Good Practice for Neural Networks

Ing. Zese Riccardo
riccardo.zese@unife.it

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Regione Emilia-Romagna

Università degli Studi di Ferrara

Outline

- Introduction to Python
- Introduction to Neural Networks
- Convolutional NN
- Recurrent NN
- Autoencoders and self supervised learning

Sources

- These slides are taken from:
  - Intel Nervana AI Academy Intructional Content

  - Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

    https://www.deeplearningbook.org/lecture_slides.html
  - Chapter "Neural Networks" of "A Course in Machine Learning" by Hal Daume III.

    http://ciml.info/
  - Some parts from "CS231n: Convolutional Neural Networks for Visual Recognition", Stanford University

    http://cs231n.stanford.edu/

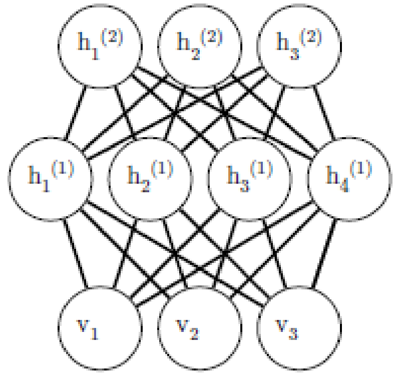
Outline

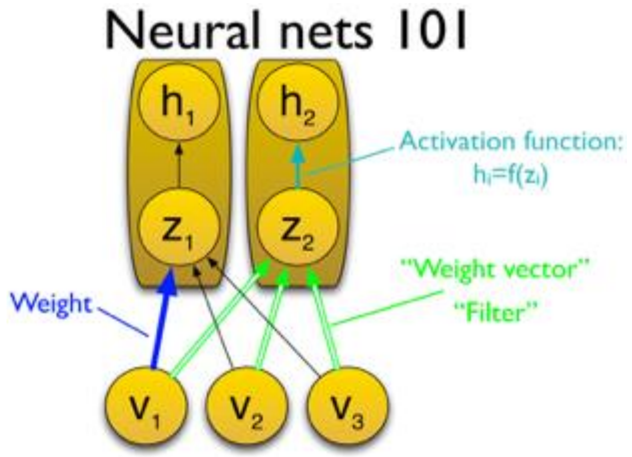
- Introduction to Python
- Introduction to Neural Networks
- Convolutional NN
- Recurrent NN
- Autoencoders and self supervised learning

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

What drives success in ML?

- Arcane knowledge of dozens of obscure algorithms?
- Mountains of data?
- Knowing how to apply 3-4 standard techniques?





Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Good practices

- Decide from the beginning the objective, error metrics to use, and the target values for this error metric.
  - Usually they are all driven by the problem that the application is intended to solve → you must analyze the scenario you are working on.
- Establish a working end-to-end pipeline as soon as possible, including the estimation of the appropriate performance metrics.
  - This would help determine possible bottlenecks in performance, what might cause overfitting, underfitting, defects in the data or software.
- Repeatedly make incremental changes such as gathering new data, adjusting hyperparameters, or changing algorithms, based on specific findings from your analysis.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Identify Needs

- High accuracy or low accuracy?
  - Surgery robot, self driving cars: high accuracy
  - Celebrity look-a-like app: low accuracy
- Which metrics?
  - Accuracy? (% of examples correct)

    → to detect rare events it is not really a good choice.
  - Coverage? (% of examples processed)

    → when we do not want to make a decision but estimate how confident we should be about a decision, especially if a wrong decision can be harmful and if a human operator is able to occasionally take over.
  - Precision? (% of detections that are right)
  - Recall? (% of objects detected)

    → when using precision and recall, it is common to plot a **PR curve**, with precision on the y-axis and recall on the x-axis, or compute the **F-score**
  - Amount of error? (For regression problems)

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

End-to-end System

- Establish a working end-to-end pipeline as soon as possible → build the simplest viable system first
- What baseline to start with though?
  - Copy state-of-the-art from related publication
- Ask yourself the following question:
  - Deep or not?
  - Which architecture?
  - Are there already defined solution for your problem?

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Deep or Not?

- Depending on the complexity of your problem, you may even want to begin without using deep learning.
  - If your problem has a chance of being solved by just choosing a few linear weights correctly, you may want to begin with a simple statistical model like logistic regression.
- Lots of noise, little structure -> not deep
- Little noise, complex structure -> deep
- Good shallow baseline:
  - *Use what you know*
  - Logistic regression, SVM, boosted tree are all good

Regione Emilia-Romagna

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

FERRARIAE UNIVERSITAS · EX LABORE FRUCTUS · 13 91

Università degli Studi di Ferrara

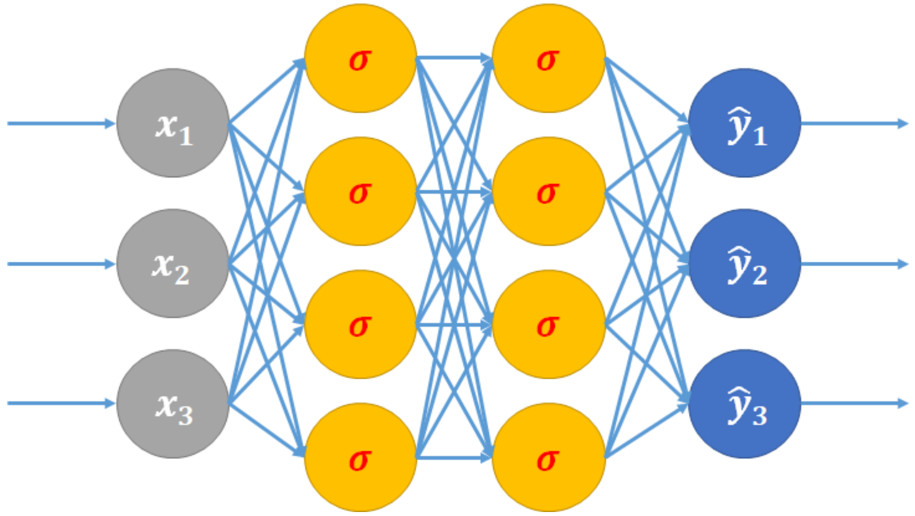
Which architecture?

- Depends on the data structure:
  - No structure → **fully connected** (feedforward network with fully connected layers)
  - Spatial structure (topological structure, e.g., images) → **convolutional** (probably beginning with linear units s.a. ReLU)
  - Sequential structure (either input or output) → **recurrent**

- A reasonable choice of optimization algorithm is SGD with momentum with a decaying learning rate

No structure

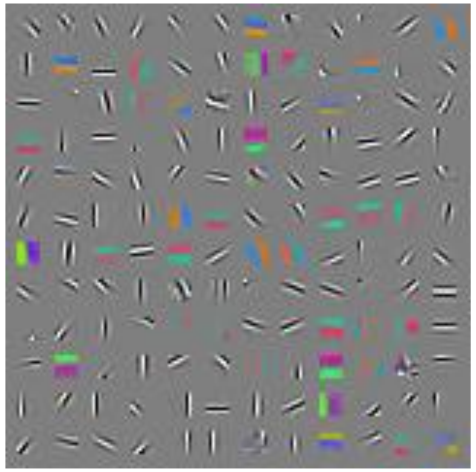
- **Fully Connected Baseline**
  - 2-3 hidden layer feed-forward neural network
    → **multilayer perceptron**
  - Rectified linear units
  - Batch normalization
  - Adam
  - Maybe dropout



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Spatial structure

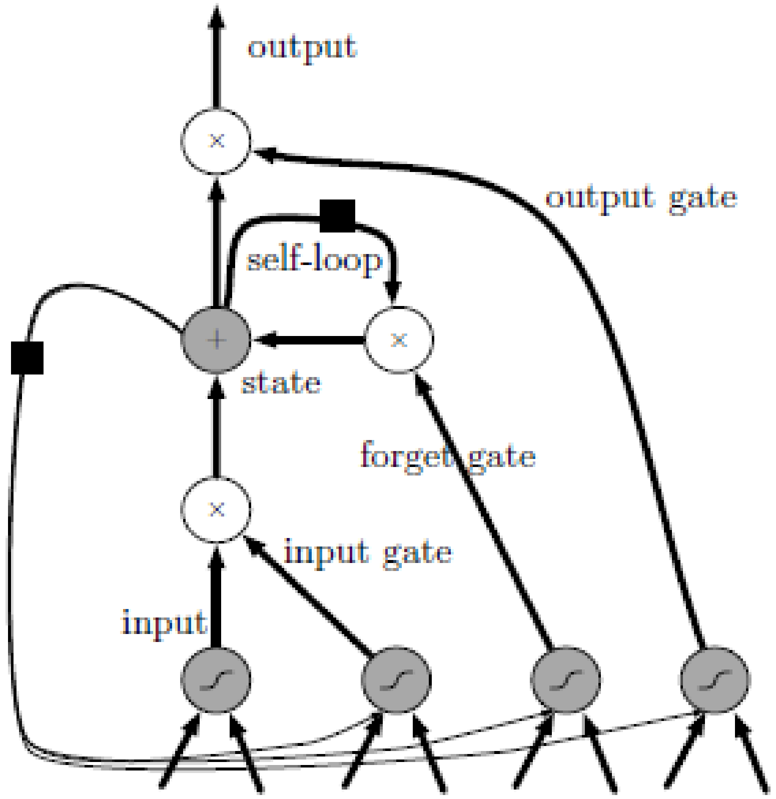
- **Convolutional Network Baseline**
  - Download a pretrained network (there could be many of them)
  - Or copy-paste an architecture from a related task
  - Or:
    - Deep residual network
    - Batch normalization
    - Adam



(Goodfellow 2016)

Sequential structure

- **Recurrent Network Baseline**
  - LSTM
  - SGD
  - Gradient clipping
  - High forget gate bias



(Goodfellow 2016)

Data-driven Adaptation

- Choose what to do based on data
- Don’t believe hype
- Measure train and test error
  - **Overfitting** versus **underfitting**
  - In case of problem, improve the algorithm or gather new data?
  - It is often much better to gather more data than to improve the learning algorithm.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Gather New Data?

