TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

BÀI TẬP LỚN GIỮA KÌ MÔN NHẬP MÔN HỌC MÁY

Người hướng dẫn: **GV.LÊ ANH CƯỜNG**
Người thực hiện: **NGUYỄN THANH TÚ – 52100349**
TRẦN THỊ VẸN –52100674
Lớp **: 21050301**
Khoá **: 25**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

BÀI TẬP LỚN GIỮA KÌ MÔN NHẬP MÔN HỌC MÁY

Người hướng dẫn: **GV.LÊ ANH CƯỜNG**
Người thực hiện: **NGUYỄN THANH TÚ – 52100349**
TRẦN THỊ VẸN –52100674
Lớp **: 21050301**
Khoá **: 25**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023

LỜI CẢM ƠN

Tôi xin gửi lời cảm ơn sâu sắc đến Thầy về việc hướng dẫn và giáo dục trong suốt môn học Máy học trong học kỳ vừa qua. Môn học không chỉ giúp tôi nắm vững các khái niệm cơ bản của Máy học mà còn mở ra cho tôi một thế giới mới về ứng dụng của nó trong thực tế.

Thầy đã tạo ra một môi trường học tập tích cực và truyền cảm hứng cho tôi để tiếp cận với những khía cạnh phức tạp và thú vị của Máy học. Những bài giảng thú vị cùng với ví dụ thực tế đã giúp tôi áp dụng kiến thức một cách hiệu quả trong thực tế.

Tôi cũng rất biết ơn về sự tận tâm và tâm huyết của Thầy trong việc hỗ trợ và giải đáp mọi thắc mắc của chúng tôi. Thầy đã không ngừng khích lệ và động viên chúng tôi để chúng tôi có thể vượt qua những thách thức và phát triển tốt hơn trong môn học này.

Môn học không chỉ giúp tôi nắm vững kiến thức mà còn giúp tôi phát triển kỹ năng phân tích và giải quyết vấn đề một cách chủ động. Tôi hy vọng có cơ hội tiếp tục học hỏi từ Thầy và áp dụng kiến thức đã học được vào thực tế công việc sau này.

Một lần nữa, xin chân thành cảm ơn Thầy về sự hỗ trợ và sự dạy dỗ tận tâm của Thầy.

Trân trọng,

ĐỒ ÁN ĐƯỢC HOÀN THÀNH
TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là sản phẩm đồ án của riêng tôi / chúng tôi và được sự hướng dẫn của GV. Lê Anh Cường;. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày 22 tháng 10 năm 2023

Tác giả

(ký tên và ghi rõ họ tên)

Nguyễn Thanh Tú

Trần Thị Vẹn

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

Phần xác nhận của GV hướng dẫn

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

Phần đánh giá của GV chấm bài

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

TÓM TẮT

Câu hỏi 3 : Tìm hiểu vấn đề Overfitting và các phương pháp giải quyết vấn đề này. Yêu cầu bao gồm mô tả lý thuyết và minh hoạ phương pháp trong lý thuyết bằng file code (.ipynb).

MỤC LỤC

LỜI CẢM ƠN i
ĐỒ ÁN ĐƯỢC HOÀN THÀNH ii
TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG ii
PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii
TÓM TẮT iv
MỤC LỤC 5
DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 6
CÁC KÝ HIỆU 6
CÁC CHỮ VIẾT TẮT 6
DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 7
DANH MỤC HÌNH 7
DANH MỤC BẢNG 7
CÂU 3 – TÌM HIỂU VỀ OVERFITTING VÀ PHƯƠNG PHÁP GIẢI QUYẾT VẤN ĐỀ 8
3.1 Vấn đề Overfitting: 8
3.2 Các phương pháp giải quyết overfitting: 10
3.2.1 Cross-Validation: 10
3.2.2. Regularization 11
3.2.3 Early stoping 13
3.2.4 Dropout 14
3.2.5 Thu thập thêm dữ liệu 14
TÀI LIỆU THAM KHẢO 16
PHỤ LỤC 17

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

CÁC KÝ HIỆU

CÁC CHỮ VIẾT TẮT

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

DANH MỤC HÌNH

DANH MỤC BẢNG

CÂU 3 – TÌM HIỂU VỀ OVERFITTING VÀ PHƯƠNG PHÁP GIẢI QUYẾT VẤN ĐỀ

3.1 Vấn đề Overfitting:

Trước khi tìm hiểu vấn đề Overfitting, ta có một số khái niệm cần nắm rõ:

- Bias: là sai số giữa giá trị dự đoán trung bình của mô hình và giá trị thực tế.

