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

LÊ THỊ MINH THÙY

DỰ ĐOÁN XÁC SUẤT THỜI GIAN XE BUÝT VỀ ĐÚNG TRẠM

Chuyên ngành: Khoa học máy tính
Mã số: 60.48.01

LUẬN VĂN THẠC SĨ

TP.Hồ Chí Minh, Ngày 20 tháng 2 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 20 tháng 2 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 thời gian 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:

IV. NGÀY HOÀN THÀNH NHIỆM VỤ:

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

Tp. HCM, Ngày 20 tháng 2 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 chân thành cảm ơn 4 năm được học tập chương trình Thạc Sĩ Khoa Học Máy Tính tại trường Đại Học Bách Khoa cùng với sự giảng dạy tận tình của thầy cô và cơ hội trao đổi kinh nghiệm quý báu với 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 dịch tiếng Việt tên là "Thống kê Công nghiệp hiện đại với ứng dụng viết trên R, MINITAB và JMP", 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 mình chỉ sử dụng chúng như là bằng chứng kham khảo cho ngày bảo vệ hội đồng Thạc Sĩ, không dùng mục đích thương mại hay bất kỳ hoạt động khoa học nào, nghĩa là, tôi cam đoan không đem nội dung Luận Văn này để mua bán, đăng báo hay thuyết trình bất cứ đâu.
Đồng thời, tôi khuyến cáo người đọc nên tiếp cận nội dung lý thuyết phiên bản tiếng Anh.

Ngày 20 tháng 2 năm 2017
Lê Thị Minh Thùy

TÓM TẮT LUẬN VĂN

ABSTRACT

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 20 tháng 2 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 Toán định lượng kinh nghiệm nhắm chừng này.
Có dữ liệu di chuyển 2/3 chặng đi từ bến xe An Sương đến bến xe Củ Chi của tuyến xe buýt 74 và kết quả tương ứng về trạm đích: đúng giờ hay trễ giờ. Sử dụng thao tác khai phá dữ liệu học có quan sát, phân chia dữ liệu 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.
Do quan sát kích thước mẫu giới hạn, ngẫu nhiên các chuyến di chuyển từ bến xe An Sương đến bến xe Củ Chi (tuyến xe buýt 74 ngẫu nhiên được chọn) kèm theo kết quả về trạm đích đúng giờ hay trễ giờ tương ứng. Vì kích thước mẫu giới hạn, không sử dụng sức mạnh tính toán hiệu năng cao trên kích thước dữ liệu siêu lớn, Luận Văn dùng kiến thức Thống Kê, làm việc trên kích thước mẫu chấp nhận được so với trên tổng thể quá to lớn nhưng đủ rút ra đặc trưng cho tổng thể để giải quyết bài toán trên.
Ngoài ra, Luận Văn dùng kiến thức xác suất, con số đo khả năng một sự kiện xảy ra vì ta chấp nhận dự đoán dựa trên 2/3 chặng đường ban đầu, trong khi đó 1/3 chặng đường sau vẫn có tác động đến kết quả dự đoán. Khi dùng xác suất, ta chấp nhận bài toán có sự tham gia không chắc chắn, có thay đổi do ngẫu nhiên.
Trên thực tế rất hiếm hai chuyến xe cùng lộ trình có những bước di chuyển cách nhau 20 giây giống nhau hoàn toàn, do đó Luận Văn dùng ước lượng tham số trong kiến thức Thống Kê để mô hình hóa di chuyển.
Công thức xác suất Bayes $\mathbb{P}(B|A) = \frac{\mathbb{P}(B)\cdot\mathbb{P}(A|B)}{\mathbb{P}(A)}$ phù hợp để giải bài toán dán nhãn dự đoán về trạm đích đúng giờ hay trễ giờ (biến cố B) cho di chuyển 2/3 chặng của tuyến đi cố định từ bến xe An Sương đến bến xe Củ Chi (biến cố A), 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.
Tổng kết lại, mục đích Luận Văn: dùng kiến thức Thống kê, xác suất, công thức xác suất Bayes để mô hình dữ liệu học, dùng mô hình này để dán nhãn đúng giờ hay trễ giờ cho dữ liệu kiểm tra.

