Selecting Hyperparameters

Sargur N. Srihari
srihari@cedar.buffalo.edu

Topics

- Overview

1. Performance Metrics
2. Default Baseline Models
3. Determining whether to gather more data
4. Selecting hyperparamaters
5. Debugging strategies
6. Example: multi-digit number recognition

Types of hyperparameters

- Hyperparameters control algorithm behavior
- Some affect the time and memory cost of algorithm
- Some affect the quality of the model recovered during training process
  - And ability to infer correct results with new inputs

Approaches to choosing hyperparams

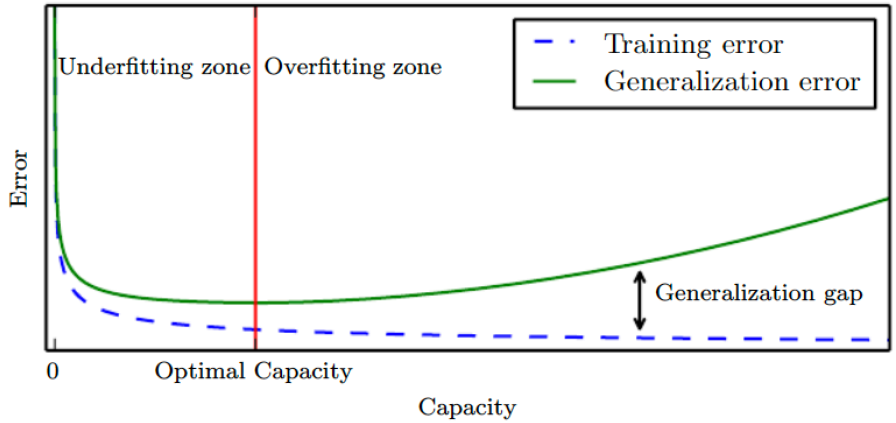
- Choosing them manually
  - Requires understanding of what they do
  - Knowledge of how they achieve good generalization
- Choosing them automatically
  - Reduce the need to understand these ideas
  - But computationally expensive

Manual hyperparameter tuning

- Need to understand relationship between
  - Hyperparameters, Training error, Generalization error, Computational resources (memory, time)
- Goal of hyperparameter search:
  - Adjust effective capacity of model to match complexity of task
  - Capacity is controlled by
    1. Representational capacity of model
    2. Ability of learning algorithm to minimize the cost
    3. Degree to which cost and training regularize model

Capacity and hyperparameters

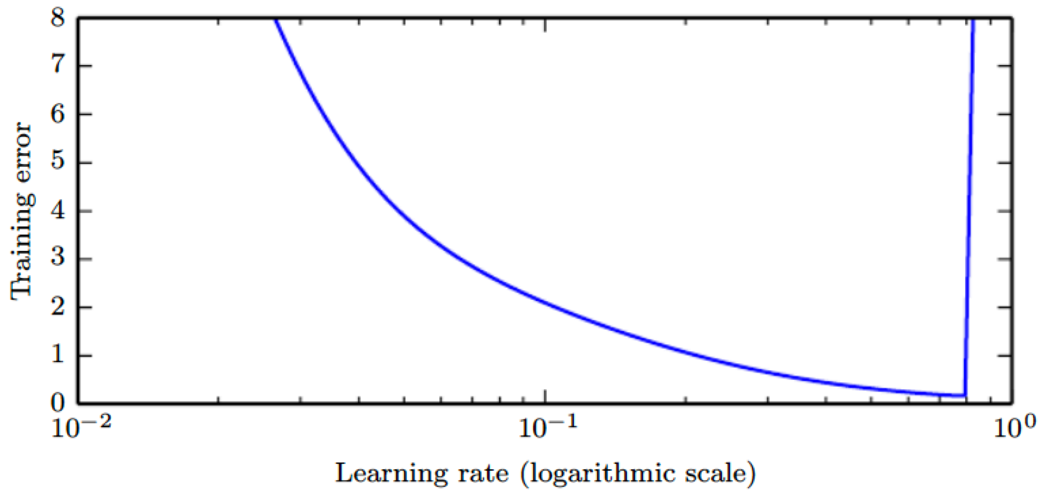
- A model with more layers and more hidden nodes per layer has higher capacity
  - But learning algorithm may not learn the function
- Generalization error is a U-shaped curve
  - Plotted as a function of one hyperparameter



  - Not every hyperparameter will be able to explore entire curve

Learