Comp1810- Dữ liệu và phân tích web
Ghi chú nghiên cứu thi giấy

1. Đánh giá nghiêm túc các thuật ngữ của Click ClickStream, liên quan đến phân tích dữ liệu web và mô tả phương pháp cốt lõi của việc trích xuất click clickstream từ một trang web.[10 Điểm]
Clickstream:

Dữ liệu ClickStream là nhật ký chi tiết về cách người tham gia điều hướng qua web trang web trong một nhiệm vụ.Nhật ký thường bao gồm các trang được truy cập, thời gian dành cho mỗi trang, Làm thế nào họ đến trên trang, và nơi họ đi tiếp theo.Từ một tổng hợp phối cảnh, dữ liệu clickstream cung cấp những hiểu biết to lớn về cách dễ dàng trang web Điều hướng, những trang nào đang gây ra sự nhầm lẫn lớn nhất và những trang nào rất quan trọng trong đạt đến một điểm đến mong muốn.
Dữ liệu Click Stream giúp kể câu chuyện của khách hàng trong khi truy cập trang web của bạn.
Bất kỳ sự kiện nào xảy ra trong khi khách truy cập trên trang web được JavaScript thu thập Mã theo dõi được đặt trên mỗi trang HTML.Điều này cho phép theo dõi tất cả các sự kiện hoặc Làm thế nào khách hàng tham gia với các trang.Kịch bản Java có thể theo dõi những điều sau Các sự kiện như phát video, tải xuống, đăng ký, mua hàng và nhiều hơn nữa.
Phương pháp trích xuất dữ liệu ClickStream:
Google Analytics là một nền tảng trực tuyến, người dùng có thể đăng ký và đăng nhập.
Thu thập ClickStream.Người dùng có thể được cung cấp mã (JavaScript) bởi Google phân tích.Mã này được đính kèm vào mỗi trang của trang web ngay sau khi đầu Tag <Head>.Mỗi lần tải trang, ID người dùng được tạo và tất cả các hoạt động cho Ví dụ về trang xem trang web, tỷ lệ thoát, trên mạng, các giao dịch mua, thời gian dành cho các trang.Tất cả những điều n được gửi vào tài khoản Analytics hoặc bất kỳ phần mềm thu thập nào khác.

Mã JS sau đây được sử dụng để theo dõi các hoạt động trong các trang web.

Các công cụ loại bỏ web để trích xuất clickstream từ trang web
Một kỹ thuật tự động được gọi là Scraping Web được sử dụng để lấy một lượng lớn dữ liệu từ các trang web.Thông thường, Internet chứa dữ liệu phi cấu trúc.AIDS cào web trong Thu thập và lưu trữ dữ liệu phi cấu trúc.Có một số phương pháp để cạo các trang web, bao gồm việc sử dụng các dịch vụ Internet, API hoặc chương trình tùy chỉnh.Xóa web được sử dụng để thu thập một lượng lớn dữ liệu (thống kê, thông tin chung, nhiệt độ, v.v.) từ Các trang web, sau đó được phân tích và sử dụng để tiến hành khảo sát hoặc R & D.Công cụ cạo web được thiết kế để giúp các doanh nghiệp thu thập dữ liệu web nguồn mở, chẳng hạn như:

- Tổng hợp
- Làm sạch
- Cấu trúc
- Xử lý
- Được phân tích bởi các đội và thuật toán.

Củ web có thể được thực hiện thủ công, nhưng nó tốn nhiều thời gian và tốn nhiều tài nguyên
Nỗ lực, rất nhiều doanh nghiệp thích sử dụng một công cụ để giúp tự động hóa quá trình này.
Hầu hết các trường hợp sử dụng phổ biến mà các doanh nghiệp hiện đang sử dụng các công cụ cạo web:
Nghiên cứu thị trường: Các ngành công nghiệp đang tìm cách giới thiệu các sản phẩm mới hoặc tham gia tập hợp thị trườ
thông tin về đối tượng mục tiêu tiềm năng trong khi cũng điều tra đối thủ cạnh tranh thành công
Các hoạt động có thể được nhân rộng/học hỏi từ.
 Dữ liệu thị trường chứng khoán: Quỹ phòng hộ, quản lý danh mục đầu tư và các nhà đầu tư mạo hiểm tất cả
thu thập dữ liệu tài chính như năng lực thị trường cổ phiếu, các bài báo truyền thông doanh nghiệp và
Tăng trưởng dựa trên số lượng nhân viên hoặc dữ liệu không gian địa lý (ví dụ: hình ảnh vệ tinh trên
tiến trình của một trang web xây dựng hoặc nhà máy).
 Tập hợp du lịch: để cạnh tranh tốt hơn, các cơ quan du lịch trực tuyến (OTA) tập hợp thực tế
thông tin thời gian về các gói kỳ nghỉ của các trang web cạnh tranh, các ưu đãi đặc biệt và
Giá cho thuê/xe/khách sạn.
 Thị trường cung cấp thực phẩm: Khi nhu cầu cung cấp thực phẩm đã tăng lên trong hai lần cuối cùng
nhiều năm, các công ty đang ngày càng tìm cách thu thập dữ liệu thực đơn nhà hàng, xu hướng
Các món ăn thông qua tìm kiếm (Trung Quốc? Nhật Bản? v.v.), và khối lượng đặt hàng dựa trên người tiêu dùng
Địa điểm định vị.