+ High bias: sai số lớn, mô hình đơn giản, tuy nhiên kết quả dự đoán chính xác không cao

+ Low bias: sai số nhỏ, mô hình phức tạp, kết quả dự đoán tốt

- Variance: là sai số thể hiện mức độ "nhạy cảm" của mô hình với những biến động trong dữ liệu huấn luyện

+ Low-variance: mô hình ít biến thiên theo sự thay đổi của dữ liệu huấn luyện

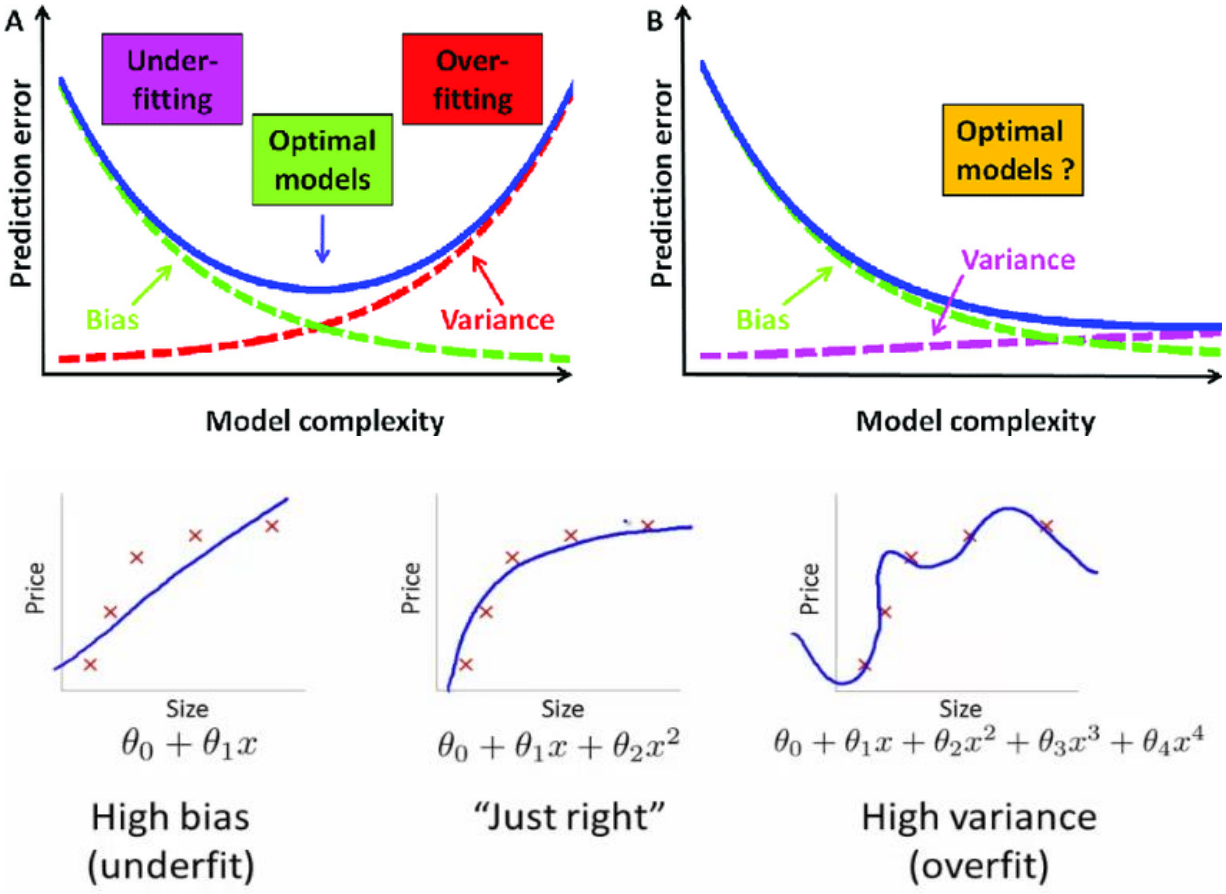
+ High-variance: mô hình biến thiên mạnh, bám sát theo sự thay đổi của dữ liệu huấn luyện. Mô hình có variance cao thường thể hiện rất tốt trên tập dữ liệu huấn luyện, nhưng không cho kết quả khả quan trên tập dữ liệu kiểm thử.

