# HOMEWORK 1

Nguyễn Thùy Trang - 22000128

October 2024

## 1 Problem 1: Machine Learning as an Optimization Problem

### 1.1 Explain why training a machine learning model can be formulated as an optimization problem. What are the objectives and constraints involved?

Mục tiêu của quá trình training một mô hình học máy là để điều chỉnh bộ tham số nhằm cực tiểu hoá hàm mất mát (độ sai lệch giữa predicted output và labels). Hàm mất mát trong học máy được đánh giá là hàm mục tiêu.

| | Linear Regression | Logistic Regression |
|---|---|---|
| Output | Numeric | Binary |
| Objective function | MSE | Binary cross-entropy (loss log) |
| Function | Linear | Sigmoid |
| Optimation Tech | Gradient descent | Gradient descent |

Bảng 1: Caption

### 1.2 Provide examples of how optimization techniques are applied in the training of models such as linear regression and logistic regression.

### 1.3

# 2 Problem 2: Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP)

Given a dataset of independent and identically distributed observations X = {x1, x2, . . . , xn} drawn from a normal distribution with unknown mean and known variance $\sigma^2$.

### 2.1 Derive the Maximum Likelihood Estimator (MLE) for the mean $\mu$.

$$P(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

Likelihood function:

$$L(\mu) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

Likelihood function:

$$\ell(\mu) = \frac{1}{2\sigma^2}(x_i - \mu)^2 - nlog\sigma - \frac{n}{2}log2\pi$$

Đạo hàm $\ell(\mu)$ với biến $\mu$:

$$\frac{d\ell(\mu)}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu).$$

$\frac{d\ell(\mu)}{d\mu} = 0$:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0.$$

$$\sum x_i - n\mu = 0$$

$$\mu = \frac{\sum x_i}{n}$$

### 2.2 Assume a prior distribution for $\mu$ that is also normally distributed with mean $\mu_0$ and variance $T$. Derive the Maximum A Posteriori (MAP) estimator for $\mu$.

Phân phối hậu nghiệm:

$$P(\mu|X) \propto L(\mu)P(\mu)$$

$$p(\mu|x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi T^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2T^2}\right)$$

Log-posterior:

$$\log P(\mu|X) = \log L(\mu) + \log P(\mu) + \text{c.}$$

$$\log P(\mu|X) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{2T^2}(\mu-\mu_0)^2 + \text{c.}$$

Khai triển:

$$\log P(\mu|X) = -\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{T^2}\right)\mu^2 + \left(\frac{\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{\mu_0}{T^2}\right)\mu + \text{hằng số.}$$

Đạo hàm theo $\mu$ và cho bằng 0:

$$\frac{d}{d\mu}\log P(\mu|X) = -\left(\frac{n}{\sigma^2} + \frac{1}{T^2}\right)\mu + \left(\frac{\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{\mu_0}{T^2}\right) = 0.$$

Suy ra:

$$\mu_{\text{MAP}} = \frac{\frac{\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{\mu_0}{T^2}}{\frac{n}{\sigma^2} + \frac{1}{T^2}}.$$

# 3 Naive Bayes Classification

You are provided with a simplified dataset of text documents classified into two categories: Sports and Politics. The vocabulary consists of the words: win, team, election, and vote.

| **Word** | **Sports Count** | **Politics Count** |
|---|---|---|
| win | 50 | 10 |
| team | 60 | 5 |
| election | 15 | 70 |
| vote | 10 | 80 |

### 3.1 Explain the Naive Bayes assumption and how it applies to text classification.

Naive Bayes assumption: $P(y|X) = \frac{P(X|y)P(y)}{P(X)}$
Gọi X là tập hợp các từ $x_i$, Y là tập các nhãn $y \in$ Sport, Politics
Ta sử dụng Naive Bayes để tính xác suất $x_i$ thuộc nhãn Sport và Politics, từ đó phân loại và nhãn có xác suất cao nhất.

### 3.2 Using the data above, calculate the probability that a document containing the words win and vote belongs to the Sports category versus the Politics category. Assume uniform class priors and apply Laplace smoothing with $\alpha = 1$.