 Bộ sưu tập các trang kết quả của công cụ tìm kiếm (SEO) / công cụ tìm kiếm
(SERP): Nhiều hành trình của người tiêu dùng bắt đầu với một truy vấn tìm kiếm đơn giản, thúc đẩy
Các doanh nghiệp lên đầu kết quả công cụ tìm kiếm.Kết quả là, họ tập hợp và phân tích hàng đầu
Kết quả tìm kiếm cho các truy vấn và từ khóa tìm kiếm có liên quan trong ngành của họ để tối ưu hóa
trang riêng của họ và xếp hạng cao hơn trong tương lai.
 Kiểm tra trang web: Các công ty xây dựng trang web/ứng dụng cho các địa lý khác nhau hoặc đó
Khởi chạy Trải nghiệm người dùng mới (UX) và Giao diện người dùng (UI) sử dụng các công cụ cạo web
Để xem kết quả mặt trước từ quan điểm của người tiêu dùng.Điều này cho phép họ cải thiện
Đảm bảo chất lượng của họ (QA) và cân bằng tải.
 Thương mại điện tử: Đây là một lĩnh vực cạnh tranh cao với nhiều giá trị có ý thức
Khách hàng.Giá sản phẩm, đánh giá của khách hàng, giá bán qua (STR) và
Điểm dữ liệu được thu thập bởi các nhà cung cấp, thị trường và thương hiệu để tối ưu hóa mặt hàng
Danh sách, thiết kế và dây chuyền sản xuất để nắm bắt tỷ lệ chuyển đổi cao hơn.
 ADTech: Các nhóm tiếp thị và các cơ quan sử dụng các công cụ cạo web để đảm bảo rằng
Các chiến dịch cục bộ được hiển thị để nhắm mục tiêu đối tượng với bản sao, hình ảnh và hình ảnh chính xác
URL.Họ cũng thu thập dữ liệu về các chiến dịch quảng cáo của đối thủ cạnh tranh để hiểu rõ hơn và
Tối ưu hóa các chiến dịch cho tỷ lệ nhấp cao hơn (CTRS).
 Phương tiện truyền thông xã hội để tiếp thị: Các công ty sử dụng các công cụ cạo web để hiểu rõ hơn về
Tình cảm xã hội của đối tượng mục tiêu của họ, để tìm những người có ảnh hưởng mà họ có thể
hợp tác và để xác định các bài đăng mà người tiêu dùng đang tham gia để họ có thể
Tham gia tường thuật và tạo ra sự quan tâm mới.

2. Cung cấp đánh giá chi tiết về các thành phần của ClickStream.[40 điểm]

Sau đây là các thành phần và các ràng buộc của dữ liệu clickstream
Phân tích.

1. Kích thước.
Một thuộc tính mô tả hoặc đặc tính của dữ liệu.Trình duyệt, trang đích và

Chiến dịch là tất cả các ví dụ về kích thước mặc định trong phân tích.Một chiều là một
Thuộc tính mô tả hoặc đặc tính của một đối tượng có thể được đưa ra các giá trị khác nhau.
Vì
ví dụ,
Một
Địa lý
vị trí
có thể
có
Kích thước
được gọi là vĩ độ, kinh độ, hoặc tên thành phố.Giá trị cho kích thước tên thành phố
Có thể là San Francisco, Berlin hoặc Singapore.Trình duyệt, trang thoát, màn hình,
và thời lượng phiên là tất cả các ví dụ về kích thước xuất hiện theo mặc định trong
Phân tích.Kích thước xuất hiện trong tất cả các báo cáo của bạn, mặc dù bạn có thể thấy khác nhau
những người tùy thuộc vào báo cáo cụ thể.Sử dụng chúng để giúp tổ chức, phân khúc và
Phân tích dữ liệu của bạn.

2. Số liệu.
Các ví dụ về số liệu dữ liệu clickstream bao gồm (1) số lượt xem trang, (2)
Mẫu trang web được truy cập, bao gồm trang thoát thường xuyên nhất và trang web trước, (3)
thời gian lưu trú trên trang web, (4) ngày và thời gian truy cập, (5) số lượng

Đăng ký điền trên 100 khách truy cập, (6) số lượng đăng ký bị bỏ hoang, (7)
Nhân khẩu học của khách truy cập đã đăng ký, (8) số lượng khách hàng với xe mua sắm,
và (9) số lượng xe mua sắm bị bỏ hoang.

3. Tỷ lệ thoát.
Đây là tỷ lệ khách hàng vào trang web và rời đi ngay lập tức mà không cần
làm bất cứ điều gì.Đó là một dấu hiệu quan trọng cho thấy tất cả không tốt với trang web.
Tại sao khách hàng lại đến trang và chỉ rời đi mà không tham gia.

4. PageView.
PageView minh họa trang phổ biến nhất mà người dùng xem và điều này sẽ kích hoạt
Các chủ sở hữu để tối ưu hóa các trang cụ thể và các câu hỏi như;những gì làm như vậy
trang phổ biến?Thông tin như vậy có thể được sử dụng để cải thiện các trang khác?Vì vậy, cái khác
Các trang cũng có thể trở nên phổ biến.Làm thế nào những trang phổ biến này có thể được thương mại hóa
để tạo thu nhập.

5. Tỷ lệ chuyển đổi
Một lợi ích quan trọng khác của việc phân tích dữ liệu clickstream là khả năng tối ưu hóa
tỷ lệ chuyển đổi.Bằng cách theo dõi hành vi của người dùng thông qua kênh chuyển đổi,
Các doanh nghiệp có thể xác định các điểm đau và tối ưu hóa hành trình của người dùng để lái xe nhiều hơn
chuyển đổi.Điều này có thể giúp tăng doanh thu và tối đa hóa lợi tức đầu tư
cho những nỗ lực tiếp thị kỹ thuật số.

