Báo cáo Lần 4

1. Logistic Regression - Cơ sở toán học

1.1. Mục tiêu và mô tả cơ bản

Logistic regression là mô hình tuyến tính dùng để mô tả xác suất của biến nhị phân $y \in \{0, 1\}$ dưới dạng hàm sigmoid của tổ hợp tuyến tính của đặc trưng x. Mục tiêu là suy ra tham số w (và bias b) sao cho

$$P(y = 1 \mid x, w, b) = \sigma(w^\top x + b) \equiv \frac{1}{1 + e^{-(w^\top x + b)}}.$$

Mô hình logistic và các phương pháp ước lượng dành cho dữ liệu nhị phân và phân loại đã được hệ thống hoá trong văn liệu thống kê cổ điển; một bài báo nền tảng cho phát triển logistic trong thống kê là Cox (1958). ([SciSpace][1])

1.2. Hàm khả năng và tối ưu (MLE)

Với tập dữ liệu độc lập $D = \{(x_i, y_i)\}_{i=1}^N$, hàm khả năng và log-likelihood theo w là

$$p(D \mid w) = \prod_{i=1}^{N} \sigma(w^\top x_i)^{y_i} \, (1 - \sigma(w^\top x_i))^{1-y_i},$$

$$\ell(w) = \sum_{i=1}^{N} \left[y_i \ln \sigma(w^\top x_i) + (1 - y_i) \ln(1 - \sigma(w^\top x_i)) \right].$$

Tối đa hoá $\ell(w)$ theo w bằng các phương pháp số (ví dụ: gradient ascent, Newton-Raphson). Thực hành và các khuyến nghị kiểm tra (diagnostics, goodness-of-fit, lựa chọn biến) được trình bày chi tiết trong sách tham khảo tiêu chuẩn về Applied Logistic Regression. ([Wiley Online Library][2])

1.3. Hạn chế của cách tiếp cận MLE thuần túy

- MLE cho một nghiệm điểm (point estimate) $\hat{w}_{MLE}$ nên **không trực tiếp cung cấp độ không chắc chắn** của tham số (mặc dù có thể xấp xỉ bằng covariance Fisher).
- Khi mẫu nhỏ, hoặc có đa cộng tuyến, hoặc dữ liệu nhiễu, MLE dễ dẫn đến **overfitting** hoặc ước lượng không ổn định. Những vấn đề này là lý do chính để cân nhắc phương pháp Bayes/regularization. ([Wiley Online Library][2])

---

2. Bayesian Logistic Regression

2.1. Ý tưởng chính

Trong khuôn khổ Bayes, ta coi tham số w là biến ngẫu nhiên với phân phối tiên nghiệm $p(w)$. Sau khi quan sát dữ liệu D, phân phối **hậu nghiệm** theo Bayes là

$$p(w \mid D) = \frac{p(D \mid w)\, p(w)}{p(D)},$$

trong đó $p(D) = \int p(D \mid w) p(w)\, dw$ là bằng chứng (marginal likelihood). So với MLE, cách tiếp cận Bayes có hai lợi ích chính: (1) **kết hợp kiến thức trước (prior)** giúp điều hoà/regularize tham số; (2) **cung cấp phân phối posterior** cho phép lượng hóa bất định trong tham số và trong dự đoán. Các nguyên lý và ứng dụng Bayesian cho dữ liệu nhị phân được mô tả trong tài liệu tổng quan và các bài báo nền tảng. ([Columbia Statistical Consulting][3])

2.2. Lựa chọn prior và ý nghĩa regularization

- Thông thường người ta chọn prior Gaussian $p(w) = \mathcal{N}(w \mid 0, \lambda^{-1} I)$. Điều này tương đương với thêm một điều khoản phạt L2 trong tối ưu (ridge), và giúp tránh trọng số quá lớn.
- Khi prior được chọn hợp lý, posterior sẽ "kéo" nghiệm về phía 0 nếu dữ liệu không đủ mạnh để khẳng định giá trị lớn cho trọng số — đây là hiện tượng regularization theo quan điểm Bayes. Các luận giải này (Bayesian regularization, model comparison) được trình bày sâu trong MacKay (1992) và sách giáo khoa về Bayesian. ([MIT Press Direct][4])