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

Luận Văn bó hẹp với đề bài cụ thể: Dùng Toán Thống Kê định lượng kinh nghiệm di chuyển trên cùng một lộ trình để dán nhãn về trạm đúng giờ hay trễ giờ.
Luận Văn không đặt mục tiêu tham vọng giải bài toán này có thể đem ra ứng dụng thực tế được vì bản thân người giải bài toán không có nhiều kinh nghiệm với Toán Thống Kê, làm việc với kích thước mẫu rất nhỏ có thể đưa ra dự đoán đúng trên kích thước quần thể.
Ngoài ra, người lái xe phải hiểu rằng kết quả dự đoán về trạm đích đúng giờ chỉ mang tính xác suất. Ví dụ, sau khi đi được 2/3 chặng, bác tài xế nhận được kết quả xác suất 98.5% xe về trạm đúng giờ. Nhưng gặp sự cố tại 1/3 chặng đường cuối, tình trạng kẹt xe nặng, xe về trạm không đúng giờ như kết quả thông báo trên với con số xác suất đúng giờ xảy ra rất lớn 98.5%. Bác tài xế phải chấp nhận xác suất 1.5% về trạm trễ giờ vẫn có khả năng xảy ra do bác tài xế không thể điều khiển tình trạng xe chạy tại 1/3 chặng đường cuối. Nhưng các con số xác suất được thông báo cho bác tài vẫn có ý nghĩa chừng mực nếu bác tài xế biết rằng các con số xác suất này định lượng kinh nghiệm những chuyến di chuyển trong quá khứ để dự báo tương lai cho chuyến đi hiện tại. Khi đã 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, người đọc chấp nhận bài toán được giải trong Luận Văn này phần nào có ý nghĩa.
Tóm lại, tác giả nhắc lại mục tiêu rất cụ thể của Luận Văn này: Dùng Toán Thống Kê định lượng kinh nghiệm di chuyển trên cùng một lộ trình để dán nhãn về trạm đúng giờ hay trễ giờ.

1.4 Phương pháp nghiên cứu

Phương pháp nghiên cứu là thu thập dữ liệu di chuyển hữu hạn, quan sát các bước di chuyển được ghi nhận cách nhau 20 giây trên 2/3 lộ trình đi từ bến xe An Sương đến bến xe Củ Chi, trực quan hóa dữ liệu này, cộng thêm kiến thức quan sát thực tế, nếu có nhiều bước di chuyển dài trong từng khoảng 20 giây thì chuyến đi đó có khả năng về trạm đúng giờ cộng thêm kết quả dán nhãn về trạm đúng giờ hay trễ giờ chỉ cho ta biết khả năng xảy ra, chứ không phải chắc chắn. Những kiến thức này giúp ta nghĩ đến kiến thức Thống Kê, xác suất chủ đạo và từ đó từng bước tìm thuật toán, công thức phù hợp để giải. Tóm lại, nhờ quan sát dữ liệu cho ta định hướng cách tiếp cận, cách nghiên cứu.

1.5 Một số kết quả thu được

1.6 Cấu trúc luận văn

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 Bến xe An Sương đến Bến xe Củ Chi 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 Bến xe An Sương đến Bến xe Củ Chi 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 Bến xe An Sương đến Bến xe Củ Chi 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 bến xe An Sương - bến xe Củ Chi, 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 $\mathcal{S}$.

Xác suất của biến cố

Thông thường chúng ta gộp tất cả các biến cố vào một tập $\mathcal{Q} := \{A : A \subset \mathcal{S} \text{ là một biến cố}\}$, 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, $\mathbb{P}(\text{A}) \geq 0$
- **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 $\geq$ 1) là các biến cố rời nhau từng đôi một thì

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

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 $\cap$ 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$\cap$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. Điều 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

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

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

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

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 $\mathcal{S}$, $\mathcal{S} = \bigcup_{j=1}^{m} B_j$, 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:

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

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 ∈ $\mathcal{S}_X$. 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

Ví dụ phân loại dựa vào công thức 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

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)

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*?
Từ bảng 2.2, ta tính thêm thông tin xác suất

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

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

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*

2.2 Thuật toán Jenks natural breaks optimization (Tùy chọn)

Kham khảo từ [3]

