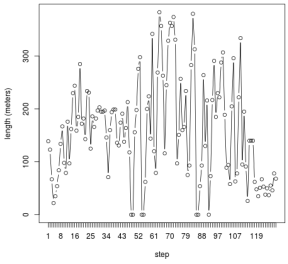
80% di chuyển chuyến xe 17 - đúng giờ



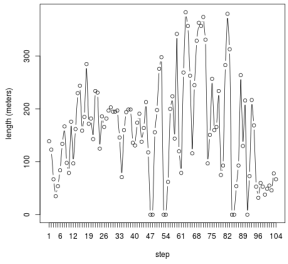
Toàn bộ di chuyển chuyến xe 18 - trễ giờ



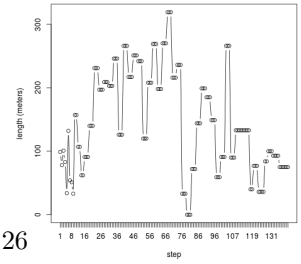
80% di chuyển chuyến xe 18 - trễ giờ



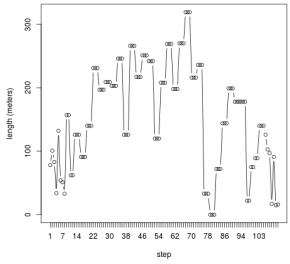
Toàn bộ di chuyển chuyến xe 19 - đúng giờ



80% di chuyển chuyến xe 19 - đúng giờ



Toàn bộ di chuyển chuyến xe 20 - trễ giờ



80% di chuyển chuyến xe 20 - trễ giờ

Trực quan hóa 20 dòng dữ liệu ngẫu nhiên

3.2 Rời rạc hóa những bước di chuyển

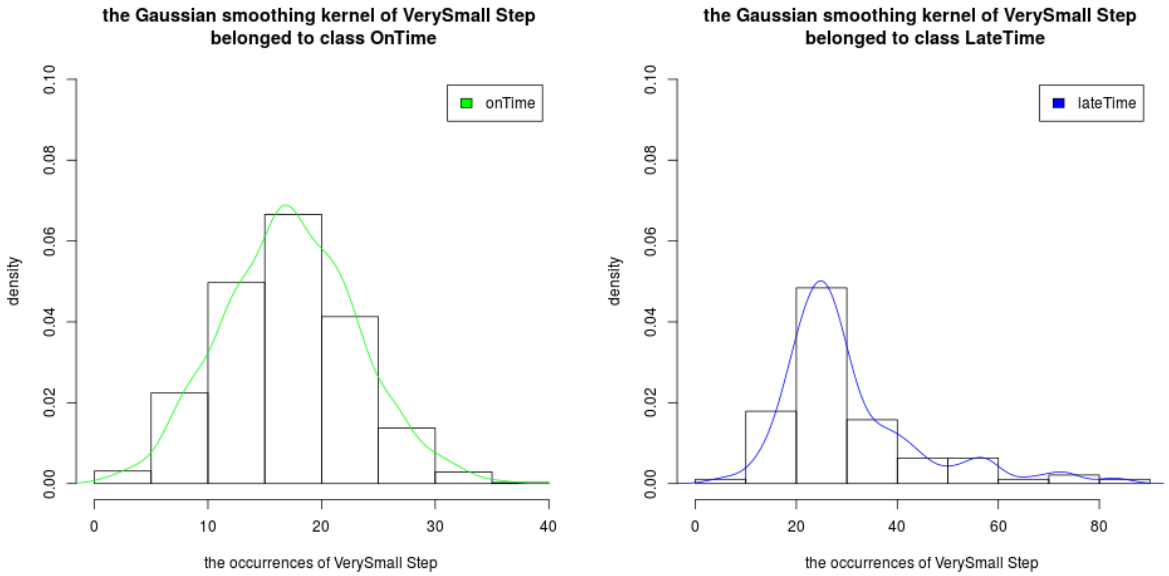
Thực hiện phân loại những bước di chuyển: biến X_1 là bước đi rất nhỏ 0-15 (km/h), biến X_2 là bước đi nhỏ 15-30 (km/h), biến X_3 là bước đi trung bình 30-45 (km/h), biến X_4 là bước đi xa 45-60 (km/h), biến X_5 là bước đi rất xa trên 60 km/h.
Bảng số liệu sau mô tả 10 dữ liệu mà ta sẽ làm việc.

Số lần thực hiện bước đi từ 0-15 (km/h)	Số lần thực hiện bước đi từ 15-30 (km/h)	Số lần thực hiện bước đi từ 30-45 (km/h)	Số lần thực hiện bước đi từ 45-60 (km/h)	Số lần thực hiện bước đi trên 60 km/h	Trạng thái đến đích
19	23	19	29	3	đúng giờ
16	14	40	21	0	đúng giờ
19	23	48	13	0	đúng giờ
14	27	43	12	0	đúng giờ
13	33	44	9	0	đúng giờ
45	26	25	25	0	trễ giờ
56	26	27	16	0	trễ giờ
29	22	41	22	1	trễ giờ
83	11	15	12	5	trễ giờ
28	32	36	21	0	trễ giờ

Như đã nói phần trên, ta có P(về đích đúng giờ) = 0.84 và P(về đích trễ giờ) = 0.16

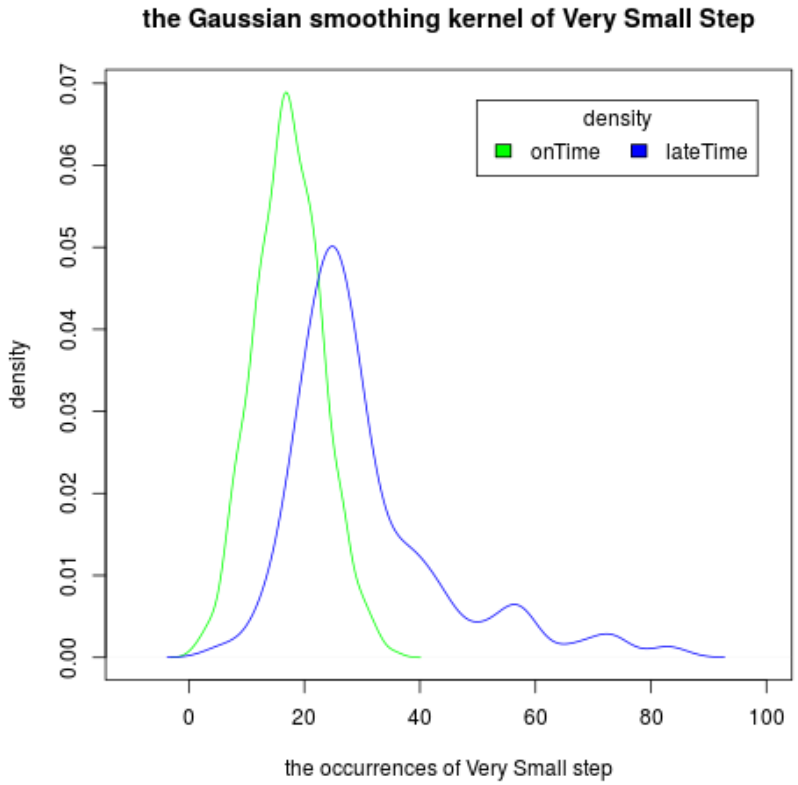
3.3 Vẽ hàm mật độ xác suất của các biến

Kham khảo **PL5**, **PL6** tại mục Phụ lục, dùng giải thuật Kernel Density Estimation để vẽ hàm mật độ xác suất của các biến dưới đây:

