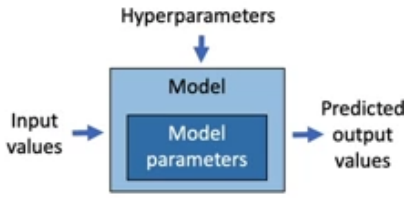
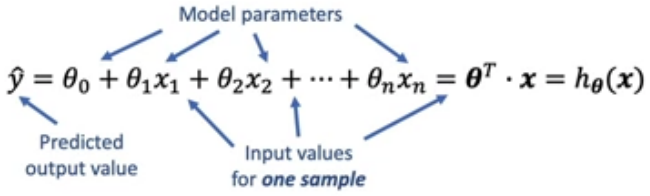
1. What are the components of a model? How do they complement each other in the processing of model prediction?

Model

- A mapping of input values to predicted outcome values
- Flexible due to model parameters
- Constrained by fixed hyper-parameters



- Example: Linear model



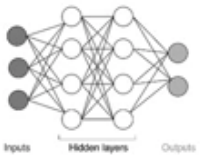


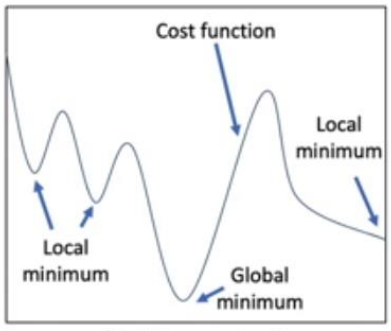
Image: Chartrand et al. RadioGraphics 2017

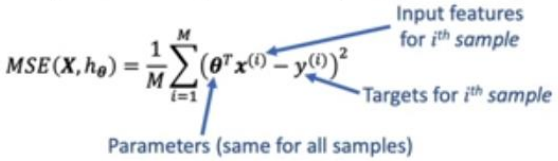
2. What is the purpose of a cost function? Please explain the global minimum and local minimum.

Cost Function = Error = Loss Function

- Measures errors or differences between predicted and target values
- Want to minimise it for training data

- Global minimum: smallest value overall
- Local minimum: smallest value in some region
- Example: Mean square error (MSE)

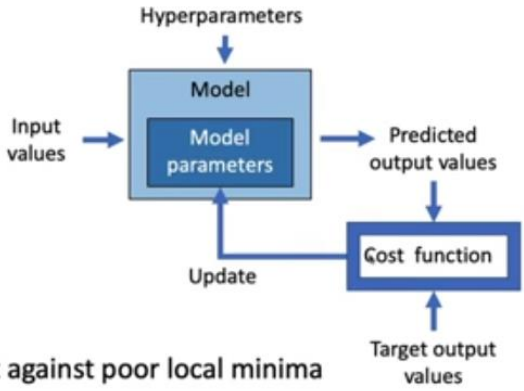




3. What is the purpose of model training? Please also briefly introduce the training process.

Training = Fitting = Optimisation

- Minimise cost function by adjusting model parameters
- Start with initial guess of model parameters
- Iteratively change model parameters & evaluate cost function
- Ideal algorithm: fast, but robust against poor local minima



4. Please briefly explain how the SGD works. What is learning rate?

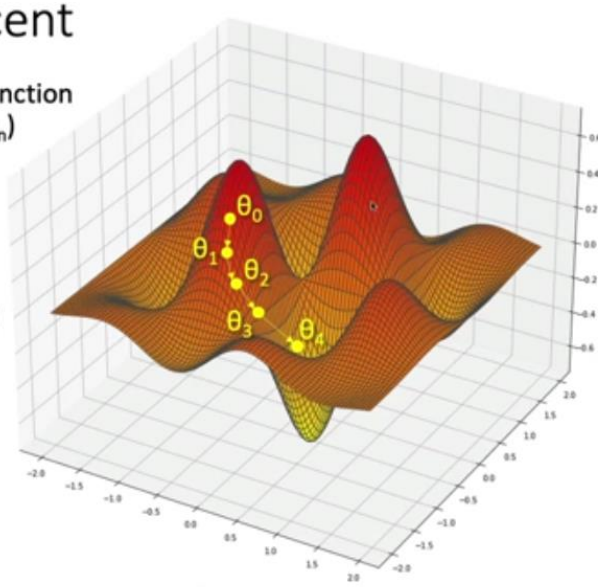
Gradient Descent

- Partial derivatives of cost function = Local gradient = $\nabla_\theta \text{MSE}(\theta_n)$
- Iteratively step downhill (negative gradient)

$$\theta_{n+1} = \theta_n - \eta \nabla_\theta \text{MSE}(\theta_n)$$

Learning rate "eta" η = sets size of step

This is an important hyper-parameter!



