CSDS503 / COMP552 – Advanced Machine Learning

Faizad Ullah

Bias and Variance

- **Is there a way to find when we have a high bias or a high variance?**
  - High Bias can be identified when we have
    - High training error
    - Validation error or test error is close to training error
  - High Variance can be identified when
    - Low training error
    - High validation error or high test-error

Bias and Variance

- **How do we fix high bias or high variance in the data set?**
- High bias is due to a simple model and we also see a high training error. To fix that we can do following things:
  - Add more input features
  - Add more complexity by introducing polynomial features
  - Decrease Regularization term
- High variance is due to a model that tries to fit most of the training dataset points and hence gets more complex. To resolve high variance issue we need to work on
  - Getting more training data
  - Reduce input features
  - Increase Regularization term

Solutions

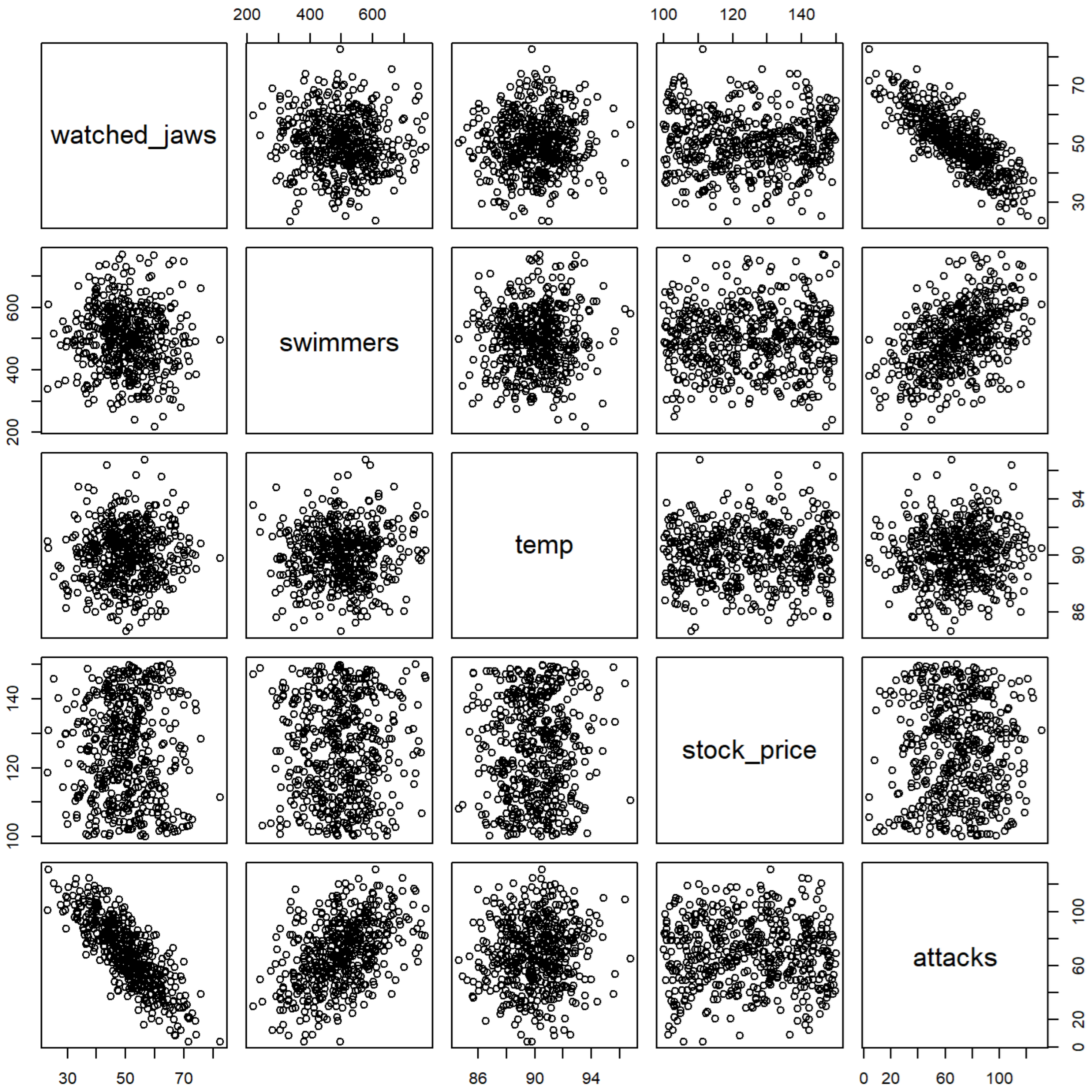
- Reduce the number of features
  - Manually select features
  - Model selection
- Regularization
  - Reduce magnitude/values of parameters θ_j.
  - Works well when we have a lot of features, each of which contributes a bit to the prediction.