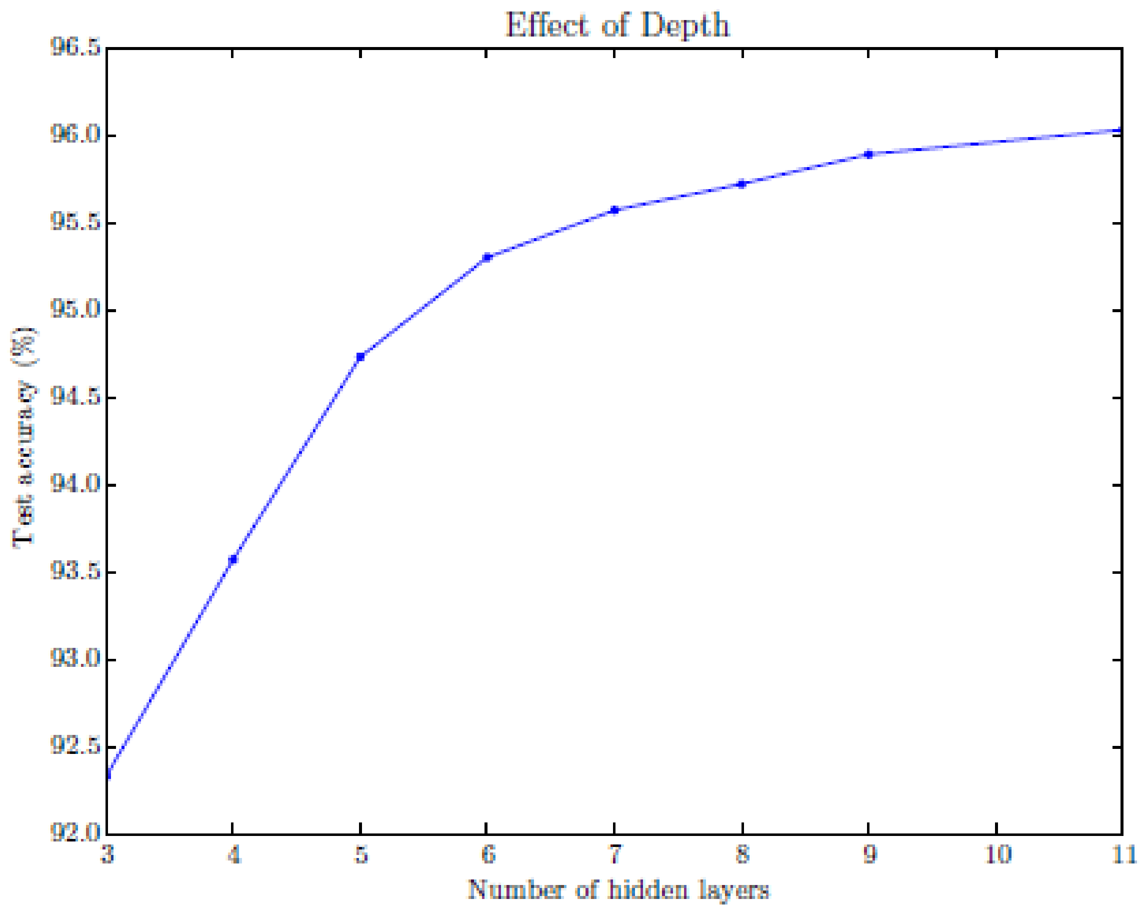
How does one decide whether to gather more data?

- Performance on the training set is poor
  - The learning algorithm is not using (correctly) the training data that is already available, so there is no reason to gather more data → try increase size of the model with more layers/hidden units, change hyperparameters, …
  - The data might be of poor quality (noise?) → collect cleaner data, inspect data for defects, perform a preprocess, …

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Gather New Data?

