Câu 1:

1. Vì sao việc huấn luyện mô hình được coi là bài toán tối ưu hóa:

* Quá trình huấn luyện có mục tiêu tìm ra các tham số của mô hình có thể dự đoán chính xác nhất trên dữ liệu đầu vào
* Các hàm mục tiêu là đo lường sự khác biệt giữa giá trị dự đoán và giá trị thực và cũng phải tối thiểu hóa loss của mô hình

1. Ví dụ áp dụng cho Linear Regression và Logistic Regression

* Linear Regression:
  + Loss: MSE =

Trong đó yi là giá trị thực và y,i là giá trị dự đoán

* + Dùng Gradient Desent để tối ưu hóa đẻ giảm giá trị cho loss function
* Logistic Regression:
  + Loss: Cross Entropy = -

Trong đó yi là giá trị thực và y,i là giá trị dự đoán

* + Dùng Gradient Desent để tối ưu hóa đẻ giảm giá trị cho loss function

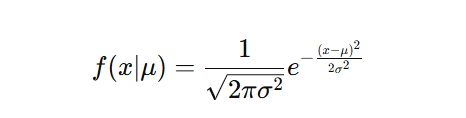
1. Vai trò Loss Function:

* Hàm mất mát là thước đo định lượng chất lượng dự đoán của mô hình. Nó xác định "mức độ sai lệch" giữa dự đoán của mô hình và giá trị thực tế trong dữ liệu huấn luyện.
* Hàm mất mát xác định **đích đến** cho bài toán tối ưu hóa, giúp mô hình đạt được hiệu suất dự đoán tốt nhất.

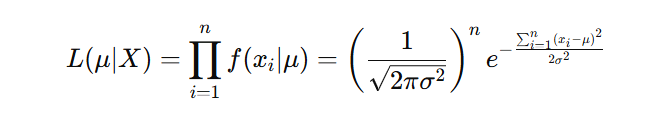
Câu 2:

1. Đạo hàm MLE với mean μ.

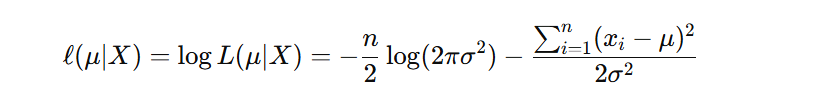
Hàm mật độ xác suất là:



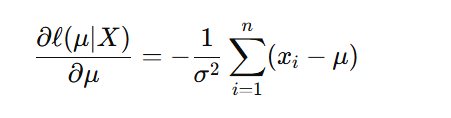
Hàm likelihood:



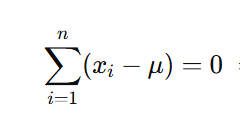
Hàm log-likelihood là:



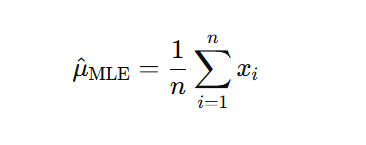
Đạo hàm tại μ tối ưu l(μ|X):



⬄

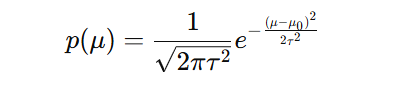


Khi đó:

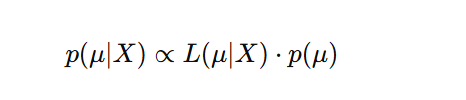


1. Đạo hàm MAP mean μ

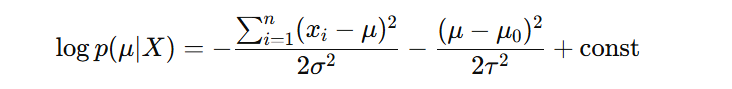
Hàm mật độ tiên nghiệm:



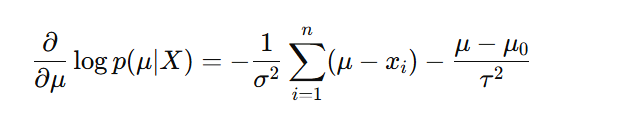
Hàm xác suất hậu nghiệm:



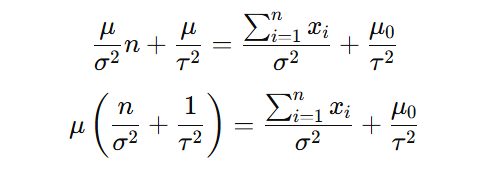
Thay l(μ|X) và log p(μ):



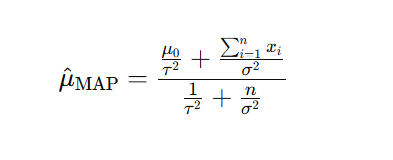
Đạo hàm tại μ:



⬄



Do đó, **MAP** cho μ:



1. So sánh MLE và MAP và ảnh hưởng μ0​ và ::

* So sánh MLR và MAP:
  + **MLE**: Chỉ dựa vào dữ liệu quan sát. Khi n→∞. MLE và MAP dần trở nên giống nhau vì ảnh hưởng của prior giảm.
  + **MAP**: Kết hợp dữ liệu quan sát và thông tin tiên nghiệm. MAP "làm mịn" ước lượng bằng cách thêm ràng buộc từ prior.
* Ảnh hưởng μ0​ và :
  + μ0​: Giá trị trung bình tiên nghiệm. MAP bị "kéo" về phía μ0\mu\_0μ0​, nhất là khi nhỏ.
  + : Độ biến thiên của prior. Giá trị lớn khiến prior ít ảnh hưởng, MAP gần giống MLE. Giá trị nhỏ làm MAP phụ thuộc nhiều vào prior, giảm ảnh hưởng của dữ liệu.