2.3 Thuật toán Kernel Density Estimation (Tùy chọn)

Chương 3

HIỆN THỰC VÀ THỬ NGHIỆM

3.1 Dữ liệu nghiên cứu

Trích ngẫu nhiên 80% thời gian di chuyển của 24 chuyến xe trong 351 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, 71, 86, 86, 145, 145, 136, 136, 104, 104, 277, 277, 240, 240, 145, 145, 97, 97, 0, 0, 171, 171, 283, 283, 283, 283, 0, 0, 214, 214, 166, 166, 307, 307, 259, 259, 208, 208, 43, 43, 144, 144, 286, 286, 287, 287, 126, 126, 152, 152, 254, 254, 139, 139, 255, 255, 294, 294, 293, 293, 269, 269, 283, 283, 267, 267, 221, 221, 165, 165, 266, 266, 259, 259, 220, 220, 168, 168, 0, 0, 226, 226, 337, 337, 394, 394, 224, 224, 208, 208, 340, 340, 359, 359, 259, 259, 297, 297, 325, 325, 303, 303, 302, 302, 284, 284, 287, 287, 268, 268, 199, 199, 112, 112, 179, 179, 279, 279, 270, 270, 139, 139, 72, 72, 195, 195, 345, 345, 332, 332, 151, 151, 149, 149, 230, 230, 261, 261, 186, 186, 170, 170, 236, 236, 243, 243, 265, 265, 151, 151, 201, 201, 29, 29, 125, 125, 121, 121, 35, 35, 29, 29, 80, 80, 46, 46, 36, 36, 132, 132, 52, 52, 44, 44, 63, 63, 63, 46, 46, 46, 40, 40, 23

Chuyến 2:...

...

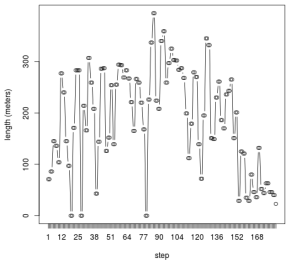
Chuyến 12: 104, 104, 145, 145, 124, 124, 166, 166, 194, 194, 205, 205, 188, 188, 158, 158, 107, 107, 121, 121, 91, 91, 152, 152, 218, 218, 93, 93, 100, 100, 0, 0, 167, 167, 167, 167, 250, 250, 160, 160, 144, 144, 223, 223, 215, 215, 131, 131, 114, 114, 241, 241, 206, 206, 193, 193, 0, 0, 0, 0, 169, 169, 212, 212, 192, 192, 200, 200, 257, 257, 223, 223, 198, 198, 166, 166, 135, 135, 184, 184, 105, 105, 90, 90, 151, 151, 202, 202, 192, 192, 168, 168, 177, 177, 217, 217, 168, 168, 243, 243, 211, 211, 109, 109, 161, 161, 272, 272, 287, 287, 281, 281, 248, 248, 257, 257, 166, 166, 261, 261, 250, 250, 152, 152, 238, 238, 267, 267, 190, 190, 111, 111, 0, 0, 117, 117, 138, 138, 209, 209, 182, 182, 123, 123, 58, 58, 151, 151, 222, 222, 248, 248, 218, 218, 110, 110, 0, 0, 0, 0, 111, 111, 211, 211, 61, 61, 114, 114, 83, 83, 109, 109, 26, 26, 232, 232, 232, 232, 232, 232, 232, 232, 232, 232, 96, 96, 37, 37, 110, 110, 48, 48, 34, 34, 121, 121, 121, 46, 46, 33, 33, 88, 88, 64, 64, 235, 235, 183, 183, 125, 125, 220, 22

Chuyến 13:...

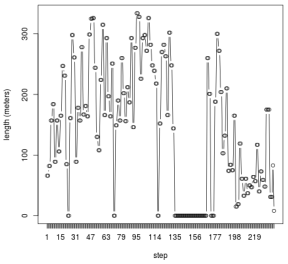
...