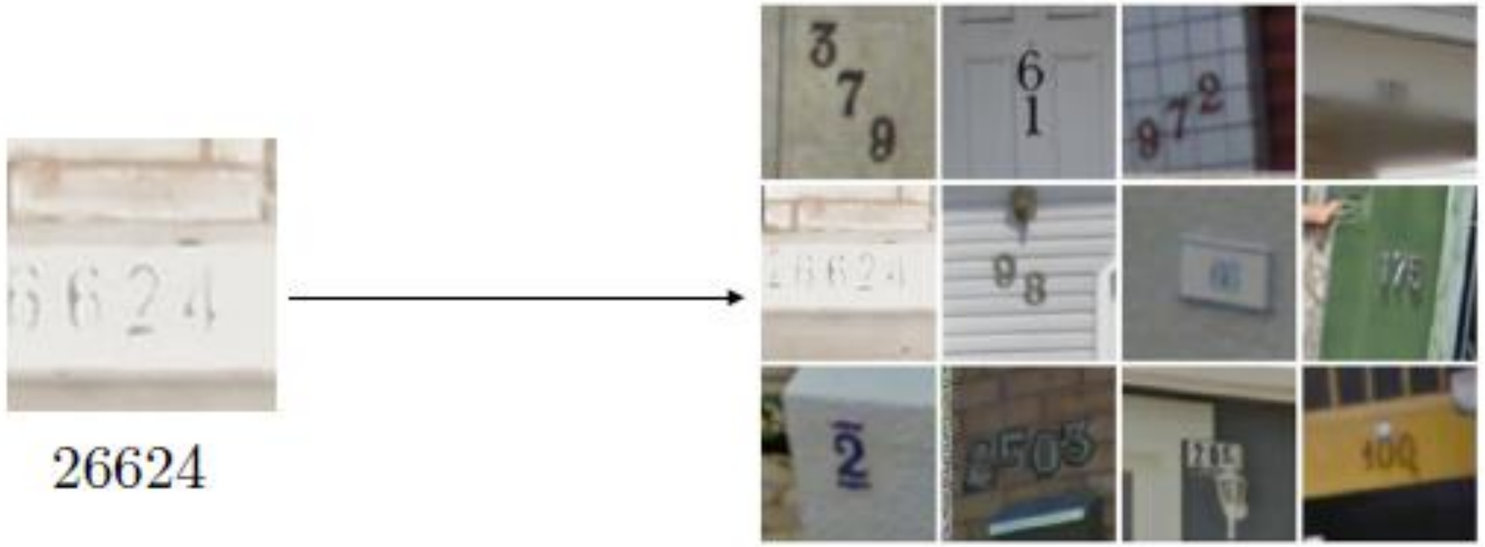
Increasing depth



(Goodfellow 2016)

Gather New Data?

Checking Data for Defects



(Goodfellow 2016)

Gather New Data?

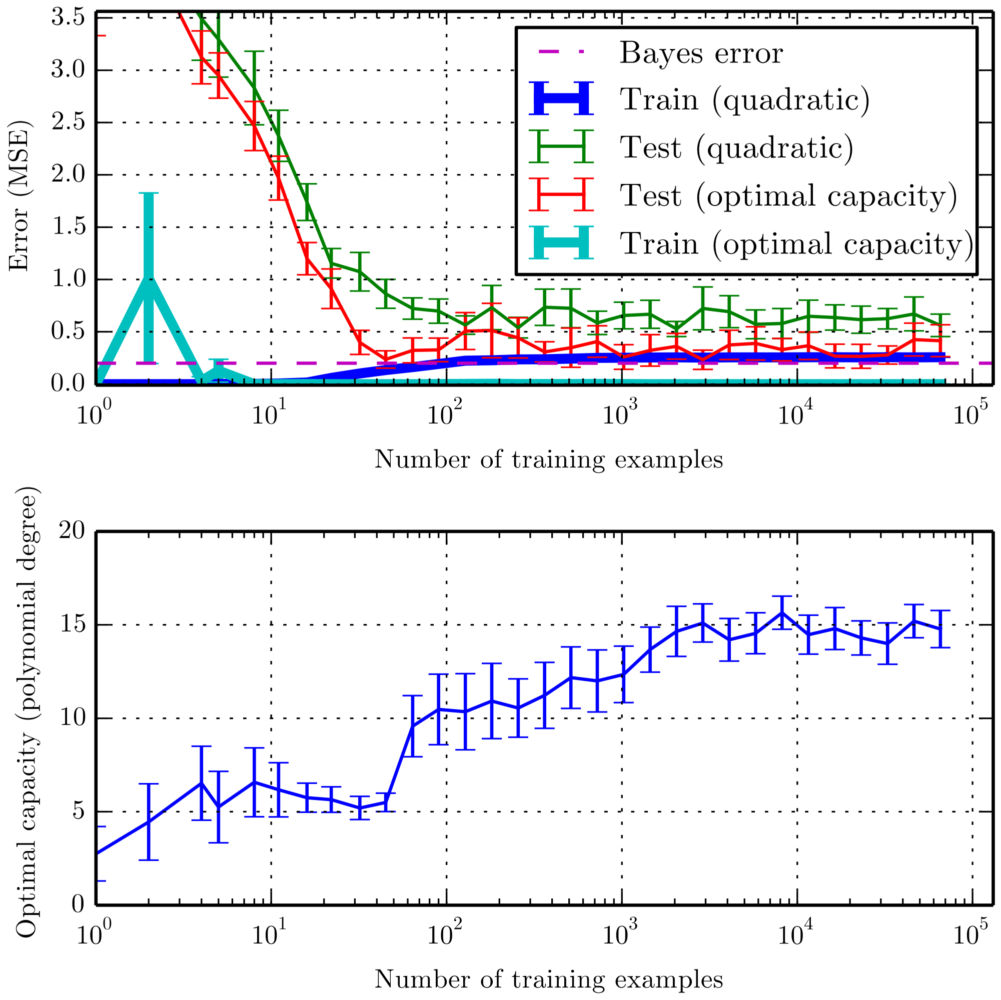
How does one decide whether to gather more data?