Khi kết hợp thì:

- Nếu có high-variance và low bias thì giá trị dự đoán của mô hình sẽ ở xung quanh hoặc gần với mục tiêu.
- Nếu có low-variance và low bias, các giá trị dự đoán của mô hình sẽ chính xác ngay tại chỗ
- Nếu có low-variance và high bias, các giá trị dự đoán sẽ bắt đầu chạm vào đâu đó chứ không phải mục tiêu. Tuy nhiên, các giá trị ở cùng một khu vực.
- Nếu có high-variance và high bias, các giá trị được dự đoán sẽ ở một nơi khác so với các giá trị được nhắm mục tiêu

Overfitting: là hiện tượng mà mô hình có low bias và high variance, lúc này mô hình trở nên phức tạp, bám sát theo dữ liệu huấn luyện. Điều này làm giản hiệu suất của mô hình khi áp dụng cho các tập dữ liệu mới.

Overfitting là hiện tượng mô hình (thuật toán) tìm được quá khớp với dữ liệu training (huấn luyện). Việc quá khớp này có thể dẫn đến việc dự đoán nhầm nhiễu, và chất lượng mô hình không còn tốt trên dữ liệu test (thực tế).



Hàm h được gọi là overfitting nếu như tồn tại một hàm g rằng [Mitchell, 1997]:

+ g nó tệ hơn dữ liệu huấn luyện

+ g tốt hơn h cho dữ liệu tương lai

Một thuật toán khác cho rằng, overfitting là làm việc rất tốt với dữ liệu huấn luyện nhưng làm việc rất tệ với dữ liệu dự đoán trong tương lai.

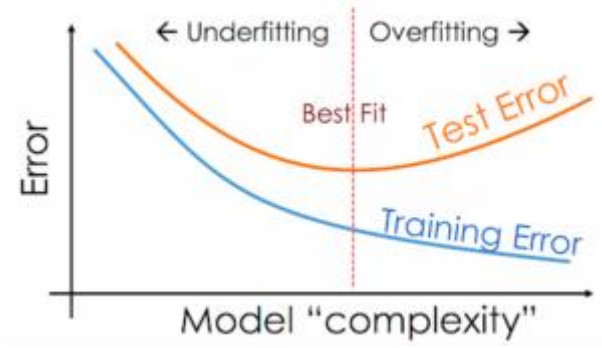
- **Quá khớp do nhiều nguyên nhân gây ra:**

+ Hàm hoặc mô hình được huấn luyện quá phức tạp hoặc có quá nhiều tham số.

+ Có nhiễu hoặc lỗi trong dữ liệu huấn luyện.

+ Kích thước dữ liệu huấn luyện quá nhỏ, không mô tả được toàn bộ không gian dữ liệu.

+ Độ phức tạp của mô hình học máy.



