Bài tập

I. Tổng quan machine learning.

Bài 1: Sự khác nhau giữa classification và regression trong bài toán học giám sát?

Regression cho đầu ra là 1 con số cụ thể (Label là một biến liên tục) còn Classification cho đầu ra là một biến phân loại (Label là một biến phân loại).

Bài 2: Sự khác nhau giữa parameter và hyperparameter?

Hyperparameter được quy định trước quá trình traning còn parameter được tạo ra trong quá trình training.

Bài 3: Giả sử dữ liệu có 5 features thì sẽ có bao nhiêu parameter?

Dữ liệu có 5 features thì sẽ có 5 parameter.

Bài 4: Tìm model là tìm cái gì?

Để tìm model f(x) là tìm bộ tham số w của model.

Bài 5: Làm thế nào để tìm model?

Để tìm model thì ta tìm bộ tham số w của model. Để tìm bộ tham số của model ta cực tiểu hàm mát mát. Bộ tham số thu được là bộ tham số làm cho hàm mất mát đạt giá trị cực tiểu.

Bài 6: Tại sao phải cần tập validation?

Do dữ liệu tập test không được thấy trong quá trình training. Mục đích của validation là để chọn hyperparameter tối ưu.

Bài 7: Dữ liệu trong mỗi group có đặc điểm gì?

Được phân chia thành các cụm , mỗi cụm có những tính chất giống nhau, là tập hợp các điểm có cùng vector đặc trưng gần nhau.

Bài 8: Tại sao phải cần giảm chiều dữ liệu?

Số lượng điểm dữ liệu và kích thước các vector đặc trưng thường rất lớn trong các bài toán thực tế. Nếu thực hiện lưu trữ và tính toán trực tiếp trên dữ liệu có số chiều lớn thì sẽ khó khăn cả về việc lưu trữ và tốc độ tính toán => Cần giảm chiều dữ liệu

Bài 9: Tại sao cần điều chỉnh hyperparameter?

Việc thay đổi giá trị các hyperparameter trên sẽ ảnh hưởng đến độ chính xác của model, công việc của chúng ta là tìm cho bằng được hyperparameter để tăng độ chính xác của mô hình.

II. Ovefitting và Underfitting

Bài 1: Overfitting là gì, nguyên nhân, giải pháp?

- Định nghĩa :Overfitting là hiện tượng mô hình tìm được quá khớp với dữ liệu tập train. Tính ghi nhớ cao khiến tính tổng quát thấp.
- Nguyên nhân:
  - Nhiễu
  - Ít dữ liệu
  - Model quá phức tạp
- Giải pháp:
  - Tăng dữ liệu
  - Sử dụng Early Stopping
  - Sử dụng regulazation
  - Sử dụng validation

Bài 2: Underfitting là gì, giải pháp?

- Định nghĩa : Underfitting là hiện tượng model không có tính ghi nhớ cũng không có tính tổng quát, tức là model cho kết quả không tốt trên cả tập train và tập test.
- Giải pháp:
  - Bổ sung thêm dữ liệu đầu vào.
  - Chọn thuật toán khác

Bài 3: Sự khác nhau giữa regularization l1 và l2?

- l1 regularization có đạo hàm không xác định tại 0 dẫn đến việc tìm nghiệm thường tốn thời gian còn l2 regularization có đạo hàm xác định mọi nơi.
- l1 regularization giúp cho mô hình ít bị ảnh hưởng bởi nhiễu hơn so với l2 regularization.
- l1 regularization khi đưa vào hàm mục tiêu, nó sẽ thường cho ra model thưa (sparse models), tức là model có parameter chứa nhiều chiều bằng 0 (tức làm cho các parameter của các features không cần thiết về 0). Còn l2 regularization thường cho ra model không thưa (non-sparse models) đầy đủ các tham số nhưng vẫn tránh được overfitting.
- l1 regularization là nó có thể cực tiểu về 0 . Giả sử $w_i = 0$ nên $w_i * x_i = 0$ nên nó sẽ triệt tiêu feature đó luôn. Còn l2 thì nó có thể gần bằng 0 chứ không bằng 0 nên feature đó vẫn còn. Nhưng độ ảnh hưởng của nó lên model sẽ giảm.

- Chú ý: Sparse-matrix hay ma trận thưa là ma trận có nhiều phần tử bằng 0 nên dùng cách này dễ dàng lưu trữ, xử lý vì ta chỉ quan tâm đến các giá trị bằng 0.

Bài 4: Tại sao khi tăng dữ liệu có thể giảm thiểu overfitting?

Khi có thêm dữ liệu thì tất nhiên ta có thêm thông tin về các mỗi quan hệ giữa feature và label . Khi có nhiều dữ liệu thì hàm mất mát trên dữ liệu training sẽ gần bằng với hàm mất mát trên toàn tập dữ liệu khi dữ liệu training càng lớn. khi hai đại lượng này trùng nhau thì overfitting hoàn toàn biến mất. Vì vậy thì khi tăng dữ liệu lên có thể giảm thiểu overfitting.

Bài 5: Tại sao có sự đánh đổi giữa tính ghi nhớ và tính tổng quát?

Khi train model ta chỉ muốn tìm ra model dự đoán tốt trên trên dữ liệu train nhưng không có nghĩa nó dự đoán tốt trên dữ liệu test. Trong quá trình traning thì dữ liệu có thể không được như mong muốn (có các điểm dữ liệu nhiễu) dẫn đến model phải bỏ qua các điểm dữ liệu này dẫn đến tính ghi nhớ bị giảm xuống, tính tổng quát tăng lên nên dự đoán model trên tập training chỉ mang tính tương đối chính xác nhưng bù lại model sẽ tổng quát hơn và dự đoán chính xác hơn trên dữ liệu test.

Khi model phức tạp tính ghi nhớ tốt nhưng nó không mô tả được quan hệ của dự liệu, nó không theo một quy luật nào cả nên tính tổng quát giảm. Ngược lại khi model đơn giản thì nó có quy luật rõ ràng ví dụ như 1 đường thẳng thì tính tổng quát nó tốt. Nhưng do nó đơn giản quá nên tính ghi nhớ nó không tốt.