Manual Feature Selection

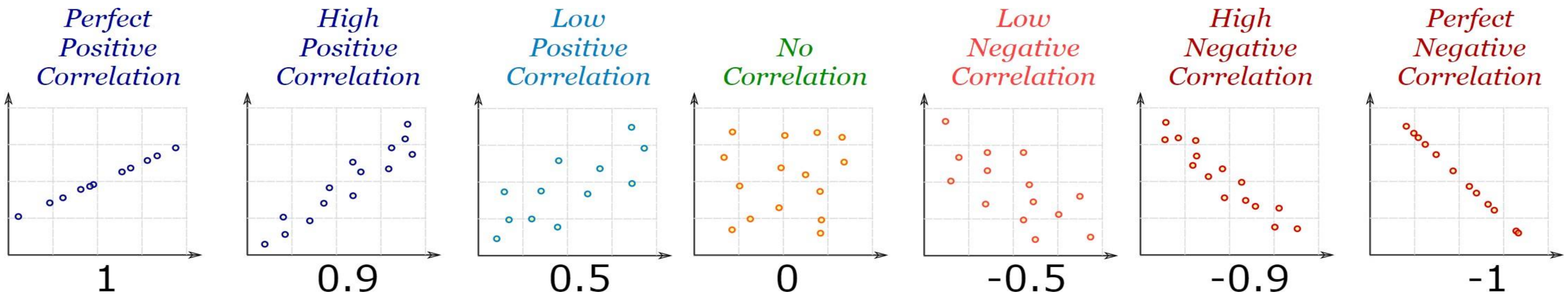
Manually select parameters

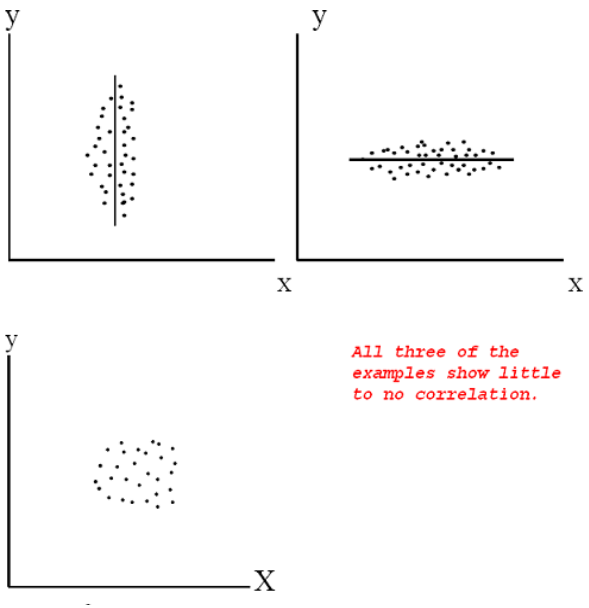
- Our data frame will consist of 1000 daily measurements of the following independent variables:
- **attacks:** Number of shark attacks (output variable)
- **swimmers:** Number of swimmers in water
- **watched_jaws:** Percentage of swimmers who watched iconic Jaws movies
- **temp:** Average temperature of the day
- **stock_price:** The price of your favorite tech stock that day (a totally unrelated variable)

Scatter Diagrams



Scatter Plots





Short-Term Goals

- Perfect models
- Training Accuracy
- Complex algorithms
- Large datasets
- …

Long-Term Goals

- We are interested in better long-term predictions.
- Find a balance between simplicity and complexity in our models.
- Prevent overfitting and to improve generalization.