6. Mục tiêu
Mục tiêu và KPI là đồng nghĩa cả hai đều được sử dụng để đánh giá hiệu suất của

trang web hoặc trang.Trong khi KPI là một số liệu đo lường, mục tiêu là số liệu ngưỡng
giá trị.Nếu KPI đang tải xuống một bộ phim.Mục tiêu 200 người tải phim
mỗi ngày.Nếu KPI là người đăng ký trên trang web của bạn.Mục tiêu 100 người
Đăng ký trên trang web mỗi ngày.

7. KPI
KPI là viết tắt của các chỉ số hiệu suất chính.Đây là một thuật ngữ được sử dụng để đánh giá
Hiệu suất của bất kỳ số liệu.Số liệu nào nên được sử dụng để đo lường hiệu suất
của các trang web?KPI là số liệu được thiết lập bởi các chủ sở hữu của doanh nghiệp.Ví dụ, nó có thể
là một giá trị cụ thể của số liệu, ví dụ như tải xuống một bộ phim trong một ngày hoặc những người đăng ký
trên trang web của bạn.Nó cũng có thể là một tập hợp các số liệu phải được theo dõi
Biết làm thế nào trang web đang hoạt động.KPI là mục tiêu cụ thể của phép đo
cho thấy trang web đang hoạt động hay không.

3. Đánh giá phê bình các điều khoản với một ví dụ thích hợp về giả thuyết null với một
Ví dụ thích hợp.[12 dấu hiệu]

Định nghĩa của giả thuyết null:
Giả thuyết khống là một loại giả thuyết giải thích dân số
Tham số có mục đích là kiểm tra tính hợp lệ của dữ liệu thử nghiệm đã cho.Cái này
Giả thuyết bị từ chối hoặc không bị từ chối dựa trên khả năng tồn tại của
cho dân số hoặc mẫu.Nói cách khác, giả thuyết null là một giả thuyết trong đó
Các quan sát mẫu là kết quả từ cơ hội.Nó được cho là một tuyên bố trong đó
Các nhà khảo sát muốn kiểm tra dữ liệu.Nó được ký hiệu là H0.
Trong các thống kê, giả thuyết null thường được ký hiệu bằng chữ H với chỉ số
‘0, (không), sao cho H0.Nó được phát âm là H-null hoặc H-Zero hoặc H.
Khi thí nghiệm được thực hiện để kiểm tra giả thuyết.Dựa trên kết quả
Kết quả có thể của giả thuyết null là một trong hai.
Từ chối giả thuyết null
Không từ chối giả thuyết null
Chúng tôi không sử dụng từ chấp nhận trong đầu ra của giả thuyết.Chấp nhận nó có nghĩa là
Kết quả là chính xác cụ thể mà không phải là trường hợp.
Công thức giả thuyết null
Ở đây, các công thức kiểm tra giả thuyết được đưa ra dưới đây để tham khảo.
Công thức cho giả thuyết null là:
H0: P = P0
Giả thuyết khống nói rằng không có mối tương quan giữa sự kiện đo được
(biến phụ thuộc) và biến độc lập.Chúng tôi không phải tin rằng
Giả thuyết null là đúng để kiểm tra nó.Ngược lại, bạn có thể sẽ cho rằng
Có một kết nối giữa một tập hợp các biến (phụ thuộc và độc lập).

Các ví dụ giả thuyết null

Ex-1 nếu một loại thuốc làm giảm nguy cơ đột quỵ tim, thì giả thuyết null nên
Hãy là thuốc không làm giảm khả năng bị đột quỵ tim.Thử nghiệm này có thể là
được thực hiện bởi chính quyền của một loại thuốc cho một nhóm người nhất định trong một
đường.Nếu cuộc khảo sát cho thấy có một sự thay đổi đáng kể ở người dân, thì

Giả thuyết bị từ chối.

Ex-2 làm thanh thiếu niên đang sử dụng điện thoại di động nhiều hơn những người trưởng thành để truy cập Internet?
Trả lời: Tuổi không có giới hạn trong việc sử dụng điện thoại di động để truy cập internet.

EX-3 Có một quả táo hàng ngày sẽ không gây sốt?
Trả lời: Có táo hàng ngày không đảm bảo không bị sốt nhưng làm tăng khả năng miễn dịch
để chiến đấu chống lại những căn bệnh như vậy.
Ex-4 làm trẻ có tính toán toán học tốt hơn so với những người trưởng thành không?
Trả lời: Tuổi không có tác dụng đối với các kỹ năng toán học.
Trong nhiều ứng dụng phổ biến, việc lựa chọn giả thuyết null không được tự động hóa,
Nhưng thử nghiệm và tính toán có thể được tự động hóa.Ngoài ra, sự lựa chọn của null
Giả thuyết hoàn toàn dựa trên kinh nghiệm trước đây và lời khuyên không nhất quán.Các
sự lựa chọn có thể phức tạp hơn và dựa trên sự đa dạng của các ứng dụng và
Sự đa dạng của các mục tiêu.