Chuyến 24: 279, 279, 17, 17, 52, 52, 48, 48, 67, 67, 141, 141, 131, 131, 131, 131, 95, 95, 157, 157, 155, 155, 196, 196, 242, 242, 219, 219, 50, 50, 172, 172, 219, 219, 243, 243, 148, 148, 74, 74, 5, 5, 105, 105, 194, 194, 68, 68, 216, 216, 224, 224, 231, 231, 61, 61, 76, 76, 228, 228, 295, 295, 306, 306, 293, 293, 260, 260, 236, 236, 279, 279, 284, 284, 232, 232, 174, 174, 75, 75, 205, 205, 103, 103, 98, 98, 232, 232, 266, 266, 257, 257, 188, 188, 60, 60, 192, 192, 228, 228, 256, 256, 236, 236, 215, 215, 23, 23, 10, 10, 100, 100, 44, 44, 132, 132, 255, 255, 290, 290, 272, 272, 216, 216, 237, 237, 281, 281, 274, 274, 272, 272, 294, 294, 189, 189, 0, 0, 41, 41, 84, 84, 65, 65, 197, 197, 247, 247, 221, 221, 121, 121, 83, 83, 138, 138, 252, 252, 223, 223, 195, 195, 135, 135, 17, 17, 0, 0, 6, 6, 15, 15, 44, 44, 121, 121, 61, 61, 206, 206, 220, 220, 148, 148, 52, 52, 164, 164, 208, 208, 248, 248, 203, 203, 179, 179, 11, 11, 0, 0, 62, 62, 112, 112, 252, 252, 307, 307, 316, 316, 168, 168, 104, 104, 87, 87, 192, 192, 148, 148, 131, 131, 91, 91, 215, 215, 215, 143, 143, 126, 126, 279, 279, 27

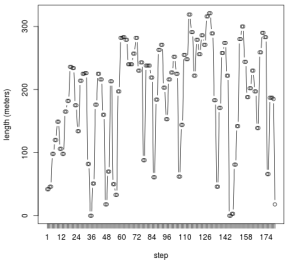
Trực quan hóa giá trị các bước di chuyển của 24 chuyến dữ liệu ngẫu nhiên, ta có



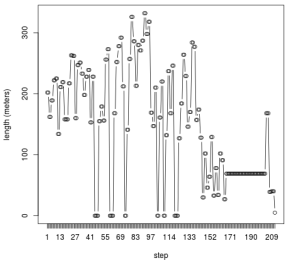
80% thời gian di chuyển
Chuyến xe 1-đúng giờ



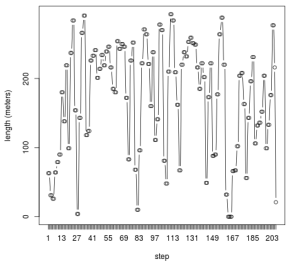
80% thời gian di chuyển
Chuyến xe 2-trễ giờ



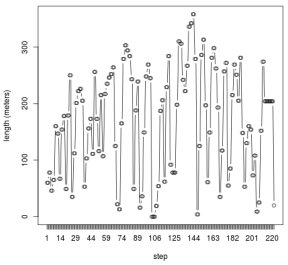
80% thời gian di chuyển
Chuyến xe 3-đúng giờ



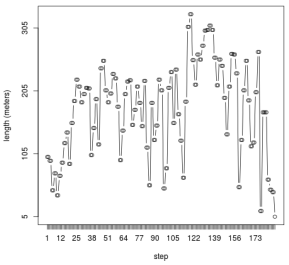
80% thời gian di chuyển
Chuyến xe 4-trễ giờ



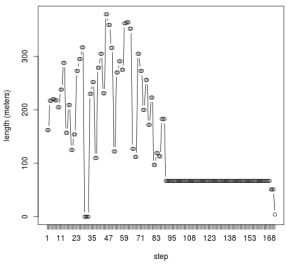
80% thời gian di chuyển
Chuyến xe 5-đúng giờ



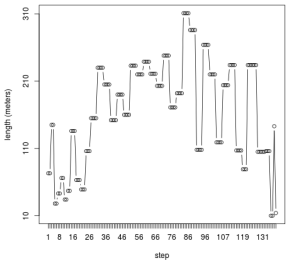
80% thời gian di chuyển
Chuyến xe 6-trễ giờ



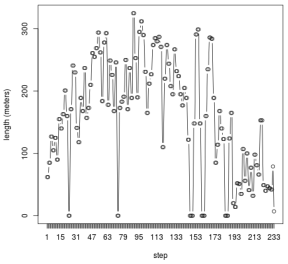
80% thời gian di chuyển
Chuyến xe 7-đúng giờ