- Performance on the training set is good, performance on the test set is poor
  - Gathering more data is one of the most effective solutions. The development of large labeled datasets is one of the most important factors in solving object recognition. In many contexts it may be costly or infeasible to gather more data.
  - In these cases, reduce the size of the model or improve regularization, by adjusting hyperparameters, or by adding regularization strategies such as dropout.
  - Or, perform data augmentation.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

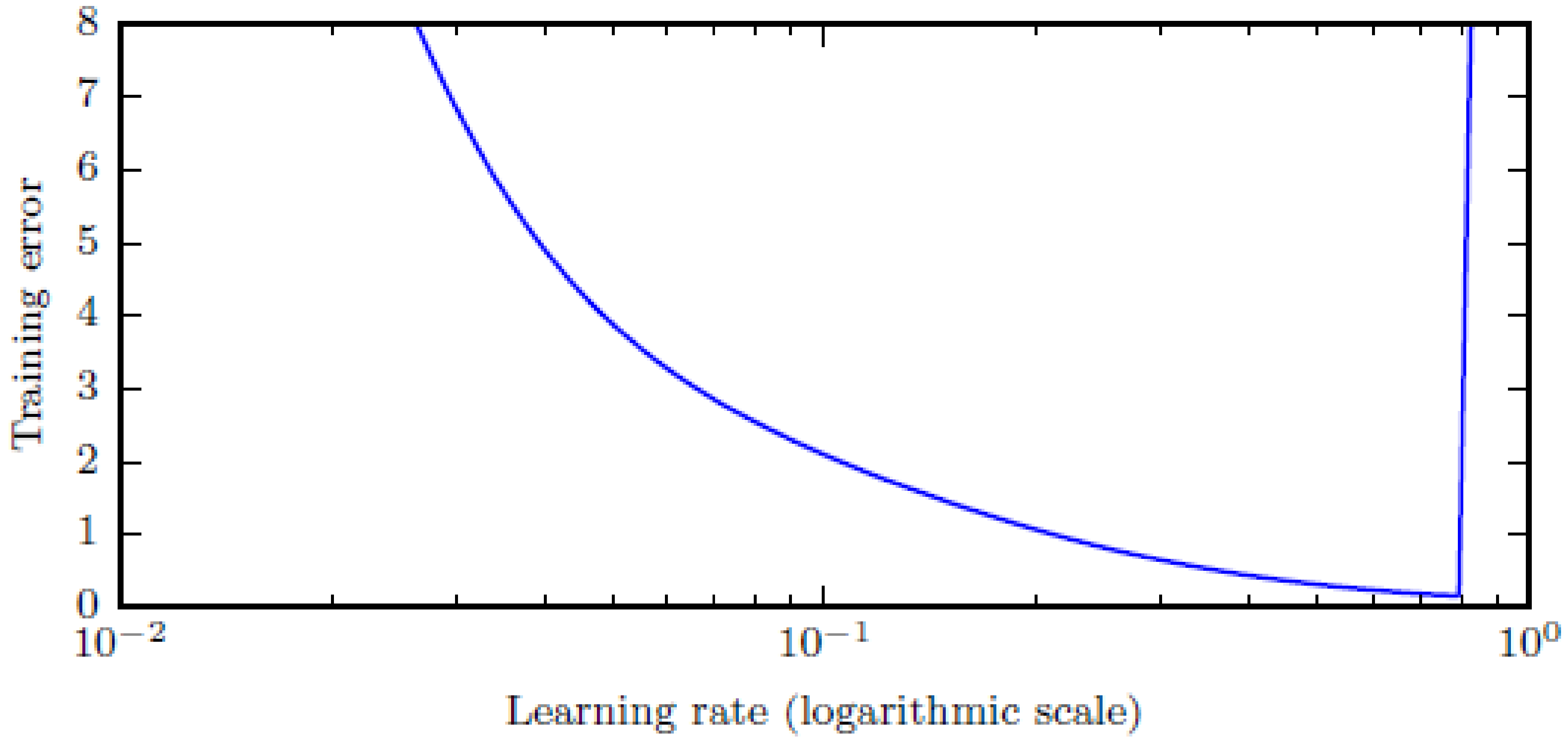
Gather New Data?