3.2 Các phương pháp giải quyết overfitting:

Tăng số lượng Training data

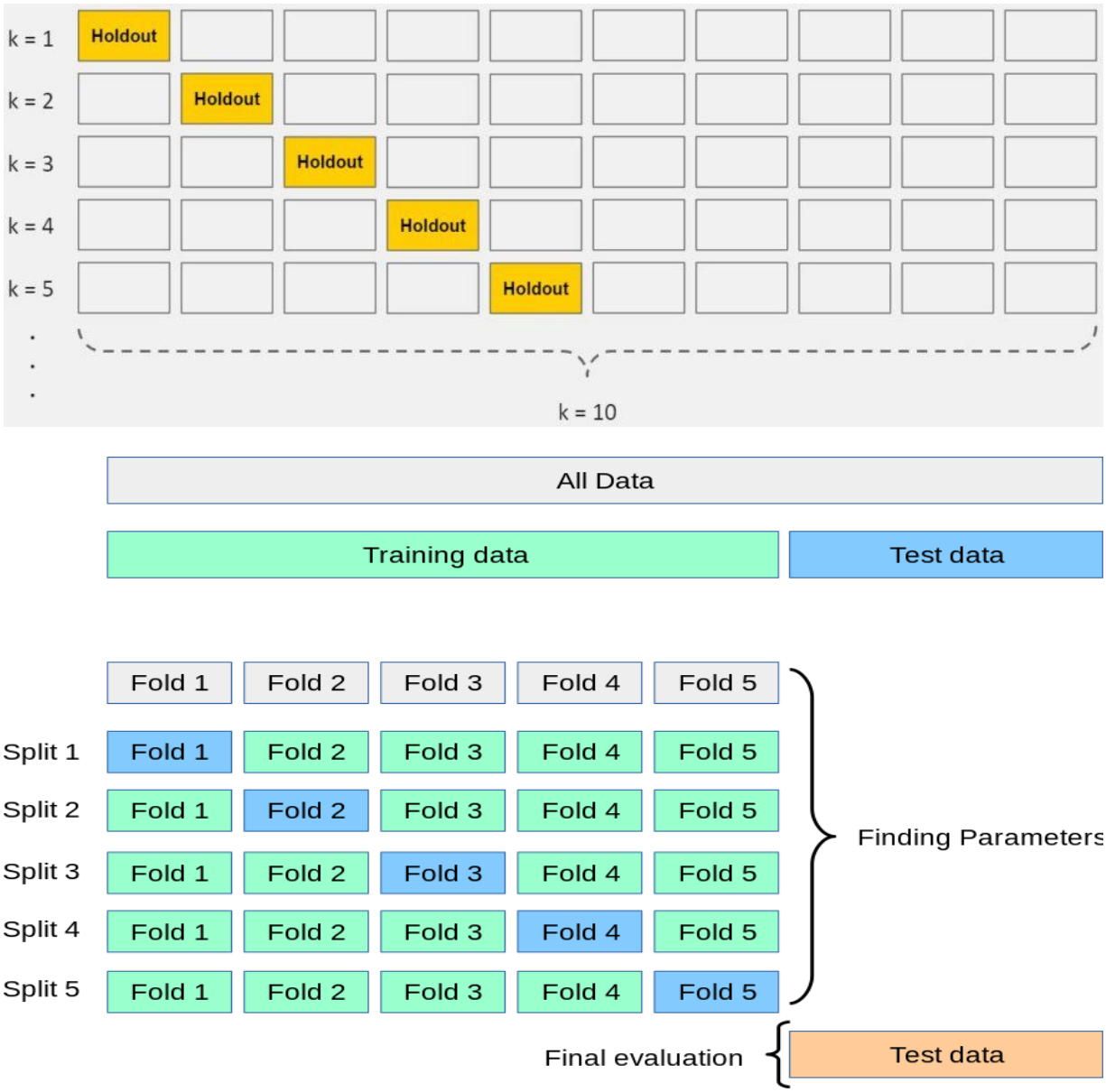
Giảm số lượng biến giải thích (feature)

Tăng độ lớn của parameter chuẩn hóa λ

3.2.1 Cross-Validation:

Ý tưởng của Cross-validation khá đơn giản: Chia tập dữ liệu huấn luyện thành nhiều phần để tạo nên các tập train-test nhỏ và sử dụng các tập train-set này để hiệu chỉnh các hyperparameter của mô hình.
Với kỹ thuật k-fold cross-validation, ta chia tập train ban đầu thành k phần tương đương nhau (k subnets), mỗi phần được gọi là folds. Sau đó, sử dụng k-1 folds để huấn luyện mô hình, fold còn lại (được gọi là holdout fold) được sử dụng như validate set. Mô hình cuối được xác định dựa trên trung bình của các train error và validation error. Cách làm này còn có tên gọi là k-fold cross validation.

Khi k bằng với số lượng phần tử trong tập training ban đầu, tức mỗi tập con có đúng 1 phần tử, ta gọi kỹ thuật này là leave-one-out.