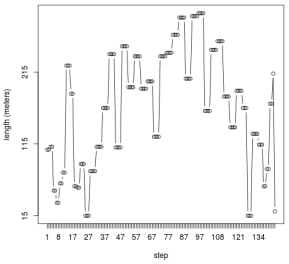
80% thời gian di chuyển
Chuyến xe 8-trễ giờ



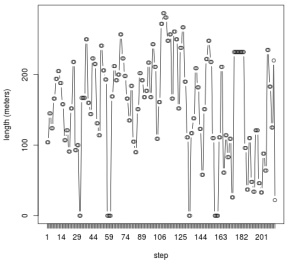
80% thời gian di chuyển
Chuyến xe 9-đúng giờ



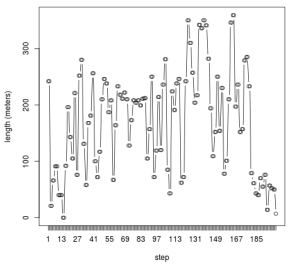
80% thời gian di chuyển
Chuyến xe 10-trễ giờ



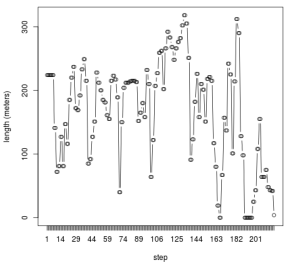
80% thời gian di chuyển
Chuyến xe 11-đúng giờ



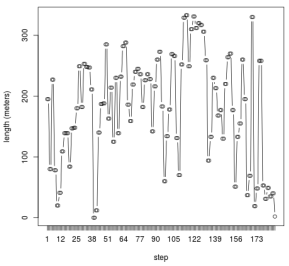
80% thời gian di chuyển
Chuyến xe 12-trễ giờ



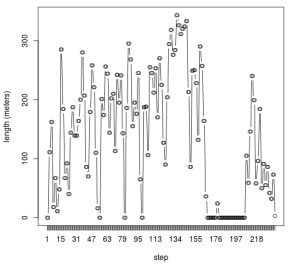
80% thời gian di chuyển
Chuyến xe 13-đúng giờ



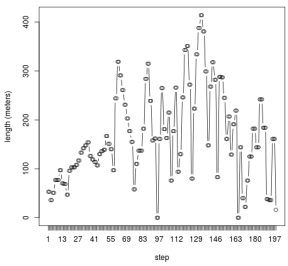
80% thời gian di chuyển
Chuyến xe 14-trễ giờ



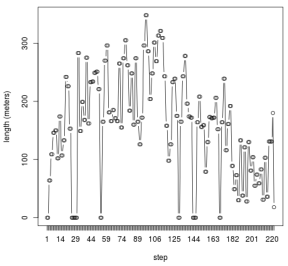
80% thời gian di chuyển
Chuyến xe 15-đúng giờ



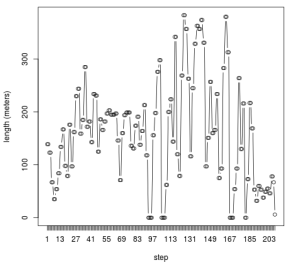
80% thời gian di chuyển
Chuyến xe 16-trễ giờ



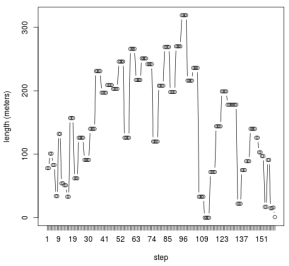
80% thời gian di chuyển
Chuyến xe 17-đúng giờ



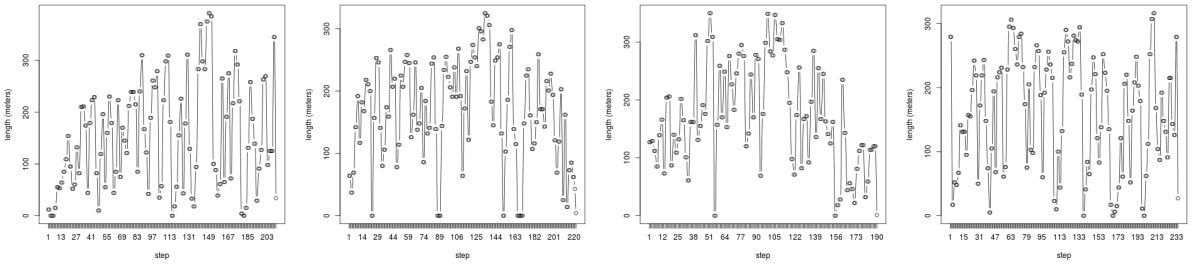
80% thời gian di chuyển
Chuyến xe 18-trễ giờ