- How much to gather?
  - It is helpful to plot curves showing the relationship between training set size and generalization error



Gather New Data?

- Tuning the learning rate



(Goodfellow 2016)

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

About Hyperparameters

Hyperparameter	Increases capacity when	Reason	Caveats
Number of hidden units	Increased	The number of hidden units directly changes the capacity of the model	Increasing the number of hidden units increases both time and memory cost.
Learning rate	Tuned optimally	Whether too high or too low, a wrong value causes optimization failure	
Weights decay coefficient	decreased	Decreasing it frees model parameters to become larger	
Dropout rate	decreased	Dropping units less often leaves them to be fit more and with more other units	

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Regione Emilia-Romagna

Università degli Studi di Ferrara

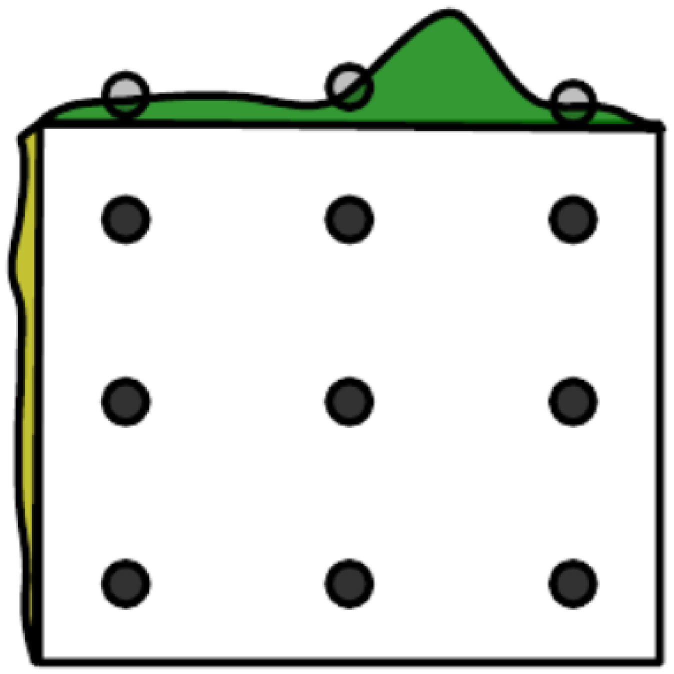
About Hyperparameters

- There are **hyperparameter optimization** algorithms that wrap a learning algorithm and choose its hyperparameters.
  - Note, they often have their own hyperparameters, such as the range of values that should be explored for each of the learning algorithm's hyperparameters.
  - However, these secondary hyperparameters are usually easier to choose.
- Two famous algorithms:
  - **Grid Search**
  - **Random Search**

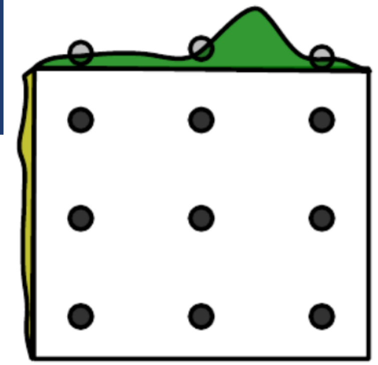
Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Grid Search

- When there are three or fewer hyperparameters.
- For each hyperparameter, the user selects a small finite set of values to explore.
- The grid search algorithm then trains a model for every joint specification of hyperparameter values in the Cartesian product of the set of values for each individual hyperparameter.