3.2.2. Regularization

Một nhược điểm lớn của cross-validation là số lượng training runs tỉ lệ thuận với k. Điều đáng nói là mô hình polynomial như trên chỉ có một tham số cần xác định là bậc của đa thức. Trong các bài toán Machine Learning, lượng tham số cần xác định thường lớn hơn nhiều, và khoảng giá trị của mỗi tham số cũng rộng hơn nhiều, chưa kể đến việc có những tham số có thể là số thực. Như vậy, việc chỉ xây dựng một mô hình thôi cũng là đã rất phức tạp rồi. Có một cách giúp số mô hình cần huấn luyện giảm đi nhiều, thậm chí chỉ một mô hình. Cách này có tên gọi chung là regularization.

Regularization, một cách cơ bản, là thay đổi mô hình một chút để tránh overfitting trong khi vẫn giữ được tính tổng quát của nó (tính tổng quát là tính mô tả được nhiều dữ liệu, trong cả tập training và test). Một cách cụ thể hơn, ta sẽ tìm cách di chuyển nghiệm của bài toán tối ưu hàm mất mát tới một điểm gần nó. Hướng di chuyển sẽ là hướng làm cho mô hình ít phức tạp hơn mặc dù giá trị của hàm mất mát có tăng lên một chút.

Regularization là 1 kĩ thuật tránh overfitting bằng cách thêm vào hàm loss 1 đại lượng lamda. f(weight) => Tối ưu model (giảm hàm loss) => giảm weight => mô hình bớt phức tạp => tránh overfitting.

a. Lasso Regularization – L1 Regularization

L1 Regularization là một phương pháp phân tích hồi quy thực hiện cả lựa chọn biến và chính quy hóa. Hồi quy Lasso sử dụng ngưỡng mềm. Hồi quy Lasso chỉ chọn một tập hợp con của các hiệp biến được cung cấp để sử dụng trong mô hình cuối cùng:

$$\text{Cost} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{m} |w_i|$$

b. Ridge Regularization – L2 Regularization

Hồi quy Ridge là một kỹ thuật để phân tích dữ liệu hồi quy nhiều lần. Khi xảy ra đa cộng tuyến, các ước lượng bình phương nhỏ nhất là không chệch. Một mức độ chệch được thêm vào các ước tính hồi quy và kết quả là hồi quy sườn núi làm giảm các sai số tiêu chuẩn. Công thức cho hồi quy sườn núi là:

$$\text{Cost} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{m} w_i^2$$

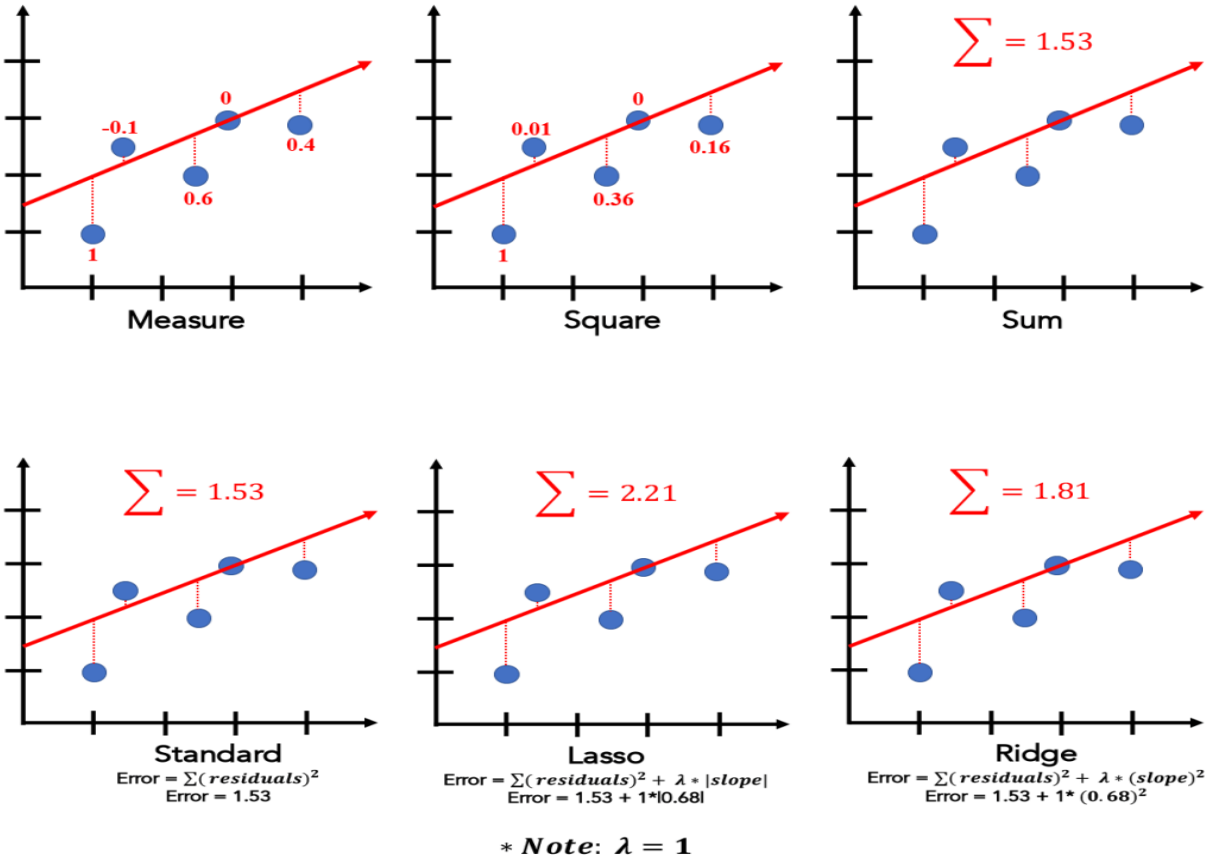
c. Elastic Net Regularization – L1 and L2 Regularization