80% thời gian di chuyển
Chuyến xe 19-đúng giờ



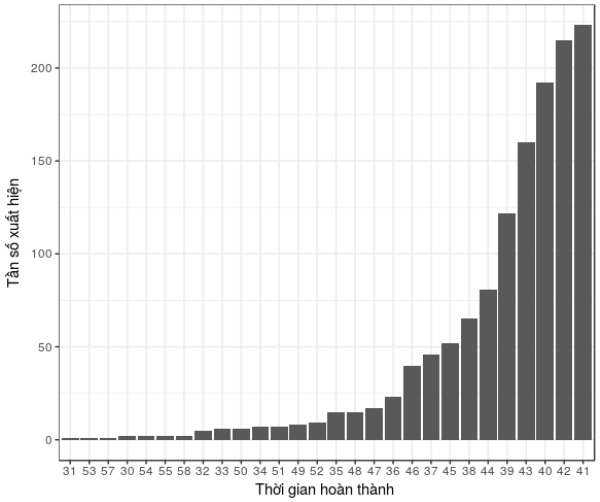
80% thời gian di chuyển
Chuyến xe 20-trễ giờ



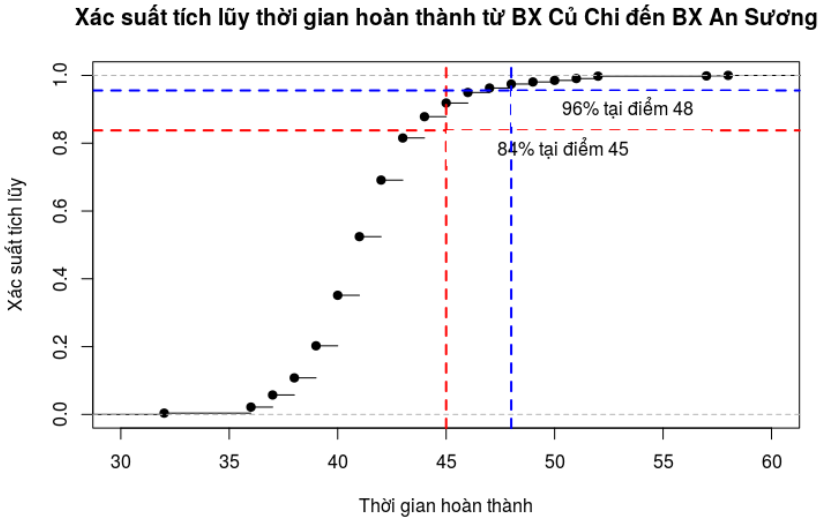
80% thời gian di chuyển Chuyến xe 21-đúng giờ | 80% thời gian di chuyển Chuyến xe 22-trễ giờ | 80% thời gian di chuyển Chuyến xe 23-đúng giờ | 80% thời gian di chuyển Chuyến xe 24-trễ giờ

Hình 3.1: Trực quan hóa 24 dòng dữ liệu ngẫu nhiên

Hình vẽ sắp xếp thời gian hoàn thành lộ trình BX Củ Chi - BX An Sương xác suất xuất hiện tăng dần.



Hình vẽ cho thấy xác suất tích lũy thời gian hoàn thành lộ trình BX Củ Chi - BX An Sương



3.2 Thu thập thông tin rút ra từ dữ liệu lịch sử

Theo bài toán đặt ra, ta sẽ lấy thông tin dữ liệu 2/3 chuyến và phân loại thời gian đến trạm đích: đúng giờ và trễ.

