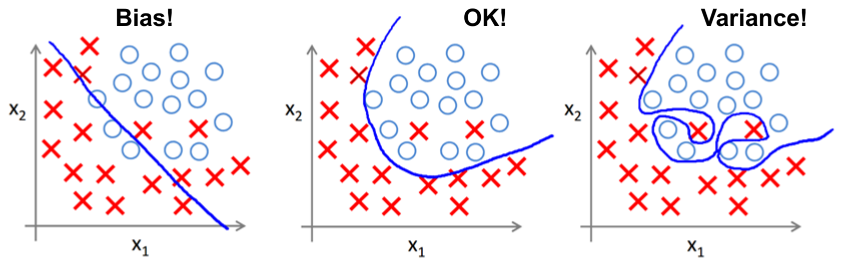
Imperial College London

Under/Over-fitting may also apply to Classification



$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$
(g = sigmoid function)

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2$
$+\theta_3 x_1^2 + \theta_4 x_2^2$
$+\theta_5 x_1 x_2)$

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2$
$+\theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2$
$+\theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \ldots)$

Source: Machne Learning Course, Andrew Ng

1

Imperial College London

Why the terms « Bias » and « Variance »?

Bias: *the hypothesis function $h_\theta(x)$ has a « pre-conception » or an « a priori » idea of the data variations which is too simple, which is biased from the start, considering the actual variability of the data (for instance it is a polynomial of too low degree).*

Variance: *the hypothesis function $h_\theta(x)$ has too many degrees of freedom – or parameters - and, as a result, can fit too many possible functions, with too much variance, considering the actual variability of the data. (for instance it is a polynomial of too high degree)*

2

Imperial College London

Underfitting and Overfitting

A good Machine Learning algorithm must:

1. Make the Training Error small. If this is not the case, we have Underfitting.
2. Make the gap between Training and Test – or Generalization - Error small. If this is not the case we have Overfitting.

Goodfellow et al, 2017

Imperial College London

One Way to Avoid Overfitting: Regularization

Regularization is any modification we make to a learning algorithm that is intended to reduce its Generalization Error but not its Training Error…

An effective regularizer is one that reduces Variance significantly while not increasing the Bias.

Goodfellow et al, 2017

Imperial College London

(L1 or L2) Regularization for Regression ML

L1 and L2 Regularizations consist of adding a new term to the objective function in order to control the variations of the parameters:

$$J(\theta) = \frac{1}{2m}\left(\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} \theta_j^2\right)$$ if L2 norm is used

Regularization Parameter, controlling the "Weight Decay"

$$J(\theta) = \frac{1}{2m}\left(\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} |\theta_i|\right)$$ if L1 norm is used

5

Imperial College London

Gradient Descent with Regularized Regression

Consider in more detail the Gradient Descent term in case of Regularization:

$$\theta_j := \theta_j\left(1 - \alpha\frac{\lambda}{m}\right) - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x_j^{(i)}$$

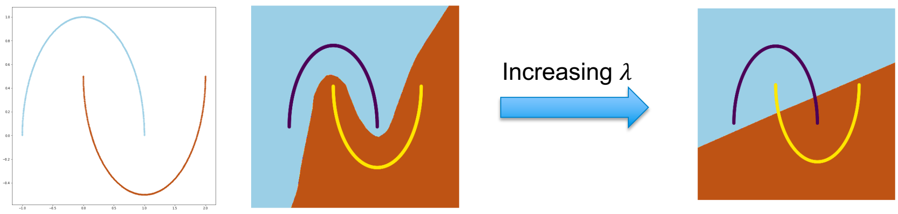
Systematic decrease of absolute value θ_j at each iteration

Same term as for optimization without regularization

6

Imperial College London

A Simple Neural Network Regularization Example (2)



$\lambda = 0$ $\lambda = 10$

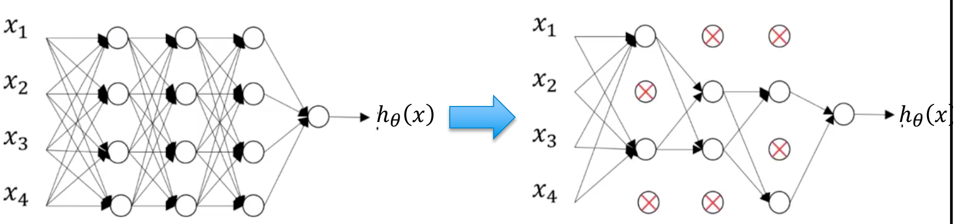
As λ increases, the values of the weights tend to zero, and the Decision Boundary first becomes a broken Line, then a single Line.

7

Imperial College London

Another Regularization Approach: Drop-Out (1)

At Training time:



For each Training example… Drop <u>hidden layers</u> units with 0.5 (or p) probability

Same approach for input layer but with p closer to 1. No change in output layer.

We are averaging the results over a set of network configurations!

https://www.youtube.com/watch?v=kAwF--GJ-ek

8

Imperial College London

Software Tools for Data Augmentation

A number of basic geometrical image transformations are likely to change the appearance of the image, but not its class:

- Rotation (to some reasonable degree)
- Translation (up, down, left, right)
- Zooming
- Cropping
- Added noise
- Changing the Brightness level

…

These are readily available in Python code (*imgaug, Albumentation* packages…).