Hồi quy ElasticNet là một phương pháp hồi quy chính quy kết hợp tuyến tính các hình phạt của phương pháp lasso và phương pháp ridge. Hồi quy ElasticNet được sử dụng để hỗ trợ máy vectơ, học số liệu và tối ưu hóa danh mục đầu tư. Hàm hình phạt được cho bởi:

$$\text{Cost} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \left((1-\alpha)\sum_{i=1}^{m} |w_i| + \alpha \sum_{i=1}^{m} w_i^2\right)$$

m – Number of Features
n – Number of Examples
y_i – Actual Target Value

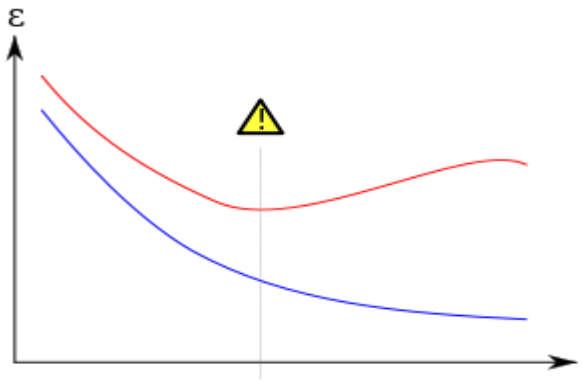
y_i(hat) – Predicted Target Value



3.2.3 Early stoping

Trong nhiều bài toán Machine Learning, chúng ta cần sử dụng các thuật toán lặp để tìm ra nghiệm, ví dụ như Gradient Descent. Nhìn chung, hàm mất mát giảm dần khi số vòng lặp tăng lên. Early stopping tức dừng thuật toán trước khi hàm mất mát đạt giá trị quá nhỏ, giúp tránh overfitting.

Một kỹ thuật thường được sử dụng là tách từ training set ra một tập validation set như trên. Sau một (hoặc một số, ví dụ 50) vòng lặp, ta tính cả train error và validation error, đến khi validation error có chiều hướng tăng lên thì dừng lại, và quay lại sử dụng mô hình tương ứng với điểm và validation error đạt giá trị nhỏ.