3.3 Tiến hành dự đoán xác suất xe buýt về trạm đúng thời gian

3.4 Kết luận

Chương 4

KẾT LUẬN

4.1 Tổng kết

4.2 Đóng góp của đề tài

Vì các giá trị xác suất ... được tính trên mẫu và Luận Văn này chưa chứng minh được các xác suất đấy có thể đúng trên tổng thể với xác suất xảy ra cao, cho nên kết quả của Luận Văn này không sử dụng được trên thực tế. Nhưng đóng góp của Luận Văn này là tìm được một trong nhiều phương pháp giải có thể cho câu hỏi "Dự báo xác suất xe về trạm đúng giờ khi xe đã đi được một 2/3 khoảng đường"

4.3 Hướng phát triển

Để giải quyết bài toán của Luận Văn này triệt để hơn có hai cách tiếp cận:

- Cách 1: Giải bài toán này với dữ liệu cực lớn để tìm gần chính xác các xác suất .....Khi đó, tôi dự đoán sẽ nảy sinh thêm bài toán mới, xử lý các vấn đề khi làm việc trên dữ liệu cực lớn.
- Cách 2: Vẫn xử lý trên dữ liệu mẫu nhỏ hơn rất nhiều so với tổng thể, người giải phải học thêm kiến thức Toán Thống Kê như các phương pháp thiết kế thí nghiệm, xác định phương pháp lấy mẫu chấp nhận được, phương pháp kiểm định giả thuyết, ..., để đủ tự tin chứng minh rằng kết quả thu được từ mẫu cũng phản ánh đúng trên tổng thể với xác suất xảy ra cao.

Chương 5

PHỤ LỤC

1. Source code R chạy giải thuật Jenks Natural breaks optimization

```
#set your working directory where your file data locates
setwd("/home/thuy/workspace/Preprocess")
mydata = read.csv("Freq_CC-AS_80.csv",header = FALSE, sep = ",")[,1]
#use library classInt to run algorithm Jenks Natural breaks optimization
library("classInt")
classIntervals(mydata, n=3, style="jenks")
#returned result
#[2,170] (170,345] (345,2578]
#118 99 1
#or use library BAMMtools to run algorithm Jenks Natural breaks optimization
library("BAMMtools")
getJenksBreaks(mydata, 3)
#returned result
#2 345 2578
```

2. Source code 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()
}
```

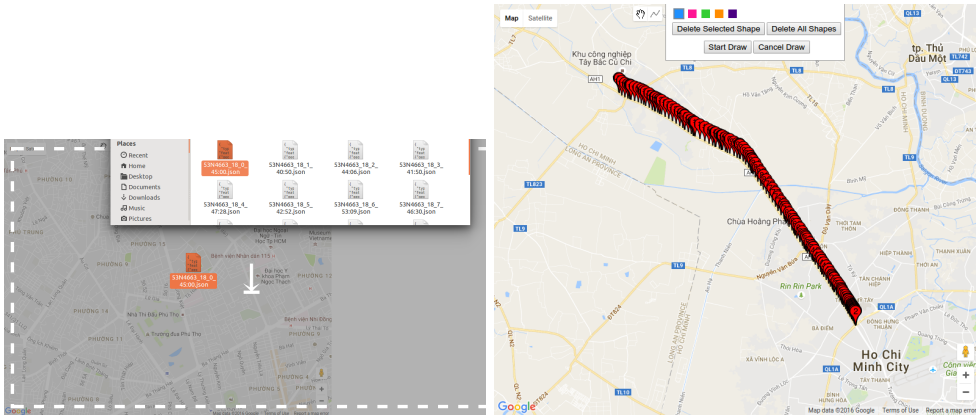
3. Source code R xuất hình thể hiện xác suất tích lũy thời gian hoàn thành BX Củ Chi - BX An Sương

```
#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)
```

4. Source code 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")
```

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



(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

Hình 5.1: Ứng dụng bản demo "Drag And Drop data layer GeoJSON"

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

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] *Jenks natural breaks optimization*, truy cập ngày 23 tháng 10 năm 2016, địa chỉ *https://en.wikipedia.org/wiki/Jenks_natural_breaks_optimization.*
Jenks natural breaks optimization example, truy cập ngày 18 tháng 2 năm 2017, địa chỉ *https://www.ehdp.com/vitalnet/breaks-1.htm.*

Developer's Website

[4] *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.*

[5] *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.*

[6] *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.*

[7] *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.*

[8] *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.*