HỌ TÊN: CAO SỸ NGUYÊN VŨ

MSV: 22001665

Câu 1:

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

Quá trình huấn luyện có mục tiêu tìm kiếm các tham số của mô hình sao cho mô hình có thể đưa ra dự đoán chính xác nhất trên tập dữ liệu. Để đạt được mục tiêu này cần đo lường sự khác biệt giữa giá trị dự đoán và giá trị thực (gọi là hàm mất mát) và tìm các tham số để hàm mất mát đạt giá trị nhỏ nhất

b) Ví dụ áp dụng cho Linear Regression và Logistic Regression
   - Linear Regression:
     - Loss: MSE = $\frac{1}{n}\sum_{i=1}^{n}(y\_r - y\_p)^2$
       Trong đó y_r là giá trị thực, y_p là giá trị dự đoán, n là số lượng dữ liệu
     - Dùng thuật toán giảm gradient để tối ưu hóa để giảm giá trị cho hàm mất mát
   - Logistic Regression:
     - Loss: Cross Entropy = $-\frac{1}{n}\sum_{i=1}^{n} y\_r(\log(\mathrm{y_p})) + (1 - y\_r)\log(1 - \mathrm{y\_p})$
       Trong đó y_r là giá trị thực, y_p là giá trị dự đoán, n là số lượng dữ liệu
     - Dùng thuật toán giảm gradient để tối ưu hóa để giảm giá trị cho hàm mất mát

c) Vai trò Loss Function:
   - Hàm mất mát (loss function) đóng vai trò cực kỳ quan trọng trong quá trình huấn luyện mô hình học máy. Nó là một thước đo số lượng để đánh giá mức độ sai lệch giữa giá trị dự đoán của mô hình và giá trị thực tế.
   - Đánh giá hiệu suất: Hàm mất mát cho phép đánh giá một cách định lượng hiệu suất của mô hình. Một hàm mất mát nhỏ cho thấy mô hình đang dự đoán khá chính xác.
   - Hướng dẫn quá trình học: Hàm mất mát đóng vai trò như một "mục tiêu" mà mô hình cố gắng đạt được. Trong quá trình huấn luyện, mô hình sẽ điều chỉnh các tham số của nó để giảm thiểu giá trị của hàm mất mát.
   - Tối ưu hóa mô hình: Các thuật toán tối ưu hóa (như giảm gradient) sẽ sử dụng gradient của hàm mất mát để cập nhật các tham số của mô hình theo hướng giảm giá trị của hàm mất mát.

Câu 2:

a) Đạo hàm MLE với mean μ.

Hàm mật độ xác suất là:

$$f(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Hàm likelihood:

$$L(\mu|X) = \prod_{i=1}^{n} f(x_i|\mu) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

Hàm log-likelihood là:

$$\ell(\mu|X) = \log L(\mu|X) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$$

Đạo hàm tại μ tối ưu l(μ|X):

$$\frac{\partial \ell(\mu|X)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

⇔

$$\sum_{i=1}^{n}(x_i-\mu) = 0$$

Khi đó:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

b) Đạo hàm MAP mean μ
Hàm mật độ tiên nghiệm:

$$p(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\mu-\mu_0)^2}{2\tau^2}}$$

Hàm xác suất hậu nghiệm:

$$p(\mu|X) \propto L(\mu|X)\cdot p(\mu)$$

Thay l(μ|X) và log p(μ):

$$\log p(\mu|X) = -\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\tau^2} + \text{const}$$

Đạo hàm tại μ:

$$\frac{\partial}{\partial\mu}\log p(\mu|X) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(\mu-x_i) - \frac{\mu-\mu_0}{\tau^2}$$

⇔

$$\frac{\mu}{\sigma^2}n + \frac{\mu}{\tau^2} = \frac{\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{\mu_0}{\tau^2}$$

$$\mu\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right) = \frac{\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{\mu_0}{\tau^2}$$

Do đó, MAP cho μ:

$$\hat{\mu}_{\text{MAP}} = \frac{\frac{\mu_0}{\tau^2} + \frac{\sum_{i-1}^{n} x_i}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}$$

c) So sánh MLE và MAP và ảnh hưởng μ0 và r^2::
- So sánh MLR và MAP:
  - MLE: Tìm giá trị của tham số làm cho khả năng (likelihood) của dữ liệu quan sát được là lớn nhất. Nói cách khác, MLE tìm ra giá trị tham số khiến cho dữ liệu hiện tại có vẻ "hợp lý" nhất.
  - MAP: Kết hợp giữa MLE và kiến thức trước đó về phân bố của tham số (prior). MAP tìm ra giá trị của tham số làm cho hậu nghiệm (posterior) của tham số là lớn nhất, tức là kết hợp cả thông tin từ dữ liệu và thông tin trước đó.
- Ảnh hưởng μ0 và r^2:
  - μ0: Giá trị trung bình tiên nghiệm. MAP bị "kéo" về phía μ0 nhất là khi r^2 nhỏ.
  - r^2: 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:

a) 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) = $\prod_{i=1}^{n} P(xi|C)$

b) Áp dụng cho data ở trên
- Công thức Laplace smoothing:

P(x|C) = $\frac{freq(x,C)+a}{Nclass+a.V}$
  - 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) = $\frac{50+1}{135+4} = \frac{51}{139}$
- P(vote|Sport) = $\frac{10+1}{135+4} = \frac{11}{139}$

⇨ P(win, vote|Sport) = $\frac{51}{139} * \frac{11}{139}$

- P(win|Politics) = $\frac{10+1}{165+4} = \frac{11}{169}$
- P(vote|Politics) = $\frac{80+1}{165+4} = \frac{81}{169}$

⇨ P(win, vote| Politics) = $\frac{51}{139} * \frac{11}{139}$

c) 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:

a) 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 + $\lambda\sum_{i=1}^{n}(|w\_i|)$

  - L1 làm cho hệ số trọng số giảm xuống đúng bằng 0. Điều này giúp 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:

      $$\text{Loss} = \text{MSE} + \lambda \sum_{i=1}^{n} w\_i^2$$

    - L2 làm giảm giá trị của các hệ số gần về 0. Điều này giúp giữ lại tất cả các đặc trưng và làm giảm tầm quan trọng của những đặc trưng ít ảnh hưởng

b) 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:

https://github.com/Vux-Cao/homework1