Generalized Linear Model

*E-mail: pima.vn@gmail.com

Mô tả Dự án

Generalized Linear Model (GLM) là những mô hình thống kê nâng cao tổng quát hóa từ mô hình Hồi quy Tuyến tính (Linear Regression, LR). GLM được phát minh bởi John Nelder và Robert Wedderburn vào năm 1972. Mô hình này được dùng rất nhiều trong các bài toán hồi quy và phân lớp dữ liệu. Trong dự án này, ta sẽ tìm hiểu về cấu trúc của GLM và áp dụng vào bài toán thực tế thông qua việc chọn mô hình, ước lượng tham số trong mô hình, dự đoán kết quả và đánh giá.

Yêu cầu cơ bản

Lý thuyết

- Tìm hiểu mô hình LR dưới góc nhìn của ước lượng tham số thống kê. Mô hình LR có áp dụng được và có phù hợp cho các bài toán sau đây hay không? Giải thích lý do.
  - Dự đoán mong muốn học tiếp sau đại học (có hay không) của các sinh viên Việt Nam.
  - Dự đoán số ca dương tính với COVID–19 tại TPHCM.
  - Dự đoán tỉ lệ học sinh lớp 12 thích Toán của một trường.
- Trình bày các thành phần trong GLM. Vì sao LR là trường hợp đặc biệt của GLM?
- Trình bày thuật toán xác định các tham số trong GLM và so sánh với LR.
- Trình bày cách đánh giá mô hình GLM.
- Cho các kiểu dữ liệu nhị phân (binary), đếm (count), tỉ lệ (proportion). Chọn **một** trong các kiểu dữ liệu trên và trình bày GLM trong trường hợp cụ thể đó.

Thực hành

- Tìm hiểu cách huấn luyện, đánh giá và sử dụng các mô hình tuyến tính (Linear Model) trong thư viện `scikit-learn`.
- Tìm một bộ dữ liệu giống với kiểu dữ liệu đã chọn ở phần Lý thuyết và áp dụng GLM phù hợp. Đánh giá mô hình và rút ra nhận xét.

Yêu cầu nâng cao

Lý thuyết

- Chứng minh nếu sử dụng Cannonical link function thì thuật toán Newton-Raphson và Fisher's Scoring là như nhau. Trong trường hợp tổng quát, hãy so sánh hai thuật toán.
- Overdispersion là gì? Giải thích nguyên nhân và trình bày giải pháp.

Thực hành

- Cài đặt GLM nhưng không sử dụng các thư viện sẵn có.

Kiến thức

- Toán : Biến ngẫu nhiên, phân phối xác suất, ước lượng tham số, vi phân nhiều biến, tối ưu hóa, ma trận.
- Một số từ khóa : Generalized Linear Model, Linear Regression, Exponential Family, Cumulant, Variance Function, Maximum Likelihood Estimation, Newton-Raphson, Fisher's Scoring, Cannonical Link Function, Deviance, Overdispersion.

Tham Khảo

[1] Các bài giảng PiMA 2021.
[2] https://en.wikipedia.org/wiki/Generalized_linear_model
[3] https://scikit-learn.org/stable/modules/linear_model.html