Câu 3:

1. Giải thích Naïve Bayes assumtion và cách áp dụng:

* Giả định chính của Naive Bayes là các đặc trưng (từ trong văn bản) là độc lập có điều kiện với nhau, tức là:

P(x1,x2,…,xn∣C)=P(x1∣C)⋅P(x2∣C)⋯P(xn∣C)

Trong đó xi là các đặc trưng và C là nhãn lớp.

* Áp dụng:
  + Trong bài toán phân loại văn bản, mỗi tài liệu được coi là một tập hợp các từ trong từ vựng.
  + Mục tiêu là tính xác suất P(C∣X)P(C | X)P(C∣X) (xác suất tài liệu thuộc một lớp C với X là tập hợp các từ), dựa trên Bayes' Rule
  + Với giả định độc lập có điều kiện, ta có thể viết:

P(X|C) =

1. Áp dụng cho data ở trên

* Công thức Laplace smoothing:

P(x|C) =

* + a = 1
  + V: Số lượng từ cần dùng
  + Nclass: tổng số từ trong văn bản
* Tổng số từ trong Sports: 135
* Tổng số từ trong Politics: 165
* P(win|Sport) =
* P(vote|Sport) =
* P(win, vote|Sport) =
* P(win|Politics) =
* P(vote|Politics) =
* P(win, vote| Politics) =

1. Đánh giá kết quả và hạn chế

* Dựa trên kết quả tính toán:
  + Nếu P(Sports∣win, vote)>P(Politics∣win, vote) tài liệu được phân loại vào Sports.
  + Ngược lại, tài liệu sẽ được phân loại vào Politics.
  + Tổng hợp hai từ này, mô hình sẽ đưa ra một kết quả phân loại dựa trên sự kết hợp xác suất, có tính đến việc điều chỉnh Laplace để đảm bảo không có xác suất bằng 0
* Hạn chế:
  + Giả định độc lập giữa các từ
    - Mô hình giả định rằng các từ trong tài liệu là độc lập, tức là xác suất của một từ không bị ảnh hưởng bởi từ khác. Tuy nhiên, trong thực tế, các từ có mối quan hệ ngữ cảnh mạnh mẽ.
  + Nhạy cảm với dữ liệu không cân bằng:
    - Nếu một danh mục (ví dụ: **Politics**) có số lượng tài liệu huấn luyện lớn hơn đáng kể so với danh mục còn lại (**Sports**), mô hình sẽ có xu hướng thiên về danh mục lớn hơn, dẫn đến sai lệch.
  + Không phân biệt được tầm quan trọng của từ:
    - Mô hình đối xử tất cả các từ như nhau, không phân biệt từ quan trọng và từ ít quan trọng đối với phân loại.

Câu 6:

1. Sự khác biệt L1 (Lasso) và L2(Ridge) Regularization

* L1 Regularization:
  + Thêm vào Loss Function một thành phần là tổng giá trị tuyệt đối của các hệ số trọng số:

Loss = MSE + λ

* + L1 làm cho hệ số trọng số giảm xuống đúng bằng 0. Lasso có thể chọn lọc đặc trưng và loại bỏ các đặc trưng không quan trọng
* L2 Regularization:
  + Thêm vào Loss Function một thành phần là tổng bình phương của các hệ số trọng số:

Loss = MSE + λ

* + L2, thay vì loại bỏ các đặc trưng thì nó vẫn có khả năng lưu lại các đặc trưng nó giúp mô hình lưu được nhiều thông tin hơn và ổn định hơn

1. Dữ liệu có nhiều đặc trưng tương quan cao, nên chọn L1 hay L2?

* Trong trường hợp các đặc trưng có tương quan cao:
  + L2 (Ridge) thường phù hợp hơn. Lý do:

Ridge Regularization có xu hướng phân phối trọng số đồng đều giữa các đặc trưng tương quan. Điều này giúp mô hình ổn định hơn khi các đặc trưng không độc lập.

* + L1 (Lasso) có thể không hiệu quả trong trường hợp này. Lý do:

Khi các đặc trưng tương quan cao, Lasso sẽ chọn ngẫu nhiên một hoặc vài đặc trưng và loại bỏ những đặc trưng khác, làm giảm độ ổn định của mô hình.

Câu 4, Câu 5: