Maximum likelihood estimation và Maximum a posteriori estimation

Một cách khái quát, quá trình giải quyết một bài toán bằng Machine Learning thường trải qua ba bước:

- Modelling: Lựa chọn và xây dựng mô hình phù hợp cho bài toán
- Learning: Tối ưu các tham số của mô hình
- Inference: Sử dụng mô hình đã được huấn luyệnn để dự đoán với các dữ liệu mới

Các mô hình xác suất là nền tảng cho rất nhiều các mô hình thống kê trong Machine Learning. Cụ thể hơn, các mô hình xác suất được xây dựng dựa trên các phân phối xác suất cơ bản. Các phân phối này gồm có các tham số (ví dụ với Bernoulli distribution, tham số là biến γ, với Multivariate Normal Distribution, tham số là mean vector μ và ma trận hiệp phương sai $\sum$), các tham số này được ký hiệu chung là $theta$, được gọi là tham số của mô hình.

Learning chính là quá trình đánh giá (estimate) bộ tham số θ sao cho dữ liệu sẵn có và mô hình khớp với nhau nhất. Quá trình đó còn được gọi là parameter estimation.

Có hai cách đánh giá tham số thường được dùng là Maximum Likelihood Estimation (MLE) và Maximum A Posteriori Estimation (MAP Estimation).

- Maximum Likelihood Estimation chỉ dựa trên dữ liệu đã biết trong tập dữ liệu huấn luyện (training data).
- Maximum A Posteriori Estimation không những dựa trên training data mà còn dựa trên những thông tin đã biết của các tham số. Những thông tin này càng rõ ràng, càng hợp lý thì khả năng thu được bộ tham số tốt là càng cao.

1. Maximum likelihood estimation (MLE)

1.1. Ý tưởng chung

Với MLE, ta sẽ bắt đầu với một bộ dữ liệu huấn luyện gồm có N phần tử $X = x_1, x_2, \ldots, x_N$. Ta giả sử rằng bộ dữ liệu này tuân theo một phân phối xác suất nào đó, và xây dựng được mô hình Machine Learning thống kê được đại diện bởi tham số θ.

Maximum Likelihood Estimation là việc đi tìm bộ tham số θ sao cho xác suất sau đây đạt giá trị lớn nhất:

$$\theta = \max_{\theta} p(\mathbf{x}_1, \ldots, \mathbf{x}_N | \theta)$$

Mục tiêu ở đây là tìm được mô hình thống kê hay cụ thể hơn là tham số θ sao cho có thể mô tả được chính xác nhất bộ dữ liệu X. Do đó, ta có $p(\mathbf{x}_1|\theta)$ là xác suất mà điểm dữ liệu x_1 xuất hiện với điều kiện là tham số θ và $p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\theta)$ là xác suất mà toàn bộ bộ dữ liệu X cùng xuất hiện với tham số θ. Giá trị xác suất này được gọi là likelihood.

Ta đi tìm tham số θ để cực đại hoá likelihood, chính là cách để ta tìm tham số θ sao cho tạo ra được mô hình xác suất phản ảnh đúng nhất bộ dữ liệu huấn luyện cho trước.

1.2. Cách giải bài toán dựa vào MLE

Việc trực tiếp giải bài toán tối ưu trên thường rất phức tạp do tính phụ thuộc của các điểm dữ liệu trong bộ dữ liệu. Nói cách khác, xác suất xuất hiện của điểm dữ liệu này có thể phụ thuộc vào điểm dữ liệu khác dẫn đến việc tối ưu hoá xác suất đồng thời của các điểm dữ liệu gặp khó khăn.

Từ đó, để đơn giản hoá quá trình tối ưu, ta cần lập một giả sử các điểm dữ liệu trong bộ dữ liệu X độc lập với nhau. Khi các điểm dữ liệu được coi là độc lập với nhau, xác suất đồng thời của các điểm dữ liệu được tính bằng tích các xác suất của từng điểm dữ liệu. Từ đó, ta có biểu thức:

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\theta) \approx \prod_{n=1}^{N} p(\mathbf{x}_n|\theta)$$

Tuy nhiên, việc tối ưu một tích các giá trị xác suất thường khó khăn hơn việc tối ưu một tổng (do tích các xác suất có thể dẫn tới lỗi số học trong máy tính). Do đó, ta sẽ biến đổi phép nhân thành phép cộng thông qua việc sử dụng hàm logarit:

- log của một tích bằng tổng của các log.
- log là một hàm đồng biến, một biểu thức sẽ là lớn nhất nếu log của nó là lớn nhất, và ngược lại.

$$\theta = \max_{\theta} \sum_{n=1}^{N} \log\left(p(\mathbf{x}_n|\theta)\right)$$

Đến đây, việc giải bài toán tối ưu này có thể được thực hiện bằng nhiều phương pháp khác nhau như sử dụng đạo hàm, phương pháp nhân tử Lagrange ...