Techniques

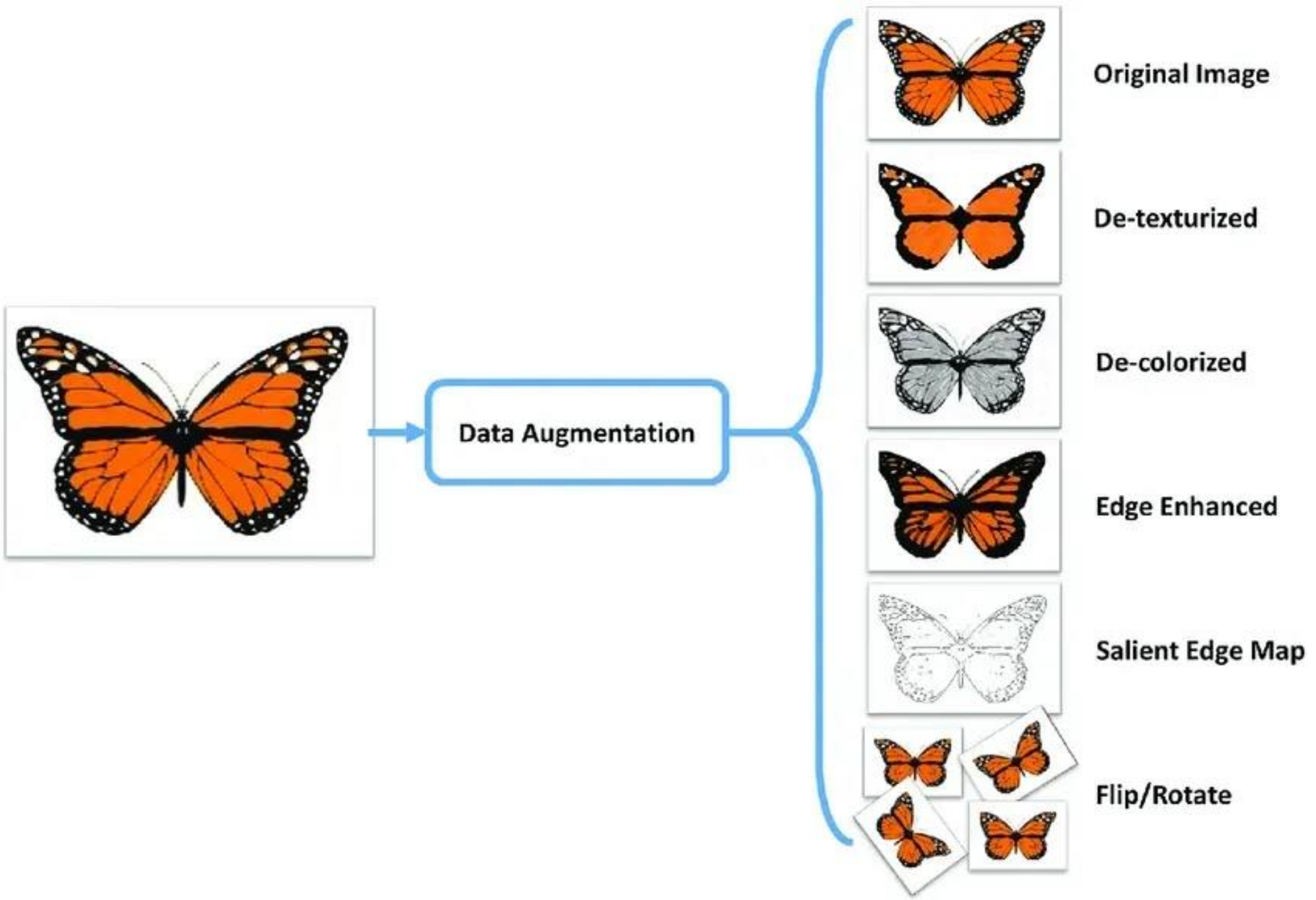
- Data Augmentation
- Early Stopping
- Dropout
- L2 regularization
- L1 regularization

Data Augmentation

Data Augmentation

- Creating new training examples by applying various transformations to the existing data
  - Rotation
  - Flipping
  - Scaling
  - Adding noise
- This increases the diversity of the training set and helps the model generalize better.

Data Augmentation - Images



Data Augmentation - Text

EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks

Jason Wei[1,2] **Kai Zou**[3]
[1]Protago Labs Research, Tysons Corner, Virginia, USA
[2]Department of Computer Science, Dartmouth College
[3]Department of Mathematics and Statistics, Georgetown University
jason.20@dartmouth.edu kz56@georgetown.edu

