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

* Classification (phân loại) : Là bài toán có các label của data được chia thành các số hữu hạn nhóm

Ví dụ như xác định người này có bị bệnh hay không?, thanh toán này có gian lận hay không? , người này là nam hay nữ ? ,….

* Regression (Hồi quy) : Là bài toán mà label của data là một số thực nằm trong dải (-∞, +∞)

Ví dụ như : Các bài toán về dự đoán giá trị sản phẩm, dự đoán chiều cao cân nặng của một người,….

Câu 2: Sự khác nhau giữa parameter và hyperparameter?

* Parameter: là các giá trị của model được sinh ra trong quá trình training

Giả sử: model f(x) = w1\*x1 + w2\*x2 + w3\*x3 thì các w1, w2, w3 được gọi là các parameter

Khi chúng ta nói tìm được model tốt nhất cho bài toán thì có nghĩa là chúng ta đã tìm được các parameter phù hợp cho tập dữ liệu ban đầu

* Hyperparameter: là các giá trị không được sinh ra từ quá trình training mà chúng được định nghĩa bởi con người. Nó không phụ thuộc vào tập dữ liệu huấn luyện

Hyperparameter được sử dụng trong quá trình training giúp tìm ra các parameter phù hợp nhất

Câu 3: Giả sử có 5 features thì có bao nhiêu parameter?

* Nếu tập dữ liệu có 5 features thì có 5 parameter

Câu 4: Tìm model là tìm cái gì?

* Tìm model tức là tìm hàm f(x) hay tìm bộ parameter

Câu 5: Làm thế nào để tìm model?

* Để tìm model ta tối thiểu hàm mất mát từ đó tìm được các parameter

Câu 6: Tại sao phải cần tập validation?

* Hyperparameter được sử dụng trong quá trình training giúp tìm ra các parameter phù hợp nhất vì vậy việc chọn các hyperparameter rất quan trong
* Tập validation vừa giúp chúng ta thử được các bộ số hyperparameter để tìm ra bộ số phù hợp vừa phù hợp với nguyên tắc dữ liệu ở tập test không được thấy trong quá trình train

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

* Dữ liệu trong mỗi group có các đặc điểm tương đồng vd như về màu sắc, Kích thước, hình dạng ,…….

Câu 8: Tại sao cần giảm chiều dữ liệu?

* Các feature trong các bài toán thực tế có thể có số chiều rất lớn hàng nghìn chiều. Ngoài ra số lượng các điểm dữ liệu rất lớn. Gây khó khăn trong việc lưu trữ và tốc độ tính toán tính toán. Giảm chiều dữ liệu nén tập dữ liệu giúp cho việc lưu trữ trở nên dễ dàng và tốc độ tính toán được cải thiện

Câu 9: Tại sao cần điều chỉnh hyperparameter?

* Hyperparameter được người tham gia qua trình huấn luyện định nghĩa vì vậy không có công thức cũng như phương pháp để tìm. Mỗi bộ hyperparameter lại cho ra bộ parameter khác nhau nên chúng ta cần điều chỉnh nó để tìm ra bộ parameter phù hợp nhất với bộ dữ liệu ban đầu

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

* Overfitting là hiện tượng model cho kết quả tốt trên tập dữ liệu train nhưng lại cho kết quả xấu trên tập dữ liệu test
* Nguyên nhân:

+ Do nhiễu

+ Quá ít dữ liệu

+ Model quá phức tạp

* Giải pháp:

+ Tăng tập dữ liệu

+ Early stopping

+ Sử dụng regulazation

Bài 2: Underfitting là gì, giải pháp?

* Underfitting là hiện tượng model cho kết quả xấu trên cả tập dữ liệu train và test
* Giải pháp:

+ Thay đổi thuật toán trong trường hợp mình sử dụng sai thuật toán hoặc thuật toán sử dụng chưa hợp lí

+ Dữ liệu đầu vào còn ít cần bổ sung thêm

Bài 3:

* **L1**: Là tổng của trị tuyệt đối của các thành phần w. Khi đưa vào hàm mục tiêu, nó sẽ thường cho ra model thưa có parameter chứa nhiều chiều bằng 0.
* **L2:** Là bình phương độ dài của vector w. Nó giúp chúng ta tính đạo hàm được dễ hơn khi tối ưu hàm mục tiêu.

Bài 4: Tại sao khi tăng dữ liệu lại giảm overfitting?

* Khi tăng tập dữ liệu train thì Ldtrain dần hội tụ về Ld. Khi Ldtrain = Ld thì hiện tượng overfitting hoàn toàn biết mất điểu này xảy ra khi tập dữ liệu Dtrain tiến tới vô cùng

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

* Một model ghi nhớ tốt tức là nó dự đoán đúng tất cả dữ liệu trên tập train vì vậy khi dữ liệu tập train quá lớn làm cho model ngày càng phức tạp. Khi đưa 1 dữ liệu mới không có trong tập train thì model dự đoán bị sai lệch.