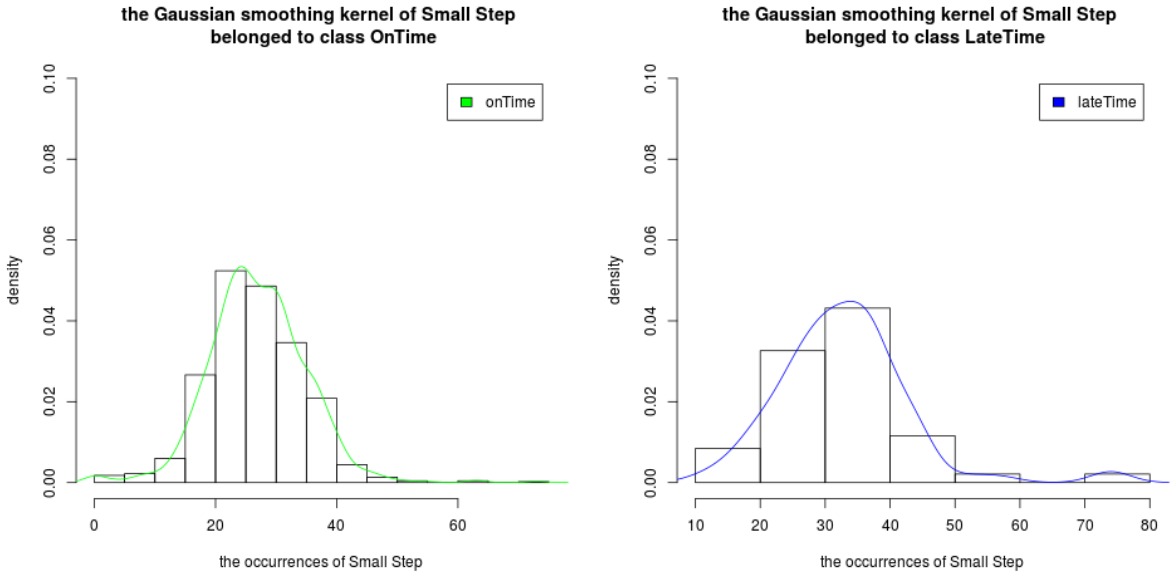


Hàm mật độ xác suất của biến X_1 bước đi 0-15 km/h

Nếu gộp lại

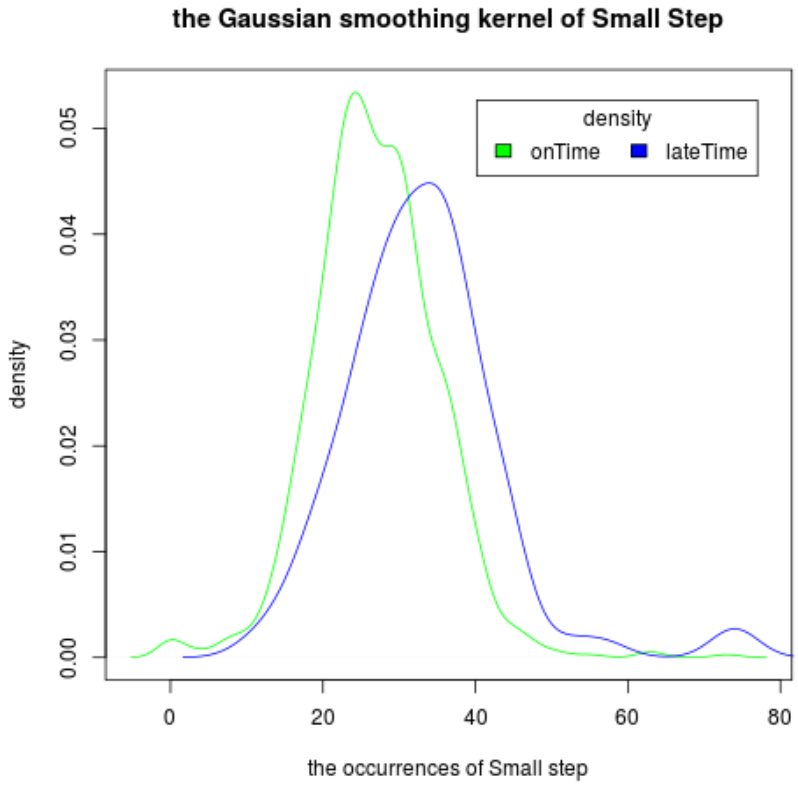


Hàm mật độ xác suất của biến X_1 bước đi 0-15 km/h

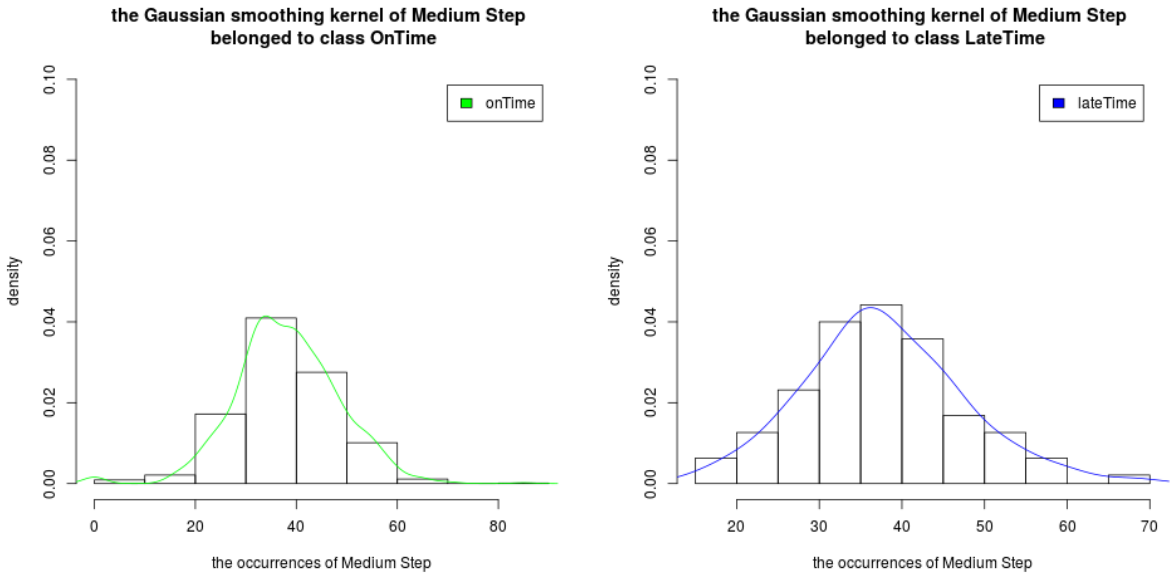


Hàm mật độ xác suất của biến X_2 bước đi 15-30 km/h

Nếu gộp lại

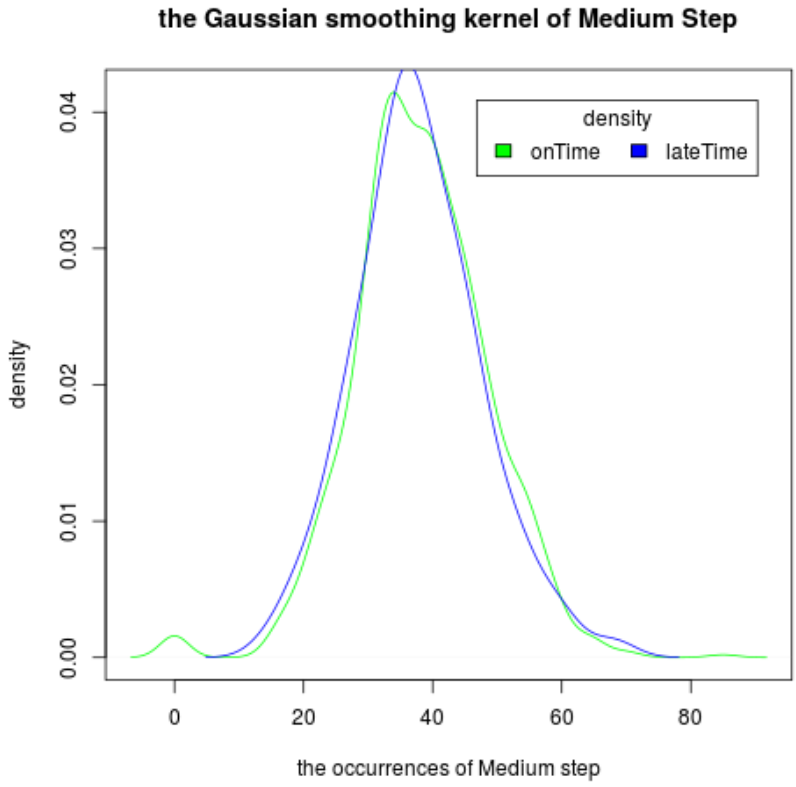


Hàm mật độ xác suất của biến X_2 bước đi 15-30 km/h

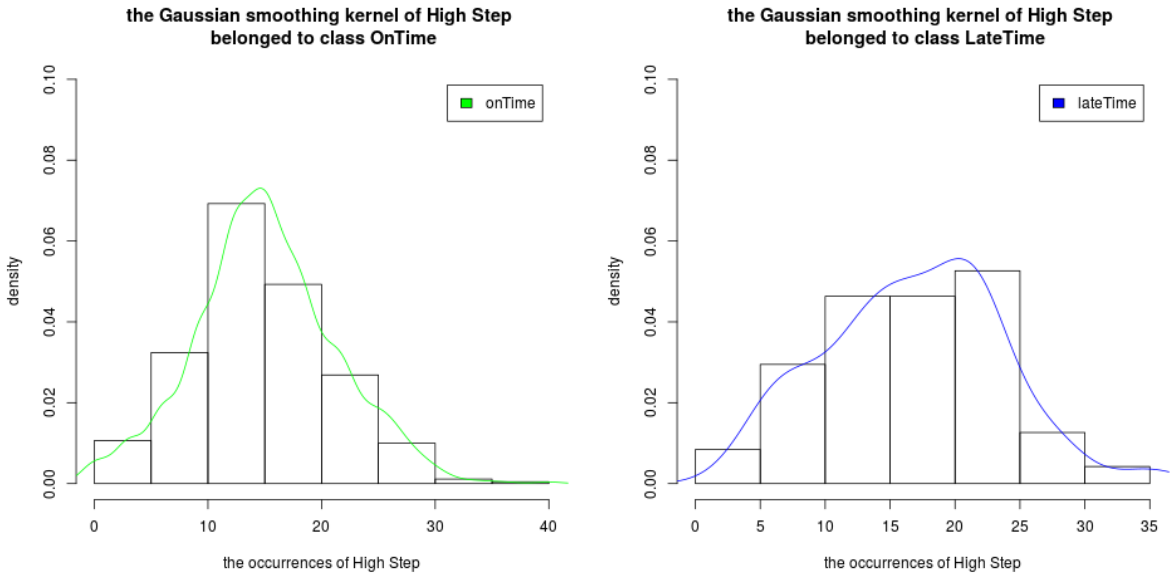


Hàm mật độ xác suất của biến X_3 bước đi 30-45 km/h

Nếu gộp lại

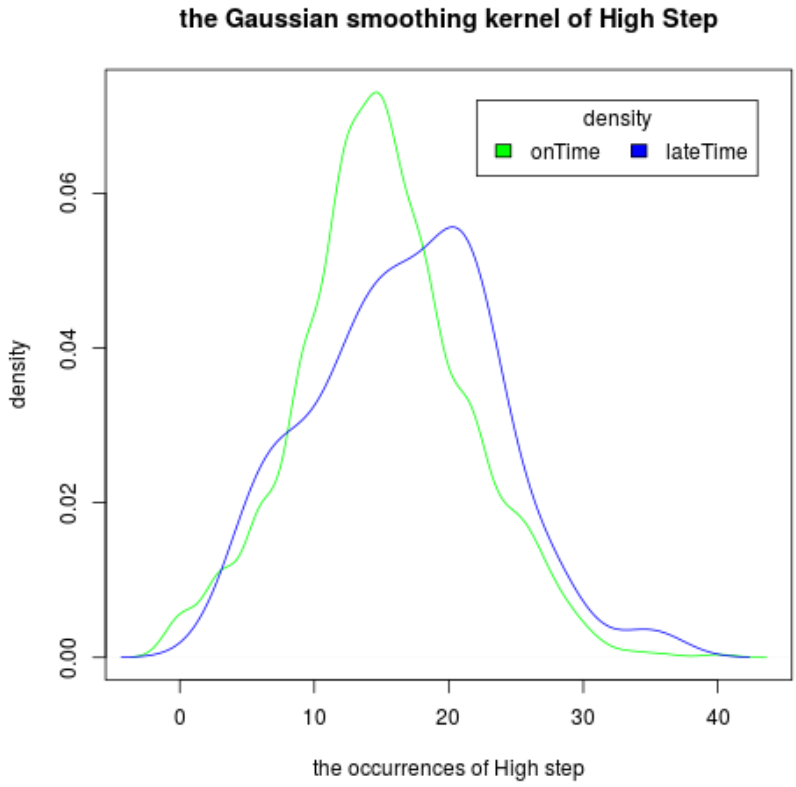


Hàm mật độ xác suất của biến X_3 bước đi 30-45 km/h

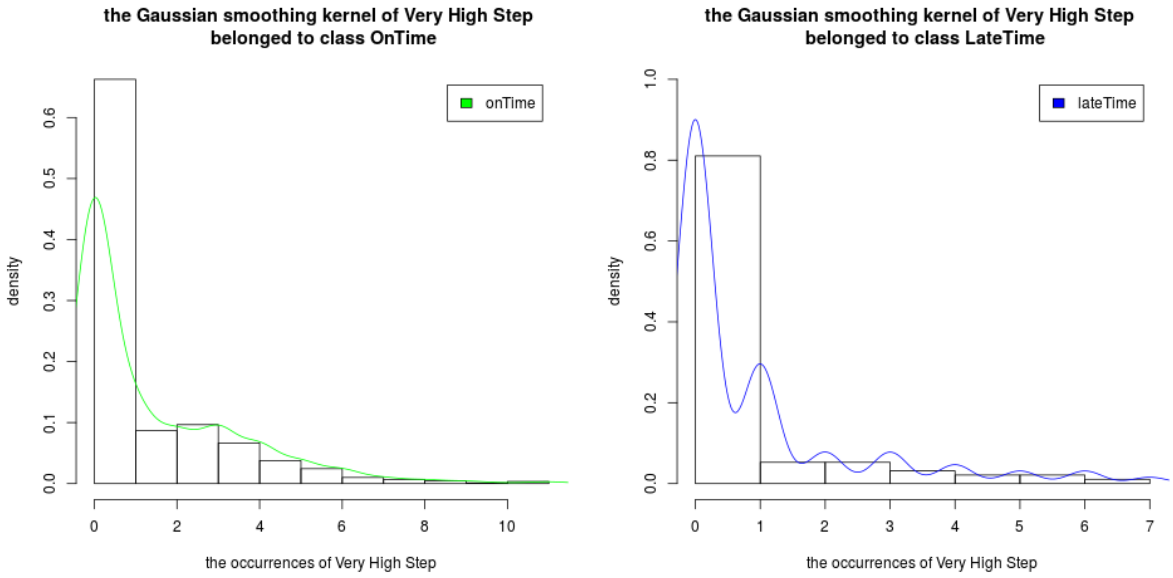


Hàm mật độ xác suất của biến X_4 bước đi 45-60 km/h

Nếu gộp lại

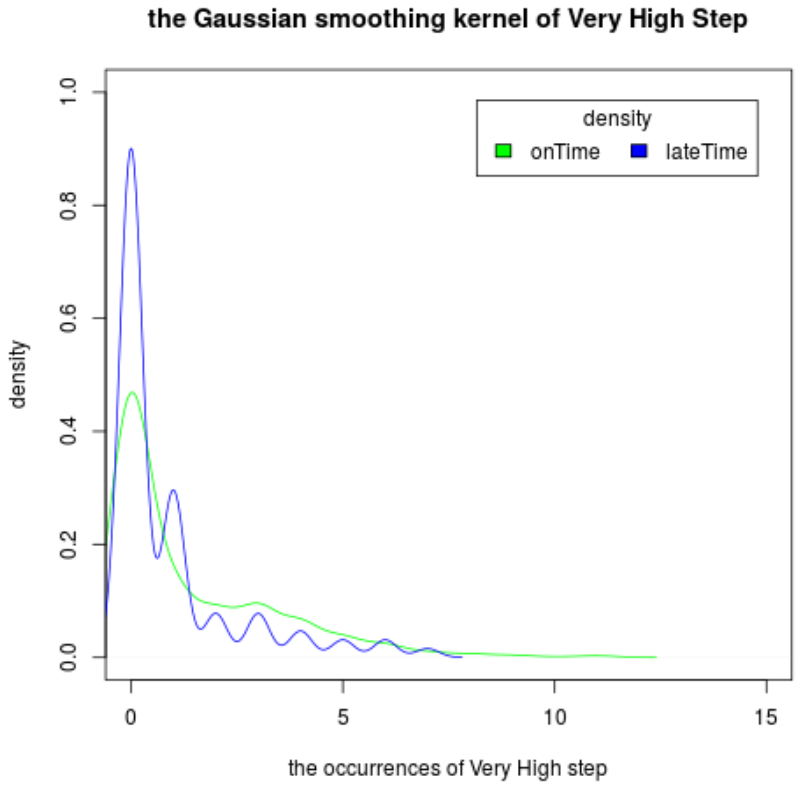


Hàm mật độ xác suất của biến X_4 bước đi 45-60 km/h



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

Nếu gộp lại



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

3.4 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

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

Nếu muốn hiện thực bằng dòng lệnh R, kham khảo **PL7**

3.5 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)
Cách giải:

- Chỉ giữ lại các bước di chuyển trong 80% thời gian đầu
- Dữ liệu làm việc là dữ liệu sau khi được đồng bộ hóa khoảng cách thời gian hồi đáp (20 giây)
- 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

- Ta chọn thời gian hoàn thành trước đúng 45 phút là về đích đúng giờ
  Như vậy ta có P(về đích đúng giờ) = 0.84 và P(về đích trễ giờ) = 0.16
- Dùng giải thuật Kernel Density Estimation để vẽ hàm mật độ xác suất của các biến
- Sử dụng công thức Bayes tổng quát

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(B_i) \cdot \mathbb{P}(A|B_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B_i) \cdot \mathbb{P}(A|B_i)}{\sum_{j=1}^{m} \mathbb{P}(B_j) \cdot \mathbb{P}(A|B_j)}$$

  các biến cố $A_1, A_2, ..., A_n$ là n biến cố **độc lập lẫn nhau** thì

$$\mathbb{P}\left[\bigcap_{i=1}^{n} A_j\right] = \prod_{i=1}^{n} \mathbb{P}[A_i]$$

Cụ thể như sau:
Cho dữ liệu kiểm tra x có giá trị x_1, x_2, x_3, x_4, x_5