Data Augmentation - Text

1. SR: synonym replacement
2. RI: random insertion
3. RS: random swap
4. RD: random deletion

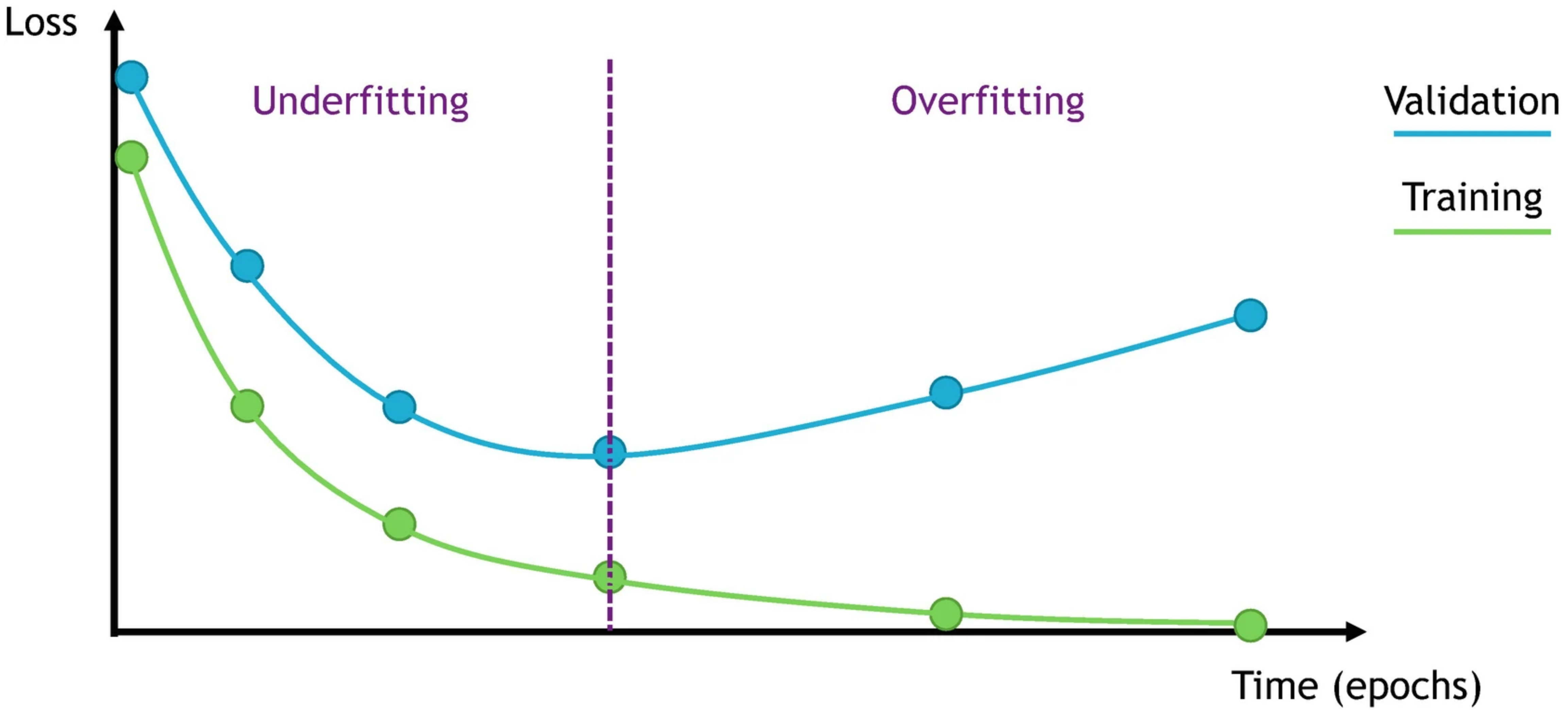
Operation	**Sentence**
None	A sad, superior human comedy played out on the back roads of life.
SR	A ***lamentable***, superior human comedy played out on the ***backward*** road of life.
RI	A sad, superior human comedy played out on ***funniness*** the back roads of life.
RS	A sad, superior human comedy played out on ***roads*** back ***the*** of life.
RD	A sad, superior human out on the roads of life.

Table 1: Sentences generated using EDA. SR: synonym replacement. RI: random insertion. RS: random swap. RD: random deletion.

Early Stopping

Early Stopping

- Early stopping is a regularization technique that monitors the model's performance on a validation set during training.
- When the validation performance stops improving, training is halted to prevent overfitting.