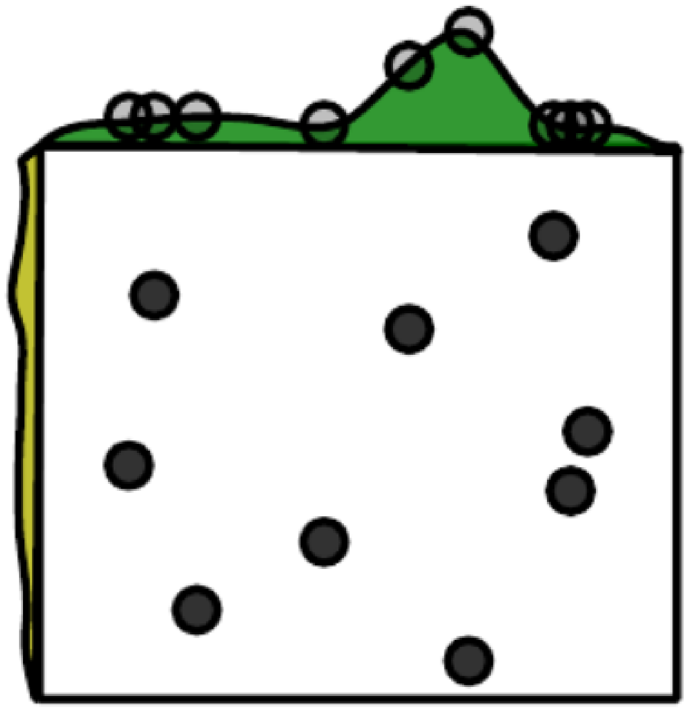
Grid Search

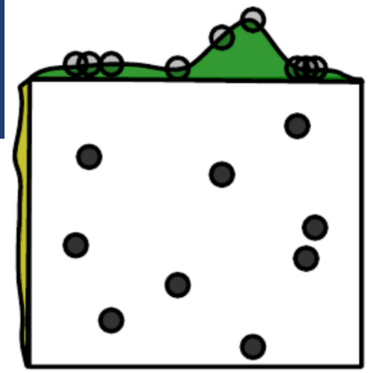


- In the case of numerical (ordered) hyperparameters, the smallest and largest element of each list is chosen conservatively, based on prior experience with similar experiments.
- Typically, a grid search involves picking values approximately on a logarithmic scale, e.g., a learning rate taken within the set $\{0.1, 0.01, 10^{-3}, 10^{-4}, 10^{-5}\}$, or a number of hidden units taken with the set $\{50, 100, 200, 500, 1000, 2000\}$.
- Grid search usually performs best when it is performed repeatedly (if we test an hyperparameter taking value in $[0,1]$ and we find the best value is 1, we should re-tune the hyperparameter in a new range, e.g., $[1,2]$

Random Search

- Simpler to program, more convenient to use, and converges much faster to good values of the hyperparameters than grid search.
- For each hyperparameter it defines a marginal distribution depending on the vales taken by the parameter (Bernoulli, Multinoulli, Uniform) and samples values from that.





Random Search

- Unlike in the case of a grid search, one should **not** discretize or bin the values of the hyperparameters.
- As with grid search, one may often want to run repeated versions of random search, to refine the search based on the results of the first run.
- The main reason why random search finds good solutions faster than grid search is that there are no wasted experimental runs, unlike in the case of grid search, when two values of a hyperparameter (given values of the other hyperparameters) would give the same result.

Gather New Data?

How does one decide whether to gather more data?

- Performance on the training set is good, performance on the test set is good
  - You are done!