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

Do các biến cố X_1, X_2, X_3, X_4, X_5 là các biến cố độc lập lẫn nhau cho nên
$\mathrm{P}(x|onTime) = \mathrm{P}(x_1|onTime)\mathrm{P}(x_2|onTime)\mathrm{P}(x_3|onTime)\mathrm{P}(x_4|onTime)\mathrm{P}(x_5|onTime)$
và $\mathrm{P}(x|lateTime) = \mathrm{P}(x_1|lateTime)\mathrm{P}(x_2|lateTime)\mathrm{P}(x_3|lateTime)\mathrm{P}(x_4|lateTime)\mathrm{P}(x_5|lateTime)$
Kham khảo **PL7** để chạy phân loại dữ liệu kiểm tra

3.6 Kết quả

Lấy ngẫu nhiên 190 mẫu trong 996 mẫu dữ liệu huấn luyện ra kiểm tra phân loại
Một số kết quả dự đoán đúng

X_1	X_2	X_3	X_4	X_5	Kết quả thực tế	Xác suất dự đoán đúng giờ
10	48	39	9	0	onTime	0.867886884048324
14	36	32	12	4	onTime	0.975384350213014
26	15	48	15	0	onTime	0.787585539441556
33	17	44	11	3	onTime	0.724957571656855
23	44	42	9	0	lateTime	0.382959767057761
30	42	34	11	1	lateTime	0.210249084300825
44	15	21	35	0	lateTime	9.45802157172484e-10
51	39	55	12	0	lateTime	6.84176417466922e-28
56	44	35	6	0	lateTime	2.02430903076165e-49
73	18	34	28	3	lateTime	7.63586294825022e-162