2.3. Posterior prediction (marginal predictive)

Để dự đoán cho một điểm mới x^*, đúng về mặt Bayes là tích hợp theo posterior:

$$p(y^* = 1 \mid x^*, D) = \int \sigma(w^\top x^*)\, p(w \mid D)\, dw.$$

Khó khăn thực tế là tích phân này *thường không thể giải biểu thức đóng* vì likelihood của logistic không là hàm liên hợp với Gaussian prior — dẫn tới posterior không có dạng đơn giản. Điều này dẫn tới nhu cầu xấp xỉ (approximation) hoặc phương pháp Monte Carlo. ([apps.olin.wustl.edu][5])

---

3. Tại sao cần xấp xỉ phân phối hậu nghiệm ?

3.1. Hạn chế tính toán

Likelihood logistic $\prod_i \sigma(w^\top x_i)^{y_i}(1-\sigma(w^\top x_i))^{1-y_i}$ kết hợp với Gaussian prior không tạo thành phân phối hậu nghiệm có dạng chuẩn; do đó các tích phân marginal (cho tham số hoặc dự đoán) không có nghiệm đóng. Vì thế ta cần:

- Dùng **MCMC** (Gibbs/Metropolis/HMC) để rút mẫu từ posterior chính xác nhưng tốn chi phí tính toán, hoặc
- Dùng **xấp xỉ xấp xỉ (approximation)** như Laplace, Variational Inference (VI), hoặc Expectation Propagation (EP) để thu được biểu thức gần đúng nhanh hơn. Những so sánh giữa các phương pháp này và các trade-off đã được thảo luận rộng rãi trong chuyên khảo Bayesian và bài báo about CV/LOO. ([apps.olin.wustl.edu][5])

3.2. Khi nào dùng xấp xỉ thay vì MCMC

- Khi cần kết quả nhanh hoặc khi mô hình lớn/đa tham số (chi phí MCMC lớn).
- Khi posterior tương đối "khỏe" (gần chuẩn) — xấp xỉ Gaussian (Laplace) sẽ chính xác hơn khi kích thước mẫu lớn theo lý thuyết limit. Tuy nhiên với posterior đa cực hoặc nặng đuôi, Laplace có thể kém chính xác. Những điều kiện và phân tích lỗi cho Laplace được trình bày trong tài liệu toán học (Tierney & Kadane, các phân tích thống kê asymptotic). ([CMU School of Computer Science][6])

---

4. Laplace Approximation — toán học và áp dụng cho logistic

4.1. Mục tiêu và ý tưởng chung

Laplace approximation (hay normal approximation) xấp xỉ một phân phối $p(w \mid D)$ bằng một phân phối Gaussian quanh điểm cực đại (mode) của log-posterior. Kỹ thuật cơ bản:

1. Tìm **mode** $w_{MAP} = \arg\max_w \log p(w \mid D) = \arg\max_w \big(\log p(D \mid w) + \log p(w)\big)$.
2. Lấy khai triển Taylor bậc hai của $\log p(w \mid D)$ quanh w_{MAP}:

$$\log p(w \mid D) \approx \log p(w_{MAP} \mid D) - \tfrac{1}{2}(w - w_{MAP})^\top H (w - w_{MAP}),$$

   trong đó $H = -\nabla^2 \log p(w \mid D)\big|_{w_{MAP}}$ là ma trận Hessian âm (positive definite under conditions).
3. Exponentiating dẫn tới xấp xỉ Gaussian:

$$p(w \mid D) \approx \mathcal{N}(w \mid w_{MAP}, H^{-1}).$$

Nguồn gốc toán học của Laplace approximation, và việc dùng nó để xấp xỉ posterior moments/marginals, được phân tích kỹ trong Tierney & Kadane (1986) và các sách về Bayesian/ML. ([CMU School of Computer Science][6])