4. Đánh giá nghiêm túc thuật ngữ với một ví dụ thích hợp về giả thuyết thay thế
với một ví dụ thích hợp.[12 dấu hiệu]

Định nghĩa của giả thuyết thay thế:
Giả thuyết thay thế định nghĩa có một mối quan hệ quan trọng về mặt thống kê
giữa hai biến.Trong khi đó, giả thuyết khống không có thống kê
mối quan hệ giữa hai biến.
Giả thuyết thay thế là một tuyên bố được sử dụng trong thí nghiệm suy luận thống kê.
Nó mâu thuẫn với giả thuyết khống và được biểu thị bằng HA hoặc H1.Chúng ta cũng có thể nói
rằng nó chỉ đơn giản là một sự thay thế cho null.Trong kiểm tra giả thuyết, một lý thuyết thay thế là
Một tuyên bố mà một nhà nghiên cứu đang thử nghiệm.
Tuyên bố này đúng theo quan điểm của nhà nghiên cứu và cuối cùng chứng minh
để từ chối null để thay thế nó bằng một giả định thay thế.Trong giả thuyết này,
sự khác biệt giữa hai hoặc nhiều biến được dự đoán bởi các nhà nghiên cứu, do đó
Mô hình dữ liệu được quan sát trong thử nghiệm không phải là do cơ hội.
Các loại giả thuyết thay thế
Về cơ bản, có ba loại giả thuyết thay thế, chúng là;
Đuôi trái: Ở đây, dự kiến tỷ lệ mẫu (π) nhỏ hơn một
giá trị được ký hiệu là π0, như vậy;
H1: $\pi < \pi 0$
Có đuôi phải: nó biểu thị rằng tỷ lệ mẫu (π) lớn hơn một số giá trị,
biểu thị bằng π0.
H1: $\pi > \pi 0$

Hai đuôi: Theo giả thuyết này, tỷ lệ mẫu (được biểu thị bằng π) là
Không bằng một giá trị cụ thể được biểu thị bằng π0.
H1: $\pi \neq \pi 0$
Lưu ý: Giả thuyết null cho cả ba giả thuyết thay thế, sẽ là H1: $\pi = \pi 0$.
Ví dụ về giả thuyết thay thế:
Ex-1 Để kiểm tra chất lượng nước của một dòng sông trong một năm, các nhà nghiên cứu đang thực hiện

quan sát.Theo giả thuyết null, không có thay đổi về chất lượng nước trong nửa đầu
của năm so với nửa thứ hai.Nhưng trong giả thuyết thay thế, chất lượng của
Nước kém trong hiệp hai khi được quan sát.
Ex-2: Rohan sẽ giành được ít hơn 100000 rupee trong trận hòa may mắn.
Ex-3: Nhân viên công ty tin rằng máy trộn bánh mì làm bánh mì có trọng lượng là 10g.
Nhưng khách hàng nghĩ rằng trọng lượng của bánh mì dưới 10g, đại diện cho điều này là null
và giả thuyết thay thế
HO: Tiết = 10g Đây là giả thuyết null.
HA: Tiết ≠ 10g Đây là giả thuyết thay thế.(Đây là một trong những điều nên
được kiểm tra) về mặt toán học đối diện với người khác.
Ex-4: Chiến dịch quảng cáo có thể hoặc không hiệu quả trong việc tăng khách truy cập
trang mạng.
HO: ở dạng null (không có sự khác biệt):
Không có sự khác biệt đáng kể giữa số lượng khách truy cập vào trang web trước đây
và sau khi quảng cáo vận động.
HA: ở dạng thay thế (khác biệt)
Có sự khác biệt đáng kể giữa số lượng khách truy cập vào trang web trước và
Sau khi quảng cáo vận động.
Cả HO và HA ở trên đều đúng.Hãy giả sử rằng khách truy cập trung bình vào trang web
Hàng tuần là 200 trước khi quảng cáo.
HO và HA cũng có thể được viết là
HO: Pha = 200
HA: Tiết ≠ hoặc> 200
Do đó, sẽ rất hữu ích khi biết những gì được sử dụng để trở thành số khách truy cập trước đây.

5. Với sơ đồ phù hợp, hãy đưa ra đánh giá chi tiết về mối tương quan thuật ngữ và
loại của họ.[10 điểm]
Định nghĩa về mối tương quan:
Phân tích tương quan giải thích mối quan hệ giữa các biến.Nó được quan tâm chủ yếu
với mối tương quan và hiệp phương sai.
Hiệp phương sai và tương quan thực tế đo lường cùng một mối quan hệ giữa hai
Các biến ngẫu nhiên, nhưng với cách tiếp cận khác nhau.
Trong khi hiệp phương sai là thước đo mối quan hệ giữa hai biến ngẫu nhiên
và từ các giá trị vô hạn -∞ đến +, đo tương quan từ mức1 đến + 1. Do đó,
Giá trị tương quan được tiêu chuẩn hóa, do đó có ý nghĩa hơn.
Tương quan đề cập đến một quá trình thiết lập mối quan hệ giữa hai
biến.
Tương quan được sử dụng để mô tả cách các bộ dữ liệu có liên quan với nhau.Tương quan
có thể được nhìn thấy khi hai bộ dữ liệu được biểu thị trên một biểu đồ phân tán, đó là một biểu đồ với
Một trục x và y và các dấu chấm đại diện cho các điểm dữ liệu.

Các loại tương quan
Biểu đồ phân tán giải thích mối tương quan giữa hai thuộc tính hoặc biến.Nó
đại diện cho cách gần hai biến được kết nối.Có thể có ba tình huống như vậy
Để xem mối quan hệ giữa hai biến
• Tương quan dương - Khi các giá trị của hai biến di chuyển theo cùng một hướng
do đó, việc tăng/giảm giá trị của một biến được theo sau bởi một

tăng/giảm giá trị của biến khác.