Một số kết quả dự đoán sai

X_1	X_2	X_3	X_4	X_5	Kết quả thực tế	Xác suất dự đoán đúng giờ
19	34	41	16	1	lateTime	0.839811262772558
13	30	60	9	0	lateTime	0.956048124887834
20	26	52	12	0	lateTime	0.926660526069996
23	34	22	23	0	onTime	0.462938246287487
36	24	25	16	6	onTime	0.291143171211406
24	29	41	16	1	lateTime	0.758174017995704
24	31	45	12	0	lateTime	0.760021675231651
23	30	44	15	0	lateTime	0.809102556309718
13	31	68	3	0	lateTime	0.882496024826019
21	42	36	13	1	lateTime	0.621148813952554

Ta thấy ở dòng 2, dù chuyến xe có xác suất dự đoán đúng giờ rất cao 0.956048124887834 (tiệm cận gần tới 1) nhưng cuối cùng về đích trễ giờ. Vậy là trên thực tế biến cố trễ giờ dù có xác suất rất nhỏ xảy ra nhưng vẫn xảy ra. Hay ở dòng 5, dù chuyến xe có xác suất dự đoán đúng giờ rất thấp 0.291143171211406 nhưng cuối cùng về đích đúng giờ.
Kết luận Xác suất phù hợp để dự đoán một biến cố xảy ra trên thực tế, chấp nhận yếu tố ngẫu nhiên xảy ra ảnh hưởng đến kết quả dự đoán.
Đánh giá kết quả dự đoán trên mẫu: Độ chính xác là 0.7631579 và sai số là 0.23684211
Kiểm tra trên tập mẫu khác, vì tác giả đã lấy hết mẫu di chuyển từ BX Củ Chi đến BX An Sương mang đi huấn luyện nên đành phải lấy mẫu di chuyển từ BX An Sương đến BX Củ Chi để kiểm tra. Với số lượng mẫu kiểm tra là 165 mẫu di chuyển từ BX An Sương đến BX Củ Chi, đánh giá kết quả dự đoán: Độ chính xác là 0.74545455 và sai số là 0.25454545

Chương 4

KẾT LUẬN

4.1 Tổng kết

Trình bày của chương mở đầu đã nói hết cho phần tổng kết. Tác giả nhắc lại đề tài Luận Văn: Dùng Thống Kê định lượng kinh nghiệm di chuyển trên một lộ trình quen thuộc, dự đoán xác suất xe buýt về trạm đích đúng giờ. Chương mở đầu đã trình bày cách nghiên cứu bài toán và nêu sơ lược các bước trong phương pháp giải. Các chương tiếp theo để khẳng định sử dụng xác suất Bayes là hợp lý để giải bài toán tìm xác suất xe buýt về trạm đúng giờ.

4.2 Đóng góp của đề tài

Tác giả không đặt mục tiêu tham vọng giải bài toán này có thể đem ra ứng dụng thực tế được vì bản thân tác giả không phải là nhà thống kê, không có nhiều kinh nghiệm làm việc với kích thước mẫu rất nhỏ có thể đưa ra dự đoán đúng trên kích thước quần thể. Nhưng thông qua Luận Văn, tác giả chứng minh được sử dụng kiến thức Thống Kê cơ bản (không chứng minh công thức Toán học vì tác giả không phải là nhà Toán học), giải được bài toán của Luận Văn.

4.3 Hướng phát triển

Tác giả không muốn đề cập cách giải bài toán của Luận Văn với dữ liệu cực lớn mà muốn đề cập đến sử dụng kiến thức Thống Kê giải với dữ liệu mẫu ngẫu nhiên, kích thước rất nhỏ. Với phương châm như vậy, tác giả chia sẽ một số phát hiện để nếu có ai hứng thú muốn phát triển đề tài của Luận Văn này. Tác giả đã phát hiện ngoài sử dụng phương pháp xác suất Bayes để phân loại, có rất nhiều thuật toán khác trong Thống Kê để phân loại. Ví dụ như thuật toán FRBCS.W giúp phân loại dữ liệu của tác giả đúng đến 75% (nhưng tác giả không trình bày bằng chứng ở đây vì nó nằm ngoài mục tiêu trình bày sử dụng phương pháp Bayes của Luận Văn). Còn muốn tìm thời gian xe buýt về trạm đích khi học các dữ liệu mẫu giới hạn trong quá khứ thì tác giả tìm được các thuật toán ước lượng hồi quy trong Thống Kê như thuật toán Wang and Mendel, giúp ta làm điều này, nhưng tác giả cũng không trình bày ở đây.
Nếu có tham vọng muốn bài toán của Luận Văn này được triển khai trên thực tế, thì người đọc cần bổ sung thêm kiến thức Thống Kê như làm thế nào tạo ra thiết kế thực nghiệm, tiêu chuẩn lựa chọn mô hình để rút ra kết luận có tính thuyết phục hơn.

Chương 5

PHỤ LỤC

PL1 Mã nguồn R vẽ histogram dữ liệu

```
#set your working directory where your file data locates
setwd("/home/thuy1/git/predictUsingProbability/Preprocess")
data=read.table(file="CC_AS_FinishTime.csv")[,1]
hist(data,breaks=12, prob=TRUE,
        xlab="Thời gian hoàn thành BX Củ Chi đến BX An Sương",
        main="Biểu đồ phân phối thời gian hoàn thành \n BX Củ Chi đến BX An Sương")
curve(dnorm(x, mean=mean(data), sd=sd(data)), add=TRUE)
```

PL2 Mã nguồn R trực quan hóa giá trị các bước di chuyển của 24 chuyến xe trên

```
#set your working directory where your file data locates
setwd("/home/thuy/workspace/Preprocess")
#count maximum column length
count.fields("data.txt", sep = ",")
maxCol <- max(count.fields("data.txt", sep = ","))
#load data with different column length
dat=read.table("data.txt", header = FALSE,
                col.names = 1:maxCol, #maxCol is maximum column length in your data row
                sep = ",",
                fill = TRUE) #set value NA for empty column
i=1
for (i in 1:nrow(dat)) {
        d=dat[i,] # get row data
        d=d[!is.na(d)] #remove column NA
        x=length(d)
        #capture image
        png(filename=paste("capture", i, ".png", sep = ""))
        #remove axes
        temp <- plot(1:x, d, type='b', axes=FALSE, xlab = "step", ylab = "length (meters)")
        #adjust axes length
        temp <- axis(side=1, at=c(1:x))
        temp <- axis(side=2, at=seq(min(d), max(d), by=100))
        temp <- box()
        print(temp)
        dev.off()
}
```

PL3 Mã nguồn R xuất hình thể hiện xác suất tích lũy thời gian hoàn thành BX Củ Chi - BX An Sương

```
#set your working directory where your file data locates
setwd("/home/thuy/workspace/Preprocess")
y = read.csv("CC_AS_Rep.csv",header=FALSE)$V1
p = ecdf(y)
plot(p,
        xlab = 'Thời gian hoàn thành',
        ylab = 'Xác suất tích lũy',
        main = 'Xác suất tích lũy thời gian hoàn thành từ BX Củ Chi đến BX An Sương' )
abline(v = 45, h = 0.83773583,col="red",lwd=2, lty=2)
legend(45, 0.83773583, '84% tại điểm 45', box.lwd = 0)
abline(v = 48, h = 0.9554717,col="blue",lwd=2, lty=2)
legend(48, 0.9554717, '96% tại điểm 48', box.lwd = 0)
```