2. Maximum a posteriori estimation (MAP)

2.1. Ý tưởng chung

Với MLE, việc xây dựng mô hình và tìm tham số θ chỉ phụ thuộc vào bộ dữ liệu huấn luyện. Khi bộ dữ liệu này có vấn đề, hiển nhiên tham số θ tìm được cũng sẽ không chính xác.

Ví dụ đối với bộ dữ liệu thống kê về kết quả thu được khi tung đồng xu. Bộ dữ liệu ghi nhận trong 10.000 lần tung đồng xu, có 8.000 lần ra mặt ngửa và 2.000 lần ra mặt sấp. Tỷ lệ ra mặt sấp lúc này là 1/5 = 20%.

Tuy nhiên, với kiến thức của chúng ta, tỷ lệ ra mặt sấp và ra mặt ngửa đối khi tung đồng xu ra tương đối cân bằng nhau, loanh quanh ngưỡng 50%. Do đó, nếu ta xây dựng mô hình MLE trên bộ dữ liệu này thì có thể ta sẽ mắc phải hiện tượng overfitting.

Trong một số trường hợp cụ thể, bên cạnh việc xây dựng mô hình và lựa chọn tham số dựa trên bộ dữ liệu huấn luyện, ta còn cần phải định nghĩa trước một số kiến thức cho mô hình, và từ đó ngăn chặn việc mô hình quá phụ thuộc và bộ dữ liệu huấn luyện dẫn đến sai sót trong kết quả.

Ngược lại với MLE, MAP có biểu thức sau:

$$\theta = \arg\max_{\theta} \underbrace{p(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_N)}_{\text{posterior}}$$

$p(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_N)$ được gọi là Xác suất hậu nghiệm (Posterior Probability). Xác suất hậu nghiệm là xác suất được điều chỉnh hoặc cập nhật của một biến cố xảy ra sau khi xem xét thông tin mới.

Vậy tại sao xác suất hậu nghiệm lại có thể giúp ta bổ sung thêm thông tin mới?

2.2. Cách giải bài toán dựa vào MAP

Áp dụng quy tắc Bayes

$$\theta = \arg\max_{\theta} \underbrace{p(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_N)}_{\text{posterior}} = \arg\max_{\theta} \left[\frac{\overbrace{p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(\mathbf{x}_1, \ldots, \mathbf{x}_N)}_{\text{evidence}}} \right]$$

Trong Maximum A Posteriori (MAP), ta có một khái niệm được gọi là prior, đại diện cho những kiến thức đã có trước của con người muốn định hướng cho mô hình.

Evidence là giá trị xác suất hiển nhiên xảy ra, độc lập với mô hình xác suất hay tham số θ. Do evidence là độc lập với tham số θ, ta có thể loại nó ra khỏi biểu thức tối ưu của θ.

$$\theta = \arg\max_{\theta} \underbrace{p(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_N)}_{\text{posterior}} = \left[\arg\max_{\theta} \underbrace{p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\theta)}_{likelihood} \underbrace{p(\theta)}_{prior} \right] = \arg\max_{\theta} \left[\prod_{i=1}^{N} p(\mathbf{x}_i|\theta) p(\theta) \right]$$

Ta thấy điểm khác biệt giữa MAP và MLE nằm ở việc bổ sung thêm prior $p(\theta)$ vào trong biểu thức tối ưu.

Đến đây, câu hỏi đặt ra là làm sao để xác định được prior $p(\theta)$. Việc xác định $p(\theta)$ cũng phải bắt đầu từ việc lựa chọn một mô hình xác suất phù hợp, sao cho quá trình tối ưu posterior và likelihood trở nên thuận lợi. Cụ thể hơn, ta chọn mô hình xác suất của prior sao cho mô hình xác suất của posterior và likelihood là không đổi, lúc này, ta gọi prior là xác suất liên hợp (conjugate distribution) của likelihood.

Sau khi lựa chọn mô hình xác suất phù hợp với prior, ta đến với việc ước lượng tham số của prior. Trong thực tế, ta có thể ước lượng và lựa chọn tham số thông qua phương pháp cross-validation. Việc lựa chọn tham số cho prior sẽ giúp mô hình hạn chế hiện tượng overfitting, đặc biệt là trong những trường hợp mà ta có rất ít dữ liệu.