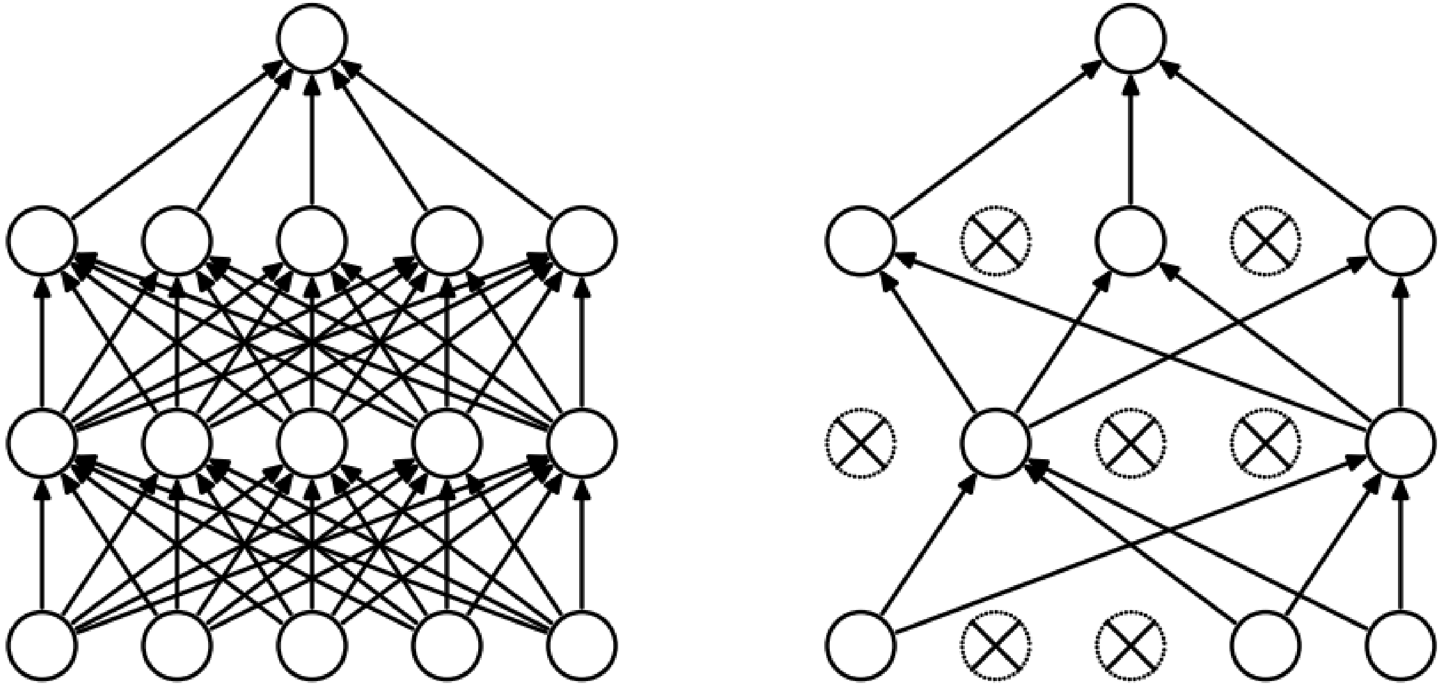

Loss
Underfitting
Overfitting
Validation
Training
Time (epochs)