• Tương quan âm - Khi các giá trị của hai biến di chuyển ở phía ngược lại
hướng để tăng/giảm giá trị của một biến được theo sau bởi
Giảm/tăng giá trị của biến khác.
• Không có mối tương quan - khi không có sự phụ thuộc tuyến tính hoặc không có mối quan hệ giữa hai biến.

Khi nhiệt độ trong ngày cao hơn kem được bán.Điều này là tích cực
mối tương quan vì cả hai đang di chuyển theo cùng một hướng (nhiệt độ cao, nhiều băng nhiều hơn
kem).Hai biến cũng có thể tương quan tiêu cực.Cả hai di chuyển theo hướng ngược lại.
Một ví dụ về mối tương quan tiêu cực là số tiền bạn chi tiêu để sưởi ấm so với
nhiệt độ trong ngày.Đó là lượng nhiệt tăng cao hơn khi nhiệt độ thấp hơn.
Tương quan được đo giữa +1 đến mức1.Biểu đồ điểm phân tán như và 10 được sử dụng để
Hiển thị mối tương quan giữa hai biến.
Tích cực và tiêu cực không phải là cách duy nhất để mô tả mối tương quan;Tương quan có thể
Cũng được mô tả bởi sức mạnh của nó.Bộ dữ liệu cũng có thể có mối tương quan hoàn hảo, mạnh mẽ
mối tương quan, hoặc tương quan yếu.Các điểm dữ liệu càng gần nhau và chúng càng nhiều
tạo thành một đường thẳng, mối tương quan càng mạnh.Nếu các điểm dữ liệu tạo thành một thẳng hoàn hảo
dòng, các bộ dữ liệu được cho là có mối tương quan tích cực hoặc tiêu cực hoàn hảo tùy thuộc vào
Theo cách nào dòng đang đi (lên và phải = tích cực, xuống và phải = âm).
Tính toán tương quan:
Công thức hệ số tương quan Pearson
Công thức phổ biến nhất là hệ số tương quan Pearson được sử dụng cho tuyến tính
Sự phụ thuộc giữa các bộ dữ liệu.Giá trị của hệ số nằm giữa -1 đến +1.Khi
Hệ số đi xuống bằng không, sau đó dữ liệu được coi là không liên quan.Trong khi, nếu chúng ta
Nhận giá trị của +1, sau đó dữ liệu có mối tương quan tích cực và -1 có mối tương quan âm.

Tầm quan trọng của mối tương quan:

Tương quan là một trong những kỹ thuật cho các lựa chọn thuộc tính trong học máy.Hầu hết
Bộ dữ liệu được tạo thành từ nhiều biến, các biến này có mối tương quan với nhau và
các nhãn mục tiêu.

6. Cung cấp một lời giải thích chi tiết và biện minh cho các kỹ thuật để thay thế một
Thiếu giá trị trong một bộ dữ liệu.[10 điểm]

Hiểu dữ liệu là rất quan trọng vì những lý do sau
Cung cấp cái nhìn sâu sắc về loại lỗi trong bộ dữ liệu.
Biết các loại lỗi sẽ giúp quyết định cách sửa chúng.
Cung cấp cái nhìn sâu sắc về phân tích ban đầu và sâu hơn để sử dụng trên bộ dữ liệu.

Hồ sơ dữ liệu
Trong việc ước tính chất lượng dữ liệu, là rất nhiều quy trình gọi là hồ sơ dữ liệu.Hồ sơ là
Quá trình đánh giá các nguồn dữ liệu, kiểm tra các loại dữ liệu, nội dung để đánh giá

chất lượng, xác định các lỗi tiềm ẩn, khác để đề xuất các kỹ thuật cho chất lượng
sự cải tiến.
Yếu tố ảnh hưởng đến chất lượng dữ liệu:
Kích thước chất lượng dữ liệu.
Lỗi dữ liệu.

Sửa lỗi và làm sạch dữ liệu

Tất cả dữ liệu được thu thập thường là "bẩn" hoặc có "lỗi":
"Dữ liệu bẩn" là một thuật ngữ được sử dụng trong việc mô tả các trạng thái khác nhau của độ thô của dữ liệu
có thể tác động đến khả năng trích xuất thông tin từ tập dữ liệu.Dữ liệu bẩn phải
sạch sẽ bằng quá trình phát hiện, sửa hoặc xóa nhiễu hoặc lỗi trong dữ liệu
bộ.
Để đặt nó vào quan điểm, điều gì làm cho dữ liệu bị bẩn?Dữ liệu bẩn được coi là có
Các vấn đề sau đây được liệt kê dưới đây trong số nhiều người khác.

Dữ liệu không đầy đủ hoặc thiếu
Nếu bất kỳ vị trí nào là một mục dữ liệu nên được để trống, không có gì được viết hoặc
Một số giá trị lẻ hoặc ký tự ở nơi chỉ ra một giá trị bị thiếu.

Dữ liệu trùng lặp: Hàng lặp lại nhầm lẫn trong một bảng nhiều lần.

Kiểu dữ liệu không chính xác: Đây là khi các loại dữ liệu sai được sử dụng.Ví dụ: nếu tuổi
của một người là 36 năm, một lỗi đã được thực hiện bằng cách nhập bảng chữ cái "wy" thay cho
36 do lỗi đánh máy.

Dữ liệu/ nhiễu/ ngoại lệ không chính xác: Đầu vào mục dữ liệu không chính xác.Ví dụ, nếu
Giá trị đúng cho tuổi là 36 tuổi, nhưng 360 đã được viết.