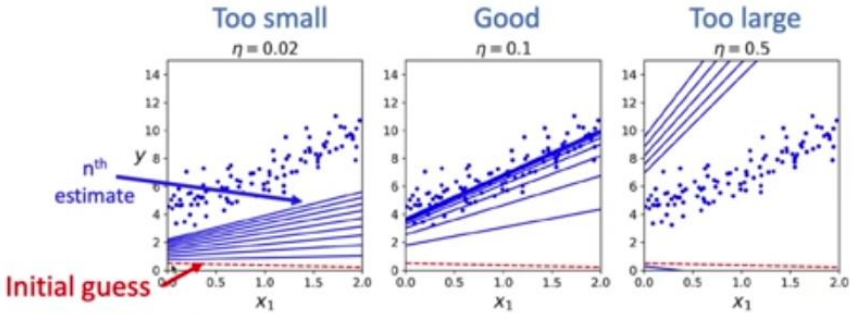
CS3317 Using Machine Learning Tools 6

5. What will happen if you have a too small or too large learning rate?

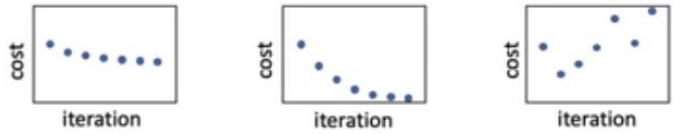
Images: Geron, Hands On

Learning rate η

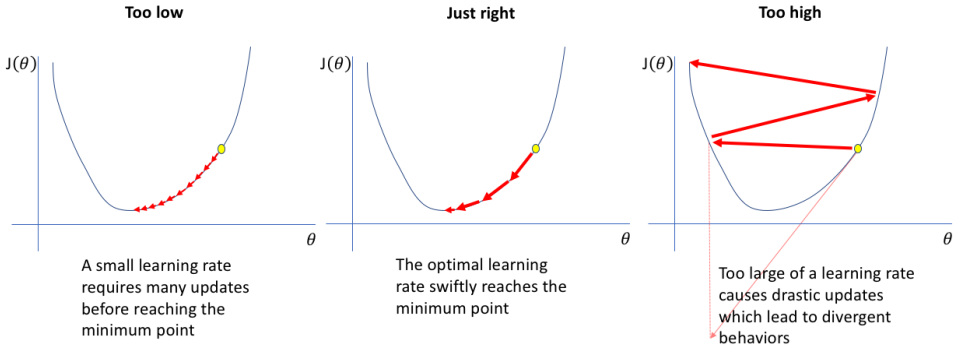
Example: trying to fit line to points



- Small η values take a long time to change, but go in the right direction
- Large η values overstep and can easily become unstable
- Can monitor the behaviour of the cost function values over iterations to spot these



8

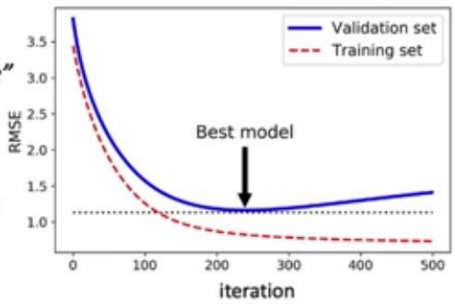


6. What are the stopping criteria and final result?

Stopping Criteria and Final Result

Stop if:

- No further improvement
  e.g. 5 iterations in a row show *"no change"*
  - `early_stopping`: turn on/off
  - `n_iter_no_change`: number of iterations
  - `tol`: if cost difference between steps is less than ε (tolerance) then treat it as *no change*
- Maximum number of iterations reached
  - `max_iter`: maximum number of iterations



- Final result: best model across the training process
  - this might *not* be the final model

Images: Geron, Hands On ML

7. What is the difference between SGD and original GD?

Stochastic Gradient Descent

- Pick one random sample
- Calculate the cost function gradient only from that sample

Better algorithm:

- Shuffle instances of the training set
- Use one instance after the other
- Adjust the learning rate η
- Reshuffle and repeat

