Xin chào cô và các bạn, em/mình là vũ văn huy để tiếp nối phần giới thiệu về các thuật toán dùng để lọc thư rác thì mình sẽ trình bầy ứng dụng của 1 phương pháp cụ thể để xử lý tin rác. Cụ thể là sử dụng phương pháp học máy. Và thuật toán mình sử dụng là Navie Bayes.

Thì thuật toán này dựa trên công thức Bayes rất nổi tiếng về việc tính xác xuất có điều kiện. P(A/B) = P(AB)/P(B). Thuật toán sẽ tính toán trên bộ dữ liệu tần suất các từ và xác xuất xuất hiện của từ đó trong bộ dữ liệu.

Khi một email mới đến, nó sẽ kiểm tra xem từ nào xuất hiện nhiều trong tin nhắn spam của tập spam, từ đó quyết định xem tin nhắn đó có phải là spam hay không.

Sau phần giới thiệu của Tuấn Long, mọi người đã có cái nhìn cơ bản về cách Naive Bayes hoạt động. Tuy nhiên, đó chỉ là trong bộ dữ liệu nhỏ. Đối với dữ liệu lớn hơn, chúng tôi sẽ sử dụng máy tính để tính toán và đưa ra kết quả.

Trước tiên thì mình sẽ giới thiệu sợ qua về bộ dữ liệu mà bọn mình sử dụng. Bộ dữ liệu gồm 5571 dòng, 2 cột, tương ứng là nhãn ham ứng với tin nhắn thông thường, và spam đối với tin nhắn spam. Cột còn lại “SMS” là nội dung, phần body của 1 đoạn email. Như mọi người thấy thì thuật toán của bọn mình ứng dụng trong bài toán supervilearning (học có giám sát) với các dữ liệu được gán nhãn.

Để hiểu hơn về dữ liệu thì bọn mình đã thực hiện 1 số phân tích và thu được kết quả như sau:

+ Đầu tiên mình sẽ tính % tin nhắn thông thường, tin spam, thì được tỷ lệ là:

ham 0.865937

spam 0.134063

Tiếp theo để trực quan hóa tần suất suất hiện của từ trong tin nhắn mình sử dụng 1 thư viện wordcloud của python để hiện thị như dưới đây. Các từ trong tin thông thường sẽ là I u, go to, ….Còn trong thư spam các từ phổ biến sẽ là: Free(chúc mừng bạn đã nhận được 1 suất học bổng free chẳng hạn), call , now, ….

Khi 1 tin nhắn mới suất hiện nếu trong tin đó từ Free mà có nhiều thì tin đó là spam

Tiếp theo mình sẽ chia tập ban đầu thành tập train và tập test theo cách thông thường với tỷ lệ 80 đối với tập train và 20% đối với tập test. Trong cả tập train và test đều giữ đúng các tỷ lệ tin thông thường và tin spam là 0.86 và 0.14.

Tiếp theo mình sẽ làm sạch dữ liệu bằng cách thay thế tất cả ký tự đặc biệt không là chữ hoặc số thành ký tự cách(space) và chuyển tất cả về lower – chữ viết thường.

Rồi cắt từng từ trong 1 câu ra: và lọc lấy cách từ khác nhau thì thu được 7783 từ khác nhau và tần suất xuất hiện các từ như bảng dưới đây.

Sau đó, chúng tôi ghép các từ duy nhất lại để tạo thành câu mới. Thì có 1 bảng như này.

Rồi mình viết code thuật toán navie bayes từng bước như bạn tuấn long vừa trình bày:

Và viết hàm dự đoán bằng cách so sánh xác xuất cuối cùng xác xuất tin đó là tin spam và xác xuất tin đó là tin thông thường rồi so sánh chúng.

Mình sẽ test với 1 câu bất kỳ:

Qua các lượt test trên thì mô hình hoàn toàn dự đoán đúng.

Khi thử lại trên tập test thì cho độ chính xác lên đến 99%. Nghĩa là cứ 100 tin nhắn nhận đến thì nó sẽ dự đoán sai 1 lần. và đúng 99 lần.

Lưu ý trên chỉ là tests trên tập huấn luyện ban đầu và cho kết quả khá tốt.

Đối với trường hợp test trên bộ dữ liệu thực tế gồm 51 bản ghi cho tỷ lệ chính xác khá cao: 92% theo đánh giá và ghi nhận trực tiếp từ con người.(tốt hơn google)