Sửa dữ liệu không đầy đủ/ thiếu

• Làm thế nào các giá trị bị thiếu được thể hiện trong các bộ dữ liệu?
• Không gian cho các giá trị bị thiếu để trống hoặc các ký tự khác được sử dụng để
thay thế nó?

Đây là lý do tại sao giai đoạn đầu tiên (sự hiểu biết về kinh doanh và dữ liệu là rất cần thiết)
Nếu biến là một cột của tuổi hoặc BMI và được biểu diễn bằng 0, đó là một chỉ dẫn
có giá trị bị thiếu vì không có người tuổi hoặc chỉ số khối cơ thể là 0. Thiếu giá trị là
thay thế bằng cách sử dụng;trung bình, chế độ trung bình hoặc giá trị mặc định hoặc thậm chí sử dụng một số không phổ
Kỹ thuật tùy thuộc vào bối cảnh và biện minh.Dưới đây là một số phương pháp của
thay thế các giá trị bị thiếu.

Trung bình hoặc trung bình là các kỹ thuật phổ biến nhất để thay thế các giá trị bị thiếu.Nhưng
Nó có thể không phải là kỹ thuật đúng đắn luôn luôn.

Trước khi quyết định thay thế giá trị bị thiếu theo giá trị trung bình, nên độ lệch
đã kiểm tra.Nếu độ lệch là> hoặc <so với 1, thì phân phối bị sai lệch ở bên phải

hoặc trái.Do đó, chính xác hơn là thay thế giá trị bị thiếu bằng trung bình thay vì
nghĩa.

7. hồi quy của người ”là gì về mặt phân tích dữ liệu cung cấp một ví dụ về tình huống
nơi mà hồi quy có thể được sử dụng.[10 điểm]

Phân tích hồi quy:
Phân tích hồi quy là một tập hợp các phương pháp thống kê được sử dụng để ước tính
Mối quan hệ giữa một biến phụ thuộc và một hoặc nhiều biến độc lập.Nó
có thể được sử dụng để đánh giá sức mạnh của mối quan hệ giữa các biến và cho
Mô hình hóa mối quan hệ trong tương lai giữa họ.
Nó cung cấp các giá trị của biến phụ thuộc từ giá trị của một độc lập
Biến đổi.Việc sử dụng chính phân tích hồi quy là xác định sức mạnh của
dự đoán, dự báo một hiệu ứng, một xu hướng, v.v.
Dự đoán là giá trị liên tục, nó có phạm vi từ tối thiểu đến tối đa, nó
Không thể là nhóm thành các loại, do đó được gọi là hồi quy học tập có giám sát.
Hồi quy được sử dụng để dự đoán đầu ra (y) được gọi là các biến phụ thuộc bằng cách sử dụng đầu vào (x)
được gọi là các biến độc lập.
Ví dụ, chúng tôi muốn dự đoán doanh số của kem dựa trên nhiệt độ của
ngày.

Bán hàng (y) là đầu ra mà lần lượt được gọi là các biến phụ thuộc.
Nhiệt độ (x) là đầu vào, lần lượt được gọi là biến độc lập.
Bán hàng (Y) vé dựa trên quảng cáo truyền hình.
Bán hàng là đầu ra mà lần lượt được gọi là biến phụ thuộc.
Quảng cáo TV (X) là đầu vào, lần lượt được gọi là biến độc lập.
Hồi quy tuyến tính:
Thước đo mức độ của mối quan hệ giữa hai biến được thể hiện bởi
hệ số tương quan.Phạm vi của hệ số này nằm giữa -1 đến +1.Cái này
Hệ số cho thấy sức mạnh của sự liên kết của dữ liệu được quan sát cho hai biến.
Một phương trình đường hồi quy tuyến tính được viết dưới dạng:
$Y = mx + b$
trong đó x là biến độc lập và được vẽ dọc theo trục x
Y là biến phụ thuộc và được vẽ dọc theo trục y
Độ dốc của đường là M và B là chặn (giá trị của y khi $x = 0$).
Hồi quy tuyến tính đơn biến
Ví dụ được gọi là hồi quy tuyến tính đơn biến vì một biến (đơn vị)
đó là quảng cáo TV (X) được sử dụng để dự đoán số lượng doanh số (Y).

X -> Các biến độc lập giải thích được sử dụng để dự đoán hoặc liên kết với y
Trục
B -> y chặn
M -> Độ dốc của các biến giải thích
E -> Thuật ngữ lỗi còn lại hồi quy

Hồi quy tuyến tính đa biến

Nó được áp dụng khi có nhiều hơn một đầu vào (x1, x2, trên .x3) được sử dụng để dự đoán
Đầu ra (y).Ví dụ;
Dự đoán việc bán (y) dựa trên ba đầu vào;
Quảng cáo truyền hình (x1),
Quảng cáo Radio (x2)
Quảng cáo báo chí (x3)

Để đánh giá hiệu suất của mô hình hồi quy là hoàn toàn khác với
phân loại.Trong hồi quy, tất cả các đánh giá đang cố gắng biện minh cho việc âm mưu tốt như thế nào
Điểm gần hơn với các điểm thực tế.

8. Phân loại của ”" về mặt phân tích dữ liệu cung cấp một ví dụ về
Tình huống mà phân loại có thể được sử dụng.[10 điểm]

Phân loại:
Phương pháp sắp xếp dữ liệu thành các lớp đồng nhất theo
Các tính năng có trong dữ liệu được gọi là phân loại.
Một hệ thống phân tích dữ liệu theo kế hoạch giúp dữ liệu cơ bản dễ dàng tìm thấy và phục hồi.
Điều này có thể được quan tâm đặc biệt cho khám phá pháp lý, quản lý rủi ro và tuân thủ.
Các phương thức và bộ hướng dẫn bằng văn bản để phân loại dữ liệu nên xác định những gì
các cấp độ và các biện pháp mà công ty sẽ sử dụng để tổ chức dữ liệu và xác định vai trò của
Nhân viên trong doanh nghiệp liên quan đến quản lý đầu vào.Khi một bản thu thập dữ liệu
Đề án đã được thiết kế, các tiêu chuẩn bảo mật quy định cách tiếp cận thích hợp
Thực tiễn cho mỗi bộ phận và các tiêu chí lưu trữ xác định vòng đời của dữ liệu
Nhu cầu nên được thảo luận.Phân loại này thuộc nhóm máy
Học tập (ML) là những gì được dự đoán được biết đến, ví dụ
• Dự đoán kết quả của cuộc bầu cử (thắng hoặc thua) (phân loại)
• Dự đoán nếu bằng sáng chế bị ung thư hay không.(Phân loại)
• Dự đoán nếu bệnh nhân mắc bệnh tiểu đường hay không.(Phân loại)
• Dự đoán nếu một khách hàng sẽ khuấy động hoặc ở lại.(Phân loại)

Phân loại nhị phân:
Nó cũng là phân loại vì dự đoán đưa họ vào các nhóm.Khi nó là hai
Các nhóm được gọi là phân loại nhị phân.Đây là một loại học máy (ML) trong đó
Những gì đang được dự đoán (nhãn lớp đầu ra hoặc mục tiêu) được biết đến và một danh mục hoặc
nhóm.Ví dụ, hãy xem xét một mặt cắt ngang của bộ dữ liệu ung thư vú Wisconsin.
Nó có chín thuộc tính và một lớp đầu ra.
• Thuật toán sẽ dạy hệ thống máy tính học cách dự đoán, ai sẽ có
Ung thư (dự đoán 1) hoặc không (dự đoán 0).
• Điều nhấn mạnh ở đây, là chúng ta biết những gì đang được dự đoán;dự đoán 1 hoặc 0
Phân loại đa lớp:
Nhiều hơn hai nhóm được gọi là phân loại đa lớp.Nó có nhiều loại
các ứng dụng.Nó có thể được sử dụng để xác định động vật từ hình ảnh và sắp xếp chúng thành
Thể loại.Các công ty an ninh mạng có thể sử dụng phân loại đaass để sắp xếp đến
Email là thư rác hay không.Nó cũng có thể được sử dụng khi phân tích tâm trạng của một cá nhân
nhiều hơn tích cực hoặc tiêu cực.Thay vào đó, nó sẽ sử dụng nhiều loại như, hạnh phúc,

Đáng buồn, chán nản, phấn khích, v.v ... Không có giới hạn về số lượng các lớp được sử dụng trong nhiều lớp phân loại.
Ví dụ: Một sự tương tự là xem xét một nghiên cứu về dân số bao gồm 1000
bệnh nhân và nếu 900 bệnh nhân trong số 1000 không có bệnh, một mô hình dự đoán tất cả 1000
vì bệnh vẫn có vẻ chính xác 90%, ngay cả khi còn lại
100 bệnh nhân mắc bệnh, và họ không được xác định.Do đó, độ chính xác có
thất bại hay đúng hơn là không đủ để ước tính hiệu suất của các mô hình phân loại do
Để mất cân bằng bản chất của bộ dữ liệu.
Nghiên cứu về các kỹ thuật xử lý bộ dữ liệu mất cân bằng hoặc giảm
Ảnh hưởng của sự mất cân bằng là một lĩnh vực nghiên cứu tích cực.
Ma trận hỗn loạn:
Nếu độ chính xác đã thất bại, thì chúng ta có thể đo lường hiệu suất của phân loại
Mô hình hóa thông qua một ma trận gọi là ma trận nhầm lẫn.
Một ma trận nhầm lẫn, còn được gọi là ma trận lỗi là một bảng được sử dụng để hiển thị
hiệu suất phân loại.

• Tích cực thực sự (TP): Thuật toán dự đoán tích cực và câu trả lời đúng
tích cực;(dự đoán chính xác);
• Tiêu cực thực sự (TN): Thuật toán dự đoán tiêu cực và câu trả lời đúng
là tiêu cực (dự đoán chính xác);
• Hậu thế giả (FP): Thuật toán dự đoán tích cực, nhưng câu trả lời đúng là
tiêu cực (dự đoán không chính xác);Và
• Tiêu cực sai (FN): Thuật toán dự đoán tiêu cực, nhưng câu trả lời đúng
là tích cực (dự đoán không chính xác).

9. Cung cấp một lời giải thích chi tiết về phương sai thuật ngữ [10 điểm].
Phương sai:
• Phương sai là một phép đo của sự lây lan giữa các số trong một tập dữ liệu.
• Nó đo lường mức độ phân tán dữ liệu xung quanh giá trị trung bình của mẫu.
• Các nhà đầu tư sử dụng phương sai để xem một rủi ro đầu tư mang lại bao nhiêu và liệu nó có thể
có lợi nhuận.
• Phương sai cũng được sử dụng trong tài chính để so sánh hiệu suất tương đối của từng tài sản trong
Một danh mục đầu tư để đạt được phân bổ tài sản tốt nhất.
• Căn bậc hai của phương sai là độ lệch chuẩn.