PL4 Mã nguồn R sắp xếp tăng dần xác suất thời gian hoàn thành BX Củ Chi - BX An Sương

```
#set your working directory where your file data locates
setwd("/home/thuy1/git/predictUsingProbability/Preprocess/")
mydata = read.csv("CC_AS_Freq.csv",sep = "|",header=FALSE)[ ,1:2]
#set column name for your data
colnames(mydata) <- c("X1","X2")
X1=mydata$X1
X2=mydata$X2
mydata$X1 <- factor(mydata$X1, levels = mydata$X1[order(mydata$X2)])
library(ggplot2)
ggplot(mydata, aes(x = mydata$X1, y = mydata$X2)) +
        theme_bw() + geom_bar(stat = "identity") +
        xlab("Thời gian hoàn thành ") +
        ylab("Tần số xuất hiện")
```

PL5 Mã nguồn R vẽ đường density chồng lên histogram

```
#set your working directory where your file data locates
setwd('/home/thuy1/git/predictUsingProbability/Preprocess')
lateTime=read.table(file="freqVerySmall_LateTime.csv")[,1]
hist(lateTime, probability = TRUE,
main="the Gaussian smoothing kernel of VerySmall Step \n belonged to class LateTime",
ylab="density", xlab="the occurrences of VerySmall Step"
)
lines(density(lateTime, kernel=c("gaussian")), col="blue")
legend("topright", inset=.05, c("lateTime"), fill=c("blue"), horiz=TRUE)
onTime=read.table(file="freqVerySmall_OnTime.csv")[,1]
hist(onTime, probability = TRUE,
main="the Gaussian smoothing kernel of VerySmall Step \n belonged to class OnTime",
ylab="density", xlab="the occurrences of VerySmall Step"
) lines(density(onTime, kernel=c("gaussian")), col="green")
legend("topright", inset=.05,
c("onTime"), fill=c("green"), horiz=TRUE)
```

PL6 Mã nguồn R vẽ kết quả đồ họa hai hàm density chồng lên nhau

```
#set your working directory where your file data locates
setwd('/home/thuy1/git/predictUsingProbability/Preprocess')
#load the occurrences of very high step length belonged to class onTime
onTime=read.table(file="freqVeryhigh_OnTime.csv")[,1]
density(onTime, kernel=c("gaussian"))
plot(density(onTime, kernel=c("gaussian")),ylim=c(0.0, 1),
    main="Density of the occurrences of very high step
            length \n using the Gaussian smoothing kernel",
    ylab="density", xlab="the occurrences of very high step length",
    col="green")
#load the occurrences of very high step length belonged to class lateTime
lateTime=read.table(file="freqVeryhigh_LateTime.csv")[,1]
density(lateTime, kernel=c("gaussian"))
lines(density(lateTime, kernel=c("gaussian")),col="yellow")
#load add comment into the picture
legend("topright", inset=.05, title="density",
      c("onTime","lateTime"), fill=c("green","yellow"), horiz=TRUE)
```

PL7 Với xs là các điểm lấy mẫu, h là bandwidth, viết hàm tính myKDE cho điểm mới t

```
d <- density(xs)
h = d$bw
myKDE <- function(t){
        kernelValues <- rep(0,length(xs))
        for(i in 1:length(xs)){
                transformed = (t - xs[i]) / h
                kernelValues[i] <- dnorm(transformed, mean = 0, sd = 1) / h
        }
        return(sum(kernelValues) / length(xs))
}
```

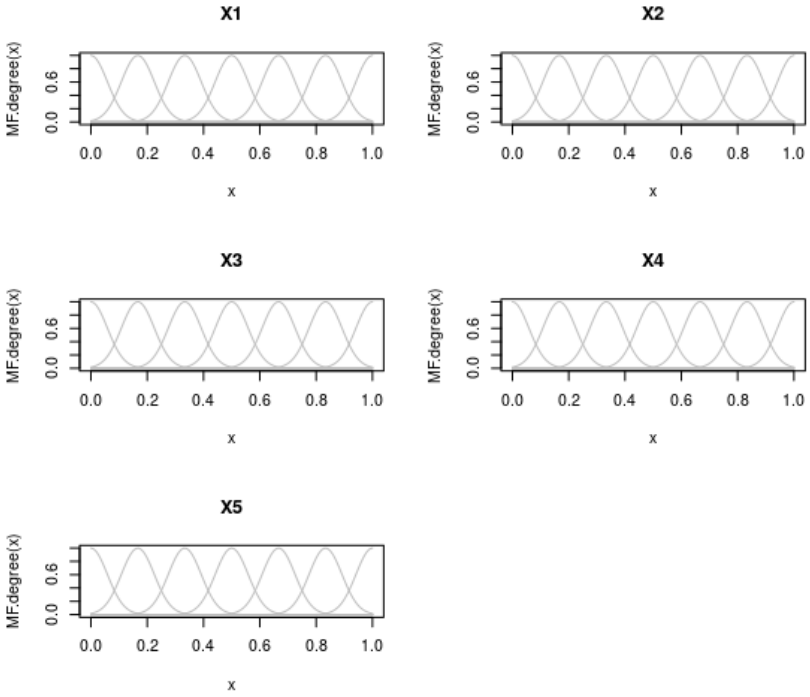
PL8 Mã nguồn R chạy giải thuật phân loại

```
setwd('/home/thuy1/git/predictUsingProbability/Preprocess')
freqVerySmall_OnTime=read.table(file="freqVerySmall_OnTime.csv")[,1]
plot(density(freqVerySmall_OnTime, kernel=c("gaussian")),
        main="the Gaussian smoothing kernel of Very Small Step",
        ylab="density", xlab="the occurrences of Very Small step",
        col="green",
        xlim=c(-10,100))
rug(freqVerySmall_OnTime)
freqVerySmall_OnTime_KDE<-density(freqVerySmall_OnTime)
freqVerySmall_OnTime_BW=freqVerySmall_OnTime_KDE$bw
freqVerySmall_OnTime_KDE_Estimation<-function(t){
    kernelValues <- rep(0,length(freqVerySmall_OnTime))
    for(i in 1:length(freqVerySmall_OnTime)) {
        transformed = (t - freqVerySmall_OnTime[i]) / freqVerySmall_OnTime_BW
        kernelValues[i] = dnorm(transformed, mean = 0, sd = 1) / freqVerySmall_OnTime_BW
    }
    return(sum(kernelValues)/length(freqVerySmall_OnTime))
}
classify <- function(x, output) {
        p_X_OnTime <- freqVerySmall_OnTime_KDE_Estimation(x[1])*
                freqSmall_OnTime_KDE_Estimation(x[2])*
                freqMedium_OnTime_KDE_Estimation(x[3])*
                freqHigh_OnTime_KDE_Estimation(x[4])*
                freqVeryHigh_OnTime_KDE_Estimation(x[5])
        p_X_LateTime <- freqVerySmall_LateTime_KDE_Estimation(x[1])*
                freqSmall_LateTime_KDE_Estimation(x[2])*
                freqMedium_LateTime_KDE_Estimation(x[3])*
                freqHigh_LateTime_KDE_Estimation(x[4])*
                freqVeryHigh_LateTime_KDE_Estimation(x[5])
        p_onTime_x=p_X_OnTime*p_onTime / (p_X_OnTime*p_onTime+p_X_LateTime*p_lateTime)
        p_lateTime_x=p_X_LateTime*p_lateTime / (p_X_OnTime*p_onTime+p_X_LateTime*p_lateTime)
        class=ifelse(p_onTime_x>p_lateTime_x,"onTime","lateTime")
        cat(paste(x[1],x[2], x[3], x[4], x[5], class, sep=","), file= output, append = T, fill = T)
}
apply(testData, 1, classify, output = 'classify.txt')
```

