1. Trình bày cách huấn luyện các Polynomial regression models.

Cách huấn luyện các Polynomial regression models giống như huấn luyện linear model với những feature mới, nhiều feature hơn, những feature mới được tạo ra bằng cách dùng feature cũ nâng bậc nó lên:

Có 2 bước huấn luyện PR models:

- Thêm vào feature bậc cao (VD: x1, x2, x3, x4, x5(x12), x6(x22),…)

- Chạy thuật toán training Linear regression để huấn luyện

1. Nêu đặc điểm learning curves của underfitting, overfitting và good-fitting models.

Learning curves là vẽ performance theo lượng dữ liệu đưa vào

Đặc điểm learning curves của underfitting: error ở tập training và validation đều cao và gần nhau

Đặc điểm learning curves của overfitting: error giảm xuống nhiều hơn underfiting, nhưng khoảng cách (gap) giữa tập training và validation lớn

Đặc điểm learning curves của good-fitting models: error ở tập training và validation đều thấp và gần nhau

1. Trình bày ý tưởng của regularization. Tại sao regularization có thể giúp ngăn overfitting?

Khi bị hiện tượng overfitting thì có nhiều cách giải quyết trong đó có cách giảm độ phức tạp model. Ngoài ra còn có cách khác là regularization, cách này sử dụng model cũ nhưng thêm ràng buộc để model bị overfitting ít dao động đi tạo hình dáng đồ thì gần giống với hình dạng đồ thị bậc thấp hơn để mô tả chính xác dữ liệu.

Regularization giúp ngăn overfitting vì một model bậc cao thì sẽ giảm dao động đi các cực trị hình parabol ở vị trí quá cao thì sẽ giảm dần đi tạo ra hình dạng đồ thị ít dao động và nó gần giống như các đồ thị bậc thấp hơn, đồ thị từ phức tạp so với dữ liệu thành trở nên đơn giản hơn mô tả chính xác dữ liệu hơn. Do đó sẽ ngăn được hiện tượng overfittingNếu cần train 1 data set có một triệu features thì nên sử dụng Stochastic gradient descent (SGD) vì thời gian nó chạy nhanh hơn nhiều so với Normal equation và Batch gradient descent (BGD) và số lượng feature lớn nếu dùng Normal equation hoặc BGD thì nó sử dụng rất nhiều bộ nhớ, thời gian nó chạy rất rất lâu và có thể nó không chạy luôn

1. Nếu cần train 1 data set có một triệu features bạn sẽ chọn thuật toán training nào: Normal equation hay Gradient descent? Tại sao?

Chọn Gradient descent vì normal equation có bước thực hiện là phải tìm ma trận nghịch đảo, đây là một phép toán phức tạp. Khi tập dữ liệu có lượng features lớn thì phép toán tìm ma trận nghịch đảo càng phức tạp và chạy lâu hơn. Mà Gradient descent thì không sử dụng ma trận nghịch đảo chỉ dùng tính chất đạo hàm để tìm ra theta để error bé nhất do đó GD nhanh hơn so với normal equation. Nếu chạy các thuật toán Gradient descent đủ lâu thì chúng không đưa đến những hypothesis giống nhau vì nó chỉ lấy 1 sample ngẫu nhiên để làm.

1. Nếu data set của bạn có các features với khoảng giá trị rất khác nhau thì những thuật toán training nào (trong 2 thuật toán: Normal equation hay Gradient descent) sẽ gặp vấn đề? Vấn đề đó là gì? Làm sao để xử lý?

Ta nên dùng Ridge regression hoặc các regularization models khác hơn là dùng Linear regression hoặc Polinomial regression vì:

- Linear thì cái hypothesis nó đơn giản, khó qua được hết các điểm và dễ bị underfitting

- Polinomial thì đôi khi chạy thì model phức tạp, dễ dẫn đến hiện tượng overfitting, đôi khi tập giá trị cách xa nhau thì error nó lớn.

- Rgularization nó ngăn được overfitting và error ở tập validation và trainning gần nhau và thấp

1. Nếu chạy các thuật toán Gradient descent đủ lâu, chúng có đưa đến các hypothesis giống nhau không? Tại sao?

Không giống nhau, nhưng các thuật toán ấy chạy ra kết quả gần xắp xỉ nhau.

Vì tùy vào thuật toán như batch gradient descent thì kết quả sẽ tìm đến chỗ cực tiểu với giá trị chính xác cao nhưng có thể bị hiện tượng là minimal local bị hiện tượng này thuật toán chạy đủ lâu nhưng không thể khắc phục được, điều này tùy thuộc chọn vị trí theta ngẫu nhiên. Thuật toán SGD thì chạy kết quả không chính xác như BGD nhưng nó khắc phục được hiện tượng local minimal khi chạy thuật toán đủ lâu thì có thể tìm đến chỗ cực tiểu ấy nhưng thuật toán ấy sẽ không dừng lại ngay đến khi đủ số vòng lặp nên có thể lần chạy sau khi tìm ra điểm cực tiểu thì kết quả có thể bị sai lệch đi. Thuật toán mini batch gradient descent thì chạy ra kết quả sẽ chính xác hơn vì chạy một lần sẽ chọn nhiều sample ngẫu nhiên, không chọn một mẫu như SGD.

1. Tại sao ta nên dùng Ridge regression, hoặc các regularization models khác, hơn là dùng Linear regression hoặc Polinomial regression?

Vì nếu sử dụng linear regression thì sẽ dễ bị hiện tượng underfitting, do đó giải quyết bài toán sẽ không tốt. Nếu sử dụng polynomial thì dễ bị hiện tượng overfitting mà khắc phục bằng cách hạ bậc phức tạp của model thì sẽ bị hiện tượng underfitting. Do đó khi dùng polynomial thì nên dùng kèm với các regularization models vì polynomial bị overfitting thì các regularization models sẽ giảm dao động của model đang bị overfitting do đó giúp ngăn hiện tượng overfitting mô tả chính xác hơn về tập dữ liệu mà vẫn giữ nguyên bậc, độ phức tạp của model.