Xác suất nhãn : $P(Politics) = P(Sports) = 0,5$
Xác suất mỗi từ thuộc nhãn Sports hoặc Politics:
$P(win|Sports) = \frac{50+1}{135+14} = \frac{51}{139}$
$P(vote|Sports) = \frac{10+1}{135+14} = \frac{11}{139}$
$P(win|Politics) = \frac{10+1}{165+14} = \frac{11}{169}$
$P(vote|Politics) = \frac{80+1}{165+14} = \frac{181}{169}$
Xác suất nhãn Sports hoặc Politics :

$$P(doc|sports) = P(win|sports)P(vote|sports) = 0,029$$

$$P(doc|politics) = P(win|Politics)P(vote|Politics) = 0,032$$

Xác suất documents thuộc nhãn "Sports"và "Politics":

$$P(Sports|doc) = P(Sports).P(doc|Sports) = 0,5.0,032 = 0,0145$$

$$P(Politics|doc) = P(Politics).P(doc|Sports) = 0,5.0,032 = 0,0156$$

Như vậy, documents chứa "Win"và "Vote"có khả năng cao hơn thuộc class "Politics".

## 4 Logistics Regression

Consider a binary classification problem where the goal is to predict whether a student will pass or fail an exam based on the numbers of hours spent fo studying and sleeping. Formulate the logistic regression model for this problem.

Input:

- $x_1$(Số giờ ngủ mỗi ngày)
- $x_2$(Số giờ học mỗi ngày)

Labels: $y \in \{0, 1\}$

- y = 0: fail
- y = 1: pass

Tổ hợp tuyến tính : $z = \theta_0 + \theta_1 x_1 + \theta_2 x_2$
Xác suất để 1 học sinh pass:

$$P(y = 1|x) = \sigma(z) = \frac{1}{1 + e^{-}z}$$

Xác suất để 1 học sinh fail:

$$P(y = 0|x) = 1 - P(y = 1|x)$$

Mô hình dự đoán Pass (y = 1) nếu:

$$P(y = 1|x) > P(y = 0|x) => P(y = 1|x) > 0,5$$

Mô hình cực tiểu hoá hàm mất mát Cross - Entropy:

$$\ell(y, \hat{y}) = -\sum_{i=1}^{n} y_i log\hat{y}_i + (1 - y_i)log(1 - \hat{y}_i)$$

Mục tiêu là tìm bộ tham số $\theta = (\theta_0, \theta_1, \theta_2)$ thoả mãn $\ell(y, \hat{y})$ đạt min.

# 5 Linear Regression and Overfitting

Code: https://github.com/aya1101/HUS$_M L$

# 6 Regularization

Regularization là một hướng tiếp cận nhằm ngăn chặn Overfitting trong mô hình. Hàm mục tiêu sau regularization có dạng:

$$J'(\theta) = J(\theta) + \lambda R(\theta)$$

| Tiêu chí | L1 Regularization (Lasso) | L2 Regularization (Ridge) |
|---|---|---|
| **Penalty** | $\sum \lvert\theta_i\rvert$ | $\sum(\theta_i)^2$ |
| **Hệ số** | Một số hệ số sẽ trở thành 0 | Tất cả hệ số sẽ giảm nhưng không bằng 0 |
| **Lựa chọn đặc trưng** | Có (feature selection) | Không |
| **Ứng dụng** | Dữ liệu có nhiều đặc trưng không quan trọng | Khi muốn sử dụng tất cả đặc trưng |

Bảng 2: So sánh L1 và L2 Regularization

## 6.1 Explain the difference between L1 (Lasso) and L2 (Ridge) regularization in the context of linear regression.

## 6.2 Given a dataset with multiple features are highly correlated, discuss which regularization method would be more appropriate and why.

Với những dữ liệu có nhiều đặc trưng tương quan lẫn nhau, sử dụng L2 Regularization hợp lý hơn. Vì L2 Regularization không không loại bỏ hoàn toàn các đặc trưng mà phân bố hợp lý các trọng số của nó. Từ đó giảm nguy cơ loại bỏ sai đặc trưng quan trọnh.
Ngoài ra, có thể sử dụng Elastic Net Regularization với ưu điểm của cả L1 và L2.