Imperial College London

Batch Normalization on $z^{(l)}$ vector (could be on $a^{(l)}$)

1. Normalize $z^{(l)}$ vector at layer (l):

$$z_{norm}^{(l)} = \frac{z^{(l)} - \mu_z}{\sqrt{\sigma_z^2 + \varepsilon^2}}$$

(with μ_z and σ_z^2 mean and variance of $z^{(l)}$ calculated separately <u>on each mini-batch</u>, and ε small parameter useful if σ_z^2 = 0)

2. Apply Linear Transform to $z_{norm}^{(l)}$, with <u>trainable</u> parameters $\beta^{(l)}$ and $\gamma^{(l)}$. For each component j of $z_{norm}^{(l)}$

$$\tilde{z}_j^{(l)} = \gamma_j^{(l)} \, z_{j\,norm}^{(l)} + \beta_j^{(l)}$$

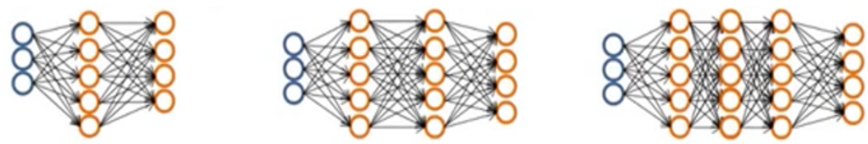
Optimize mean and variance of $\tilde{z}^{(l)}$

3. Use $\tilde{z}^{(l)}$ as input to the activation function: $a^{(l)} = g\left(\tilde{z}^{(l)}\right)$

Imperial College London

Change of Network Architecture

From one to three Hidden Layers



How can we choose the "Best" Network Architecture?

Machine Learning Diagnostic:
A test to gain insight about the performance of a Training algorithm.

Imperial College London

Use Validation Set to Optimize Hyperparameters

For each Possible Choice of Neural Network Hyperparameters:

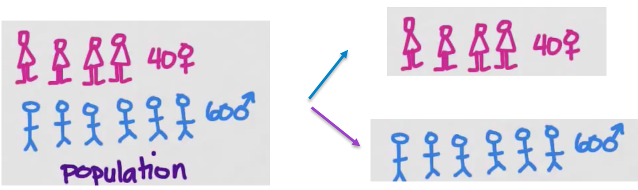
1. Train Neural Network Parameters on Training Set
2. Test Performance on the Validation Set
3. Pick the Hyperparameters that give the best performance on the Validation Set.
4. Possibly retrain the new Neural Network on Training+Validation Set.
5. Test the Neural Network on the Test Set

The Test Set is the final measure of performance but must never be used in the Training!

Imperial College London

Stratified Sampling: *Create Validation Set of 20%*

Sort Data into two "Strata"



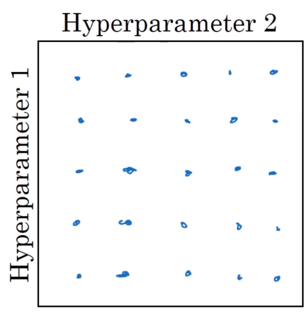
Randomly extract 20%, or 8

Randomly extract 20% or 12

First sort the population into "strata", then sample from each strata!

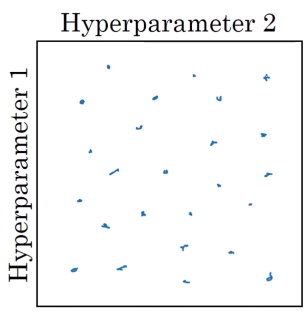
Imperial College London

Grid Search to Optimize Multiple Hyperparameters



Systematic Grid Search

Map the Validation Set Accuracy for each Hyperparameter pair.



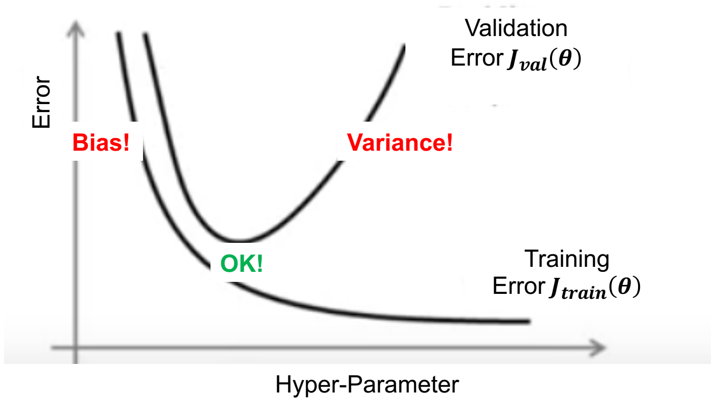
Random Grid Search

If one parameter is less sensitive than the other, the random search provides a richer exploration of the values of the sensitive parameter.

https://www.youtube.com/watch?v=AXDByU3D1hA

Imperial College London

How to Identify Bias vs Variance Problems



15

Imperial College London

General Guidelines for Improving Training

To address Bias problems

Try using more input features in order to increase the number of parameters
Try decreasing the regularization

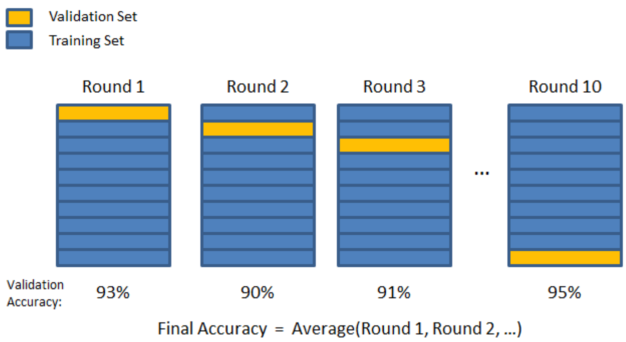
To address Variance problems

Try decreasing the number of features
Try increasing the regularization

16

Imperial College
London

What is k-Fold Validation?



https://towardsdatascience.com/cross-validation-70289113a072