Dropout

Dropout

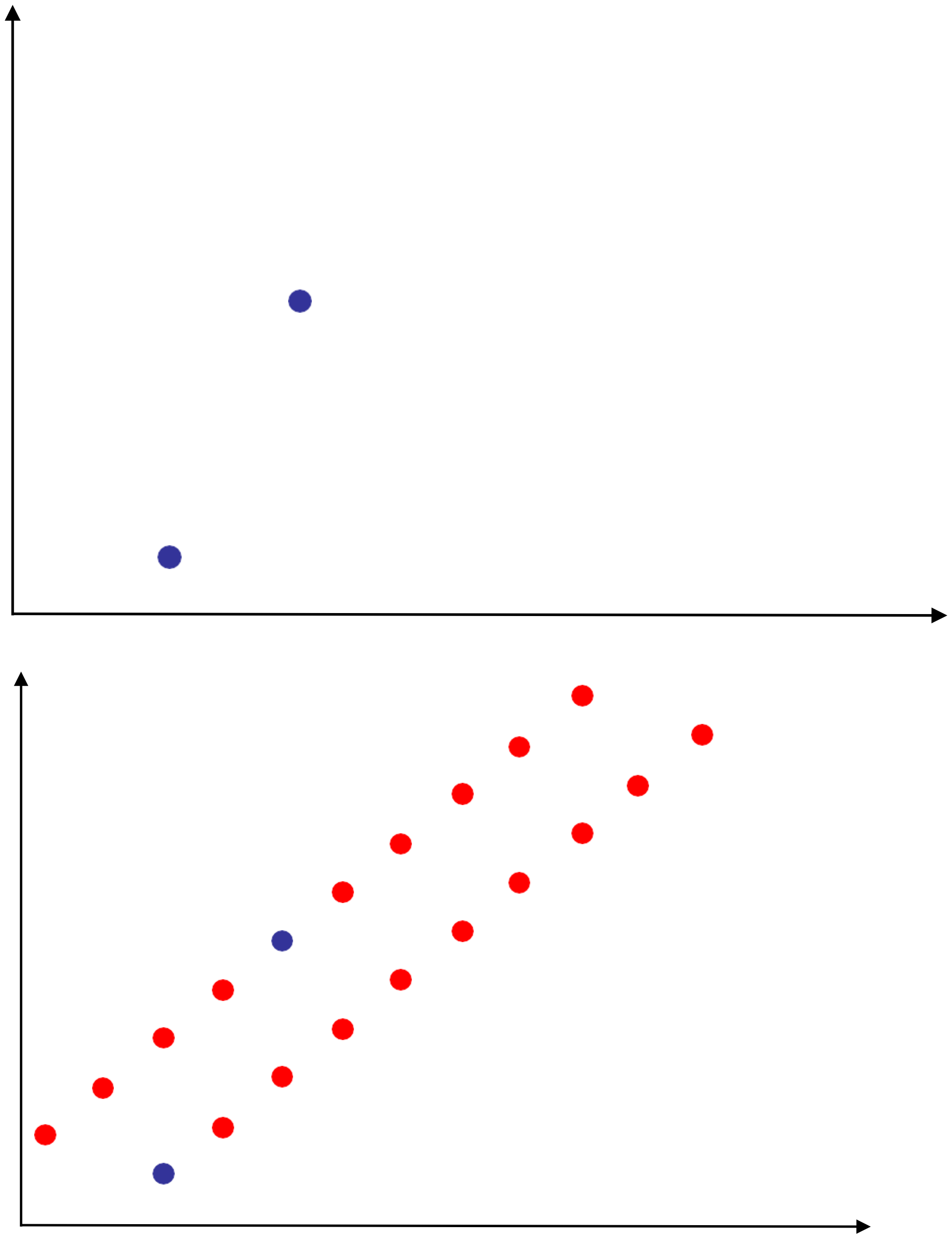
- Dropout is a regularization technique used exclusively in neural networks.
- During training, dropout randomly deactivates a fraction of neurons (typically 20-50%) in each layer by setting their outputs to zero.

Dropout



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

Regularization



$$h_\Theta(X) = \Theta^T X = \theta_0 + \theta_1 x_1$$

$$J(\Theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$\min_{\theta} \frac{1}{2m}\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)^2$$

The goal is to update theta_1

What if theta_1 is negative?

Regularization

$$y = mx + b$$

- Add penalty…

- How severe the penalty is?

Regularization

$$h_\Theta(X) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

$$\min_\theta J(\theta) = \min_\theta \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$\min_\theta J(\theta) = \min_\theta \frac{1}{2m} \left[\sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \right]$$

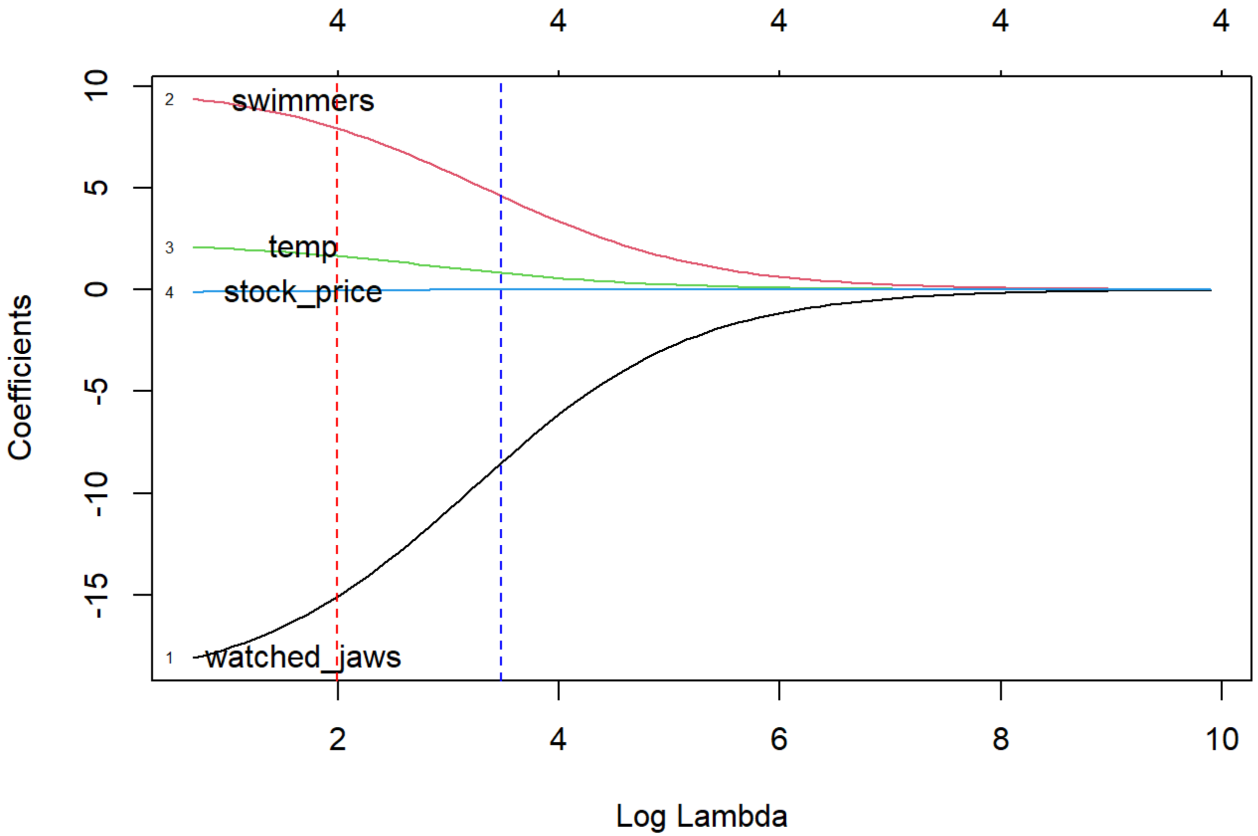
$$\min_\theta J(\theta) = \min_\theta \frac{1}{2m} \left[\sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} \left|\theta_j\right| \right]$$

How to find the optimal value of λ?

L2 Regularization or Ridge Regression

$$\min_{\theta} J(\theta) = \min_{\theta} \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} \theta_j^2\right]$$

- L2 regularization forces the weights to be small but does not make them zero
- L2 is not robust to outliers as square terms blows up the error differences of the outliers and the regularization term tries to fix it by penalizing the weights
- Ridge regression performs better when all the input features influence the output and all with weights are of roughly equal size