Trong thống kê, phương sai đo lường sự thay đổi từ trung bình hoặc trung bình.Nó được tính toán bởi
Lấy sự khác biệt giữa mỗi số trong tập dữ liệu và giá trị trung bình, sau đó bình phương
sự khác biệt để làm cho chúng tích cực, và cuối cùng chia tổng của các hình vuông cho
Số lượng giá trị trong tập dữ liệu.
Phương sai được tính bằng cách sử dụng công thức sau:

Ví dụ:
• Tìm giá trị trung bình, độ lệch chuẩn và phương sai cho các dữ liệu sau: 6, 7,10, 12, 13,
4, 8, 12.
Giải pháp:
Dữ liệu đã cho: 6, 7,10, 12, 13, 4, 8, 12

Tìm kiếm ý nghĩa:
Chúng tôi biết rằng trung bình là tỷ lệ của tổng quan sát so với tổng số
quan sát.
Trung bình = tổng quan sát / tổng số quan sát.
Trung bình = (6+7+10+12+13+4+8+12)/8
Trung bình = 72/8
Có nghĩa là = 9.
Tính toán phương sai

Do đó, phương sai = 74/8
Phương sai = 9,25
Tìm độ lệch chuẩn:
Chúng ta biết rằng phương sai là bình phương của độ lệch chuẩn.Do đó, độ lệch chuẩn
có thể được tìm thấy bằng cách lấy căn bậc hai của phương sai.
• Do đó, độ lệch chuẩn = √variance
• Độ lệch chuẩn = (9,25) = 3.041.
• Do đó, giá trị trung bình, phương sai và độ lệch chuẩn của dữ liệu đã cho là 9, 9,25, 3.041
tương ứng.

10.
Giải thích tích cực sai lệch và phản ánh về giá trị trung bình, trung bình và chế độ.[10 điểm].
Và
Giải thích sai lệch và phản ánh về giá trị trung bình, trung bình và chế độ.[10 điểm].
Skewness:
Độ lệch là mức độ bất đối xứng quan sát được trong phân phối xác suất.
Khi các điểm dữ liệu trên đường cong chuông không được phân phối đối xứng sang trái và phải
Các mặt của trung vị, đường cong chuông bị lệch.Phân phối có thể là tích cực và đúng-
sai lệch, hoặc tiêu cực và sai trái.Một phân phối bình thường thể hiện độ lệch không.
Các loại sai lệch
Tiêu cực, hoặc bị ghép trái đề cập đến một đuôi dài hơn hoặc béo hơn ở phía bên trái của
Phân phối, trong khi tích cực, hoặc xu hướng phải, đề cập đến một đuôi dài hơn hoặc béo hơn ở bên phải.
Hai độ lệch này cho thấy hướng hoặc trọng lượng của phân phối.
Ba phân phối xác suất dưới đây được giảm giá phải ở một mức độ ngày càng tăng.
Giá trị trung bình của dữ liệu sai lệch tích cực sẽ lớn hơn so với trung bình.Trong một cái ghép trái
Phân phối, giá trị trung bình của dữ liệu sai lệch tiêu cực sẽ nhỏ hơn so với trung bình.
Phân phối tích cực của một phần mềm hoặc (+ve) có nghĩa là đuôi của nó rõ rệt hơn
mặt phải hơn bên trái.Vì phân phối là tích cực, giả định là
Giá trị của nó là tích cực.Như vậy, hầu hết các giá trị kết thúc bên trái của giá trị trung bình.Điều này có nghĩa là
rằng các giá trị cực đoan nhất là ở phía bên phải.

Tiêu cực hoặc (-ve) Tai lệch trái có nghĩa là đuôi rõ rệt hơn ở bên trái thay vì
hơn quyền.Hầu hết các giá trị được tìm thấy ở phía bên phải của giá trị trung bình trong âm tính
độ lệch.Như vậy, các giá trị cực đoan nhất được tìm thấy ở bên trái.

Ý nghĩa của độ lệch:

Đây là mức độ mà một bộ dữ liệu bị sai lệch, trong ví dụ trên giả sử cho một số
Lý do chiều cao hoặc huyết áp của nam giới trong thành phố rất cao hoặc cao hơn nhiều
hơn bình thường vì một số di truyền bất thường hoặc một số khiếm khuyết, sau đó các phân phối
sẽ không có hình dạng đối xứng hoặc chuông.Đúng hơn là nó sẽ nghiêng về phía bên trái bị lệch hoặc
sang phải hoặc +vely sai lệch.

Độ lệch thường là do sự hiện diện của sự thiên vị trong bộ dữ liệu.Nếu đó là điểm kiểm tra,
Dữ liệu bình thường sẽ hiển thị một số ít có điểm rất cao và một số ít có điểm rất thấp
trong khi đa số nên có giá trị trung bình trong khoảng 68, 95 và 99 % (1,2 và
3) Từ độ lệch chuẩn.
Nhưng khi phân phối bị sai lệch hoặc phải hoặc phải.Sau đó, học sinh vượt qua rất tốt hoặc
thất bại quá nhiều.
Nếu những người trong thành phố có áp suất cao hoặc thấp so với bình thường, thì phải có
Có gì đó không đúng.

Nếu tỷ lệ tử vong ở trên phân phối bình thường, thì một số điều là sai.
Skewness là một trong những thử nghiệm cho tính bình thường trong các tình huống thực tế.
Độ lệch là đo từ -1 đến + 1
Dữ liệu phân tán bình thường có độ lệch = 0
Dữ liệu sai lệch là
-1> x <1