PL9 Mã nguồn R chạy giải thuật FRBCS.W

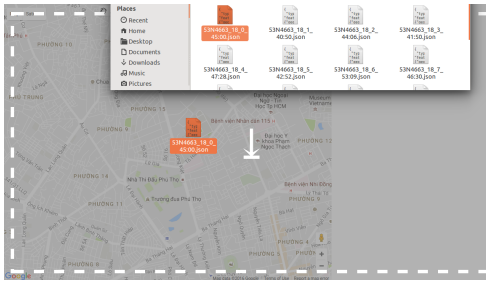
```
library(frbs)
#set your working directory where your file data locates
setwd('/home/thuy1/git/predictUsingProbability/Preprocess')
data=read.table(file="classifyRoute.csv", header=FALSE, sep=",",
        col.names = c("X1","X2","X3","X4","X5","Clazz"));
#The dataset is shuffled
dataShuffled <- data[sample(nrow(data)),]
#the last column is the output variable/attribute and it must be expressed in numbers (numerical data)
dataShuffled[,6] <- unclass(dataShuffled[,6])
#dataset divided into training and testing data
#100 first records is training data
tra.data <- dataShuffled[1:100,]
#90 last records is testing data and their output variable/attribute is removed
tst.data <- dataShuffled[100:nrow(dataShuffled),1:5]
#make the output variable/attribute of testing data an 1-column matrix
real.data <- matrix(dataShuffled[100:nrow(dataShuffled),6], ncol = 1)
#return the range [min, max] of every input variable/attribute
range.data.input <- matrix(apply(data[, -ncol(data)], 2, range), nrow = 2)
range.data.input
#         [,1]    [,2]    [,3]    [,4]    [,5]
#[1,]    4       8       15      3       0
#[2,]    83      74      68      35      11
# Set the method and its parameters. In this case we use FRBCS.W algorithm
method.type <- "FRBCS.W"
control <- list(num.labels = 7, type.mf = "GAUSSIAN", type.tnorm = "MIN",
                type.snorm = "MAX", type.implication.func = "ZADEH")
# Learning step: Generate fuzzy model
object.cls <- frbs.learn(tra.data, range.data.input, method.type, control)
# Predicting step: Predict newdata
res.test <- predict(object.cls, tst.data)
# Display the FRBS model
summary(object.cls)
# Plot the membership functions
plotMF(object.cls)
```

Hình vẽ thể hiện giải thuật FRBCS.W

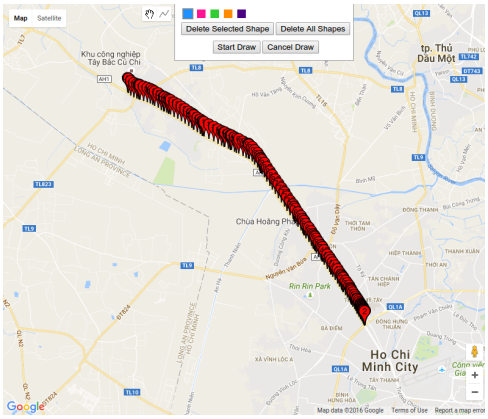


PL10 Ứng dụng bản demo "Drag And Drop data layer GeoJSON" [1] của Google Map API để kiểm tra lộ trình

Hình 5.2: Ứng dụng bản demo "Drag And Drop data layer GeoJSON"



(a) Thả file có định dạng lớp dữ liệu GeoJSon vào bản đồ



(b) Lớp dữ liệu GeoJson vừa thả xuất hiện trên bản đồ

PL11 Cách thức kết nối R với Java
Thiết lập trên hệ điều hành Ubuntu:

- Cài Oracle JDK 8:

```
sudo add-apt-repository ppa:webupd8team/java
sudo apt-get install oracle-java8-installer
sudo update-alternatives –config java
#Chọn đường dẫn Java mặc định /usr/lib/jvm/java-8-oracle/jre/bin/java
```

[1] https://developers.google.com/maps/documentation/javascript/examples/layer-data-dragndrop

- Cài môi trường R

```
  sudo apt-get install r-base
#cài đặt công cụ lập trình R, ví dụ RStudio, sau đó chạy các câu lệnh sau
  install.packages("rJava")
#Nếu bước cài đặt môi trường JDK không đúng, sẽ báo lỗi
#"configure: error: Cannot compile a simple JNI program.
#Make sure you have Java Development Kit installed and correctly registered in R."
  R.home()
#Lấy đường dẫn của R_HOME
```

- Thiết lập kết nối R sử dụng công cụ lập trình Eclipse

```
#Thêm thư viện R, sử dụng "Configure Build Path"
  Thêm /home/thuy1/R/i686-pc-linux-gnu-library/3.3/rJava/jri/JRI.jar
  vào tab Libraries
#Trước khi chạy cần phải thiết lập biến môi trường kết nối đến R,
#sử dụng Run Configuration
  Thêm -Djava.library.path=/home/thuy1/R/i686-pc-linux-gnu-library/3.3/rJava/jri/
  vào tab VM Arguments
  Thêm R_HOME vào tab Environment
```

PL12 Mã nguồn Java tính khoảng cách giữa 2 tọa độ

```
public static float distFrom(double lat1, double lng1, double lat2, double lng2) {
    double earthRadius = 6371000; //meters
    double dLat = Math.toRadians(lat2-lat1);
    double dLng = Math.toRadians(lng2-lng1);
    double a = Math.sin(dLat/2) * Math.sin(dLat/2) +
               Math.cos(Math.toRadians(lat1)) * Math.cos(Math.toRadians(lat2)) *
               Math.sin(dLng/2) * Math.sin(dLng/2);
    double c = 2 * Math.atan2(Math.sqrt(a), Math.sqrt(1-a));
    float dist = (float) (earthRadius * c);
    return dist;
}
```

PL13 Mã nguồn Java tính gom nhóm theo chuẩn 20 giây

```
function filterStandardResponse(List<Long> responseDurationList) {
    List<Long> nonStandardList = new ArrayList<Long>();
    for (Long duration: responseDurationList) {
        if (duration % 20 == 0) {
            break2StandardResponse(duration);
        } else {
            nonStandardList.add(duration);
        }
    }
    sort(nonStandardList);
    accumulate2ElmntToStandardResponse();
}
```

```
function accumulate2ElmntToStandardResponse(List<Long> sortedNonStandardDurationList) {
    List<Long> blockIndexList = new ArrayList<Long>();
    int maxIndex = sortedNonStandardDurationList.size();
    for (int i = 0; i <= maxIndex - 2; i++) {
        for (int j = 0; j <= maxIndex - 1; j++) {
            long duration1 = sortedDuration.get(i);
            long duration2 = sortedDuration.get(j);
            long sumDuration = duration1+duration2;
            if (sumDuration % 20 == 0 && !blockList.contains(i) && !blockList.contains(j)) {
                blockList.add(i);
                blockList.add(j);
                break2StandardResponse(duration);
            }
        }
    }
    List<Long> newSortedNonStandardList = new ArrayList<Long>();
    for (int i = 0; i < maxIndex; i++) {
            if (!blockList.contains(i)) {
                newSortedNonStandardList.add(sortedNonStandardDurationList.get(i));
            }
    }
    if (newSortedNonStandardList.size() >= 3) {
            accumulate3ElmntToStandardResponse(newSortedNonStandardList);
    }
```