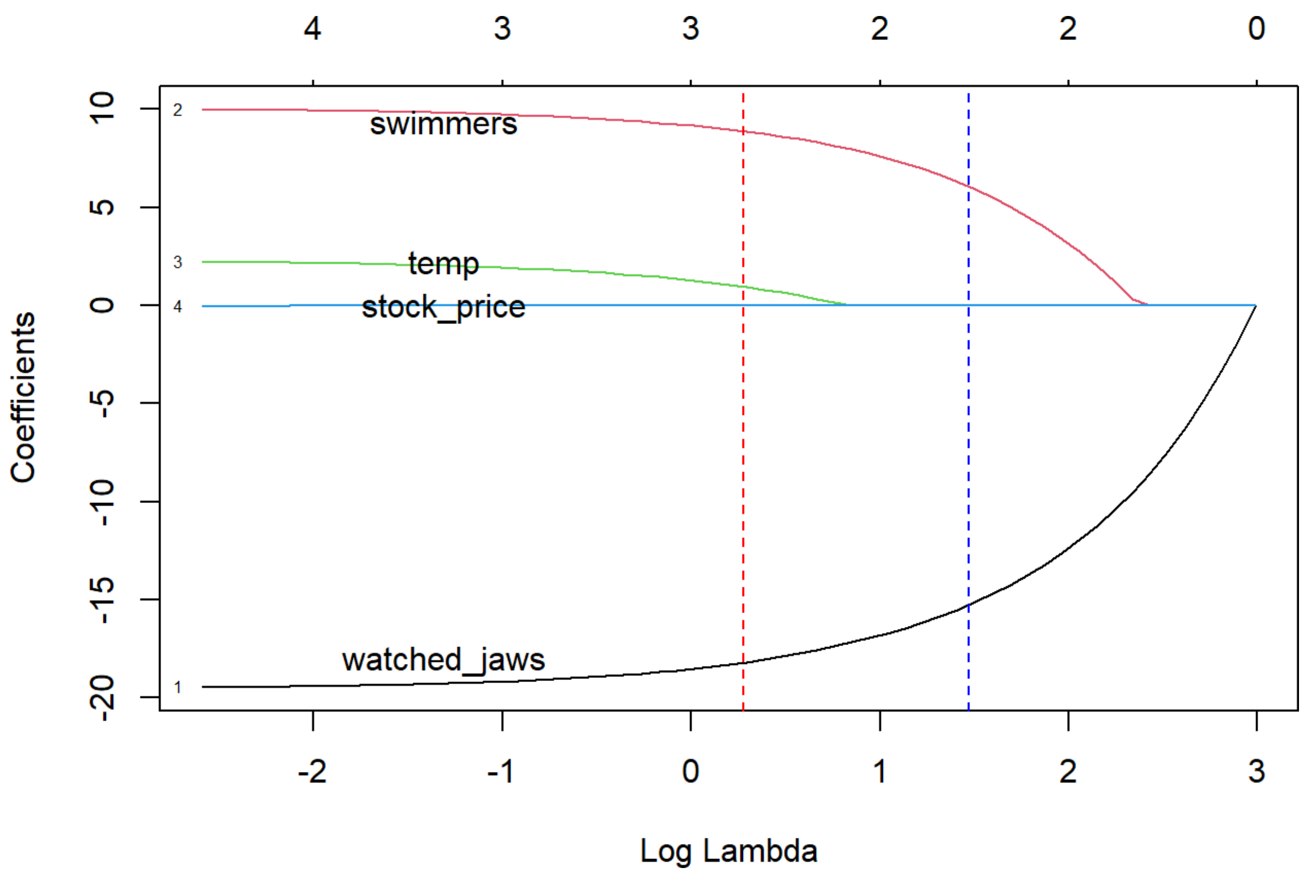


4
4
4
4
4
10
5
0
-5
-10
-15
2
swimmers
3
temp
4
stock_price
1
watched_jaws
Coefficients
2
4
6
8
10
Log Lambda

L1 Regularization or Lasso Regression

$$\min_{\theta} J(\theta) = \min_{\theta} \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n}\left|\theta_j\right|\right]$$

- L1 norm shrinks the parameters to zero.
- Not all input features have the same influence on the prediction. L1 norm will assign a zero weight to features with less predictive power.
- L1 regularization does feature selection. It does this by assigning insignificant input features with zero weight and useful features with a non-zero weight.



4
3
3
2
2
0
swimmers
temp
stock_price
watched_jaws
Coefficients
Log Lambda
10
5
0
-5
-10
-15
-20
-2
-1
0
1
2
3

Elastic net regularization

- Elastic net regularization is a combination of both L1 and L2 regularization

$$\min_{\theta} J(\theta) = \min_{\theta} \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_{\theta}(x^{(i)}) - y^{(i)}\right)^2 + \lambda_1 \sum_{j=1}^{n}|\theta_j| + \lambda_2 \sum_{j=1}^{n}\theta_j^2\right]$$

Sources

- Machine Learning, Andrew Ng, on Coursera by Stanford – a https://www.coursera.org/learn/machine-learning
- Deep Learning Specialization, Andrew Ng, on Coursera by deeplearning.ai
  - – https://www.coursera.org/specializations/deep-learning
- STATQUEST!!! An epic journey through statistics and machine learning, Josh Starmer, https://statquest.org/, https://www.youtube.com/channel/UCtYLUTtgS3k1Fg4y5tAhLbw
- https://towardsdatascience.com/cohens-kappa-9786ceceab58#:~:text=Cohen's%20kappa%20measures%20the%20agreement,raters%20may%20agree%20by%20chance.
- https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229
- https://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers#:~:text=Krippendorff's%20alpha%20(%CE%B1)%20is%20a,assign%20computable%2_0values%20to%20them.
- https://thenewstack.io/cohens-kappa-what-it-is-when-to-use-it-and-how-to-avoid-its-pitfalls/
- https://en.wikipedia.org/wiki/Cohen%27s_kappa