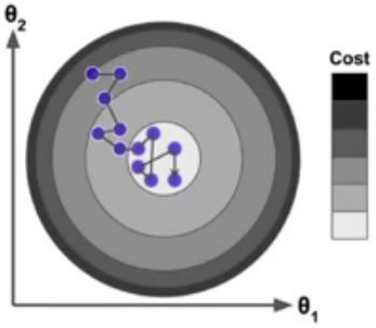


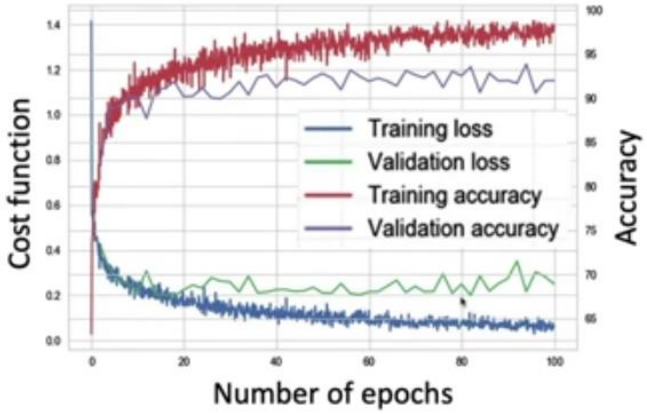
Image: Geron, Hands On ML

Pros: Fast, low memory, randomisation can help escape local minimum

Cons: Very noisy & no guarantee that minimum is reached

8. What does the below image tell us? What is epoch?

Learning Curve = Training Curve



- Epoch = one pass through whole dataset ≅ iteration
- Shows both training ***and*** validation performance
  - allows both underfitting and overfitting to be seen

Image: Chartrand et al. RadioGraphics 2017

9. What are the sources of generalization errors? Please explain them in detail.

Sources of Generalisation Error

- Variance
  - *Irreducible error*
    - Due to randomness in the data itself
  - *Overfitting* leads to over-sensitivity to small variations in the data
    - Too many model parameters

- Bias or systematic error
  - *Sub-optimal model choice or hyperparameter choice*
    - Especially underfitting
  - *Representativeness of data*
    - Lack of data coverage (model extrapolations are usually bad)
    - Bias in the data (e.g. due to limitations or bias in sampling)
    - Imbalances in the data (e.g. due to nature of problem)
      - disease vs healthy ; suspicious vs normal transactions

10. What is regularization? What is the purpose of it?

Regularisation

- Add a term to the cost function that tries to prevent overfitting
  - usually controlled by an adjustable weight α

$$\text{Cost} = \text{data\_term} + \alpha * \text{regularisation\_term}$$

- Purpose
  - Prevent overfitting by penalising large parameter values or lack of smoothness in outputs
  - Add a-priori knowledge (desired properties) to an underdetermined problem
- A form of multi-objective optimisation

11. What is L2 regularization? Why do we need it?

L2 (Ridge/Tikhonov) Regularisation

- Effect: Keep model parameters small

Cost | Data term | Regularisation term

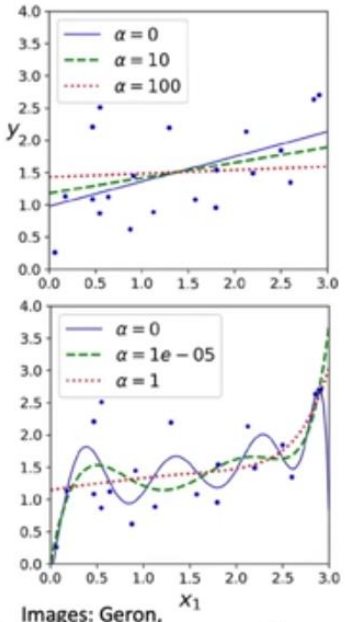
$$J(\boldsymbol{\theta}) = \mathrm{MSE}(\boldsymbol{\theta}) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2$$

L2 Norm

Regularisation strength α

Model parameters, in this case except θ_0

- Scaling of data important for setting α
- Scikit learn: `penalty` parameter "`l2`"



Images: Geron, Hands On ML

12. Similarly, what is L1 regularization? Why do we need it?

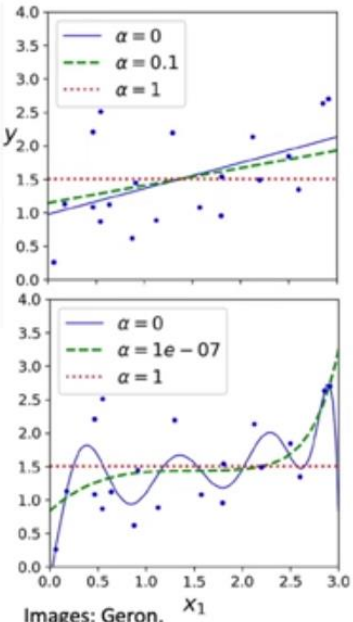
L1 (Lasso) Regularisation

- LASSO = Least Absolute Shrinkage and Selection Operator
- Effect: Keep model parameters small
- Also tends to eliminate least important features (i.e. sets some $\theta_i = 0$)

$$J(\boldsymbol{\theta}) = \mathrm{MSE}(\boldsymbol{\theta}) + \alpha \sum_{i=1}^{n} |\theta_i|$$

L1 Norm

- Scaling of data important for setting α
- Scikit learn: `penalty` parameter "`l1`"
- Not differentiable at 0, but most optimisers can cope with this



Images: Geron, Hands On ML