Các hàm accumulate3ElmntToStandardResponse và accumulate4ElmntToStandardResponse tương tự như hàm accumulate2ElmntToStandardResponse để thực hiện gom nhóm 3 thành phần, 4 thành phần thành bội số 20, cuối cùng với những thành phần chưa được gom nhóm. Tiếp theo, thực hiện việc gom nhóm các thành phần liên tiếp thành bội số 20. Cuối cùng, thực hiện gom nhóm các thành phần liên tiếp sao cho chia cho 20 có số dư nhỏ nhất

```
function accumulateSequenceElmtToStandardResponse(List<Long> sortedNonStandardDurationList) {
    int i = 0;
    int j = 1;
    int maxIndex = sortedNonStandardDurationList.size();
    while (i<maxIndex) {
        long sumDuration = sortedDuration.get(i);
        while (j < maxIndex) {
            sumDuration += sortedDuration.get(j);#sum of continuous elements
            if (sumDuration%20==0) {
                break2StandardResponse(sumDuration);
                i=j+1;#create new caculation with start index equal j+1
                j=i+1;
                break;
            } else {
                j++;#continue until sumDuration%20 equal 0
            }
        }
        if (j == maxIndex) {
            break;
        }
    }
    List<Long> notAccumulateElmtList = new ArrayList<Long>();
    while (i<maxIndex) {
        notAccumulateElmtList.add(sortedNonStandardDurationList.get(i));
        i++;
    }
    accumulateSequenceElmtWithMinRedundant(notAccumulateElmtList);
```

```
function accumulateSequenceElmtWithMinRedundant(List<Long> sortedNonStandardDurationList) {
    int i = 0;
    int j = 1;
    int stopIndex = 0;
    int maxIndex = sortedNonStandardDurationList.size();
    while (i<maxIndex) {
        long sumDuration = sortedNonStandardDurationList.get(i);
        stopIndex = i;
        long compareNumber = 20;
        if (sumDuration/20>0) {#if sumDuration > 20
            compareNumber = 20*(sumDuration/20);
        }
        long min = Math.abs(sumDuration - compareNumber);
        while (j < maxIndex) {
            sumDuration += sortedNonStandardDurationList.get(j);
            long currentCompareNumber = 20;
            if (sumDuration/20>0) {#if sumDuration > 20
                currentCompareNumber = 20*(sumDuration/20);
            }
            long currentMin = Math.abs(sumDuration - currentCompareNumber);
            if (currentMin <= min) {
                stopIndex = j;
                min = currentMin;
            }
            j++;
        }
        if (stopIndex != i) {
            sumDuration = 0;
            for (int k = i; k <= stopIndex; k++) {
                sumDuration += sortedNonStandardDurationList.get(k);
            }
            break2StandardResponse(sumDuration);
        }
        i = stopIndex + 1;
        j = i + 1;
    }
```

LÝ LỊCH TRÍCH NGANG

Họ và tên: Lê Thị Minh Thùy
Ngày sinh: 22/01/1986
Nơi sinh: Đồng Nai
Địa chỉ liên lạc: 741 Trương Công Định Phường 9, TP Vũng Tàu
Email: thuyltm2201@gmail.com

QUÁ TRÌNH ĐÀO TẠO

Thời gian	Trường đào tạo	Chuyên ngành	Trình độ đào tạo
2004 – 2009	Trường Đại học Bách Khoa TP.HCM	Công nghệ thông tin	Cử nhân
2013 – 2017	Trường Đại học Bách Khoa TP.HCM	Khoa học máy tính	Thạc sĩ

QUÁ TRÌNH CÔNG TÁC

Thời gian	Đơn vị công tác	Chuyên ngành
2014 – 2016	Công ty gia công phần mềm Tường Minh	Lập trình viên

Tài liệu tham khảo

Tài liệu trong nước

[1] 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

Tài liệu nước ngoài

[2] Jiawei Han, Micheline Kamber, Jian Pei (2012), *Data Mining: Concepts and Techniques (3rd ed.)*, Morgan Kaufmann Publishers, USA.

Website

[3] *2 ways of using Naive Bayes classification for numeric attributes*, truy cập ngày 1 tháng 3 năm 2017, địa chỉ *http://www.simafore.com/blog/bid/107702/2-ways-of-using-Naive-Bayes-classification-for-numeric-attributes.*.

[4] *The Gaussian classifier*, truy cập ngày 2 tháng 3 năm 2017, địa chỉ *http://www.svcl.ucsd.edu/courses/ece271A/handouts/GC.pdf*.

[5] *Naive Bayes 3: Gaussian example*, truy cập ngày 2 tháng 3 năm 2017, địa chỉ *https://www.youtube.com/watch?v=r1in0YNetG8*.

[6] *L7: Kernel density estimation*, truy cập ngày 26 tháng 3 năm 2017, địa chỉ *http://www.stat.washington.edu/courses/stat539/spring13/Handouts/tamu-csce-666-pr_l7-annotated.pdf*.

[7] *Kernel density estimation Wikipedia*, truy cập ngày 26 tháng 3 năm 2017, địa chỉ *https://en.wikipedia.org/wiki/Kernel_density_estimation*.

[8] *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/*.

[9] *Find-the-probability-density-of-a-new-data-point-using-density-function-in-r*, truy cập ngày 4 tháng 6 năm 2017, địa chỉ *https://stackoverflow.com/questions/28077500/find-the-probability-density-of-a-new-data-point-using-density-function-in-r*

Developer's Website

[10] *Google Maps 3 API - Data Layer: GeoJSON*, truy cập ngày 6 tháng 11 năm 2016, địa chỉ *https://developers.google.com/maps/documentation/javascript/examples/layer-data-style*.

[11] *Google Maps 3 API - Click on feature (from geojson)*, truy cập ngày 6 tháng 11 năm 2016, địa chỉ *http://stackoverflow.com/questions/29309856/google-maps-3-api-click-on-feature-from-geojson-and-check-if-it-contains-loc*.

[12] *Google Maps 3 API - Data Layer: Drag and Drop GeoJSON*, truy cập ngày 6 tháng 11 năm 2016, địa chỉ *https://developers.google.com/maps/documentation/javascript/examples/layer-data-dragndrop*.

[13] *Google Maps 3 API - Waypoints in directions*, truy cập ngày 6 tháng 11 năm 2016, địa chỉ *https://developers.google.com/maps/documentation/javascript/examples/directions-waypoints*.

[14] *Google Maps 3 API - Distance Matrix*, truy cập ngày 6 tháng 11 năm 2016, địa chỉ *https://developers.google.com/maps/documentation/javascript/examples/distance-matrix*.