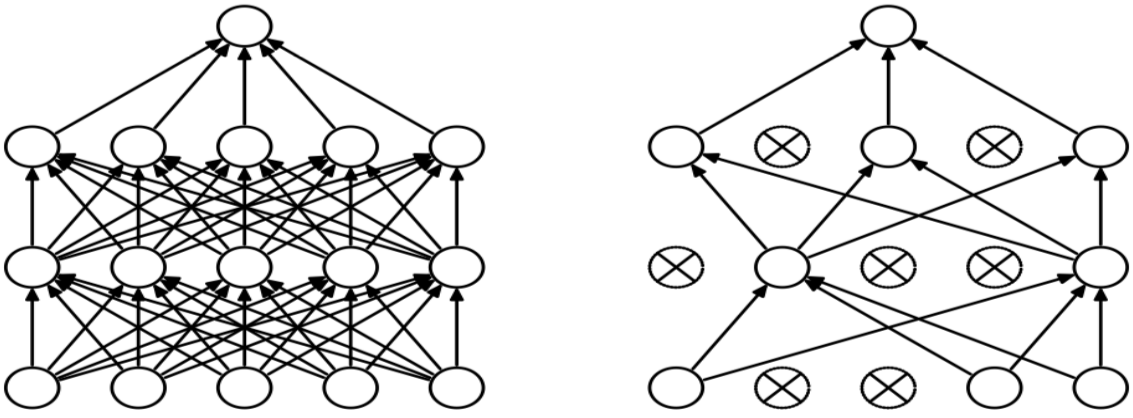
Early Stopping. Đường màu xanh là train error, đường màu đỏ là validation error. Trục x là số lượng vòng lặp, trục y là error. Mô hình được xác định tại vòng lặp mà validation error đạt giá trị nhỏ nhất.

3.2.4 Dropout

Một trong những nguyên nhân khiến model của bạn trở nên overfitting là: model của bạn quá sâu, phức tạp (chẳng hạn như nhiều layer, node) trong khi chỉ có chút xíu dữ liệu.

Ví dụ như bạn chỉ có <1 triệu nhưng bạn đòi mua nhà thì không có khả năng ....Cách giải quyết ở đây : một là tăng thêm dữ liệu như ở trên, 2 giảm độ phức tạp của model bằng cách bỏ đi 1 số layer, node.

Dropout là kĩ thuật giúp tránh overfitting cũng gần giống như regularization bằng cách bỏ đi random p% node của layer => giúp cho mô hình bớt phức tạp (p thuộc [0.2, 0.5]) .



(a) Standard Neural Net (b) After applying dropout.

3.2.5 Thu thập thêm dữ liệu

Dữ liệu ít là 1 trong trong những nguyên nhân khiến model bị overfitting. Vì vậy chúng ta cần tăng thêm dữ liệu để tăng độ đa dạng, phong phú của dữ liệu (tức là giảm variance).

Thu thập thêm dữ liệu: chúng ta phải craw thêm dữ liệu hay tới thực tiễn để thu thập, quay video, chụp ảnh,..Tuy nhiên trong nhiều trường hợp thì việc thu thập thêm dữ liệu là infeasible nên phương pháp này không được khuyến khích.

Data Augmentation: Augmentation là 1 phương thức tăng thêm dữ liệu từ dữ liệu có sẵn bằng cách rotation, flip, scale, skew ,... images. Phương pháp này được sử dụng rất phổ biến trong xử lý ảnh cho Deep learning.

Một số phương pháp data Augmentation:

+ Use tf.image

+ Use keras preprocessing layers

+ Use albumentation

+ Use openCV

Ngoài ra thì còn 1 số phương pháp cũng hay được sử dụng để giải quyết vấn đề overfitting:

Sử dụng mô hình ensemble: Kết hợp nhiều mô hình nhỏ thành một mô hình ensemble có thể giảm overfitting. Các mô hình ensemble thường tổng hợp kiến thức từ nhiều mô hình nhỏ để tạo ra một dự đoán tổng hợp.

Lựa chọn cẩn thận các tính năng (Feature Selection): Loại bỏ hoặc giảm số lượng các tính năng không quan trọng hoặc không cần thiết trong dữ liệu đào tạo. Điều này có thể giúp mô hình tập trung vào các thông tin quan trọng hơn và giảm overfitting.

Kiểm tra lại việc tiền xử lý dữ liệu: Đảm bảo rằng dữ liệu đã được tiền xử lý đúng cách, bao gồm chuẩn hóa, xử lý nhiễu và mã hóa dữ liệu…

TÀI LIỆU THAM KHẢO

Tiếng Việt

Tiếng Anh

PHỤ LỤC