4.2. Ứng dụng cụ thể cho logistic

- **B1 — Tìm MAP**: tối ưu $\log p(D \mid w) + \log p(w)$. Với Gaussian prior $\mathcal{N}(0, \lambda^{-1} I)$, bài toán tương đương tối đa hoá log-likelihood với điều khoản phạt L2. Có thể giải bằng Newton-Raphson (vì log-likelihood logistic có Hessian).
- **B2 — Tính Hessian**: Hessian của $-\log p(w \mid D)$ là

$$H = X^\top W X + \lambda I,$$

  trong đó W là ma trận chéo với phần tử $W_{ii} = \sigma(w^\top x_i)(1 - \sigma(w^\top x_i))$. (Đây là biểu thức chuẩn từ tối ưu logistic với penalization.)
- **B3 — Xấp xỉ predictive**: Thay phân phối $p(w \mid D)$ bằng Gaussian $\mathcal{N}(w_{MAP}, H^{-1})$ trong tích phân predictive:

$$p(y^* = 1 \mid x^*, D) \approx \int \sigma(w^\top x^*)\, \mathcal{N}(w \mid w_{MAP}, H^{-1})\, dw.$$

  Tích phân trên vẫn không có dạng đóng hoàn toàn vì $\sigma(\cdot)$ không phải là hàm tuyến tính, nhưng ta có thể xấp xỉ tiếp (ví dụ: sử dụng công thức xấp xỉ tích phân của một sigmoid với Gaussian, hoặc dùng Gaussian integral approximations; Bishop (PRML) trình bày cách triển khai Laplace trong bài toán phân loại tuyến tính). Một công thức gần đúng thường dùng (delta method / probit approx) cho predictive probability. ([Microsoft][7])

4.3. Ưu/nhược điểm của Laplace

- **Ưu điểm**: tính toán nhanh hơn so với MCMC; dùng được cho mô hình có nhiều tham số; trong nhiều trường hợp (posterior gần chuẩn) cho kết quả tốt.
- **Nhược điểm**: nếu posterior đa cực hoặc phân phối có đuôi dày, Laplace (dịch xấp xỉ bằng một Gaussian đơn) có thể sai lệch; các mối tương quan phi tuyến mạnh giữa tham số có thể khiến xấp xỉ kém. Khi độ chính xác cần cao, hoặc khi posterior nhiều modal, người ta ưu tiên MCMC hoặc phương pháp tiên tiến như Expectation Propagation (EP) / Variational Inference / INLA tuỳ trường hợp. Các so sánh và phân tích mức độ phù hợp của Laplace được trình bày trong các bài báo phương pháp. ([CMU School of Computer Science][6])

---

5. Các phương pháp thay thế / bổ sung

- **MCMC** (HMC, NUTS) — ưu về độ chính xác (xấp xỉ posterior đúng), tốn thời gian tính toán; khi cần interval/tin cậy chính xác thì dùng MCMC (ví dụ Stan). ([Columbia Statistical Consulting][3])
- **Variational Inference (VI)** — tối ưu một họ phân phối gần giống posterior; nhanh, phù hợp mô hình lớn nhưng có thể dẫn đến underestimation of uncertainty.
- **Expectation Propagation (EP)** — thường cho xấp xỉ tốt hơn Laplace cho một số mô hình logistic; trade-off về độ phức tạp.
- **INLA** (Integrated Nested Laplace Approximation) — một phương pháp mở rộng dùng cho latent Gaussian models, rất hiệu quả cho mô hình không quá phức tạp cấu trúc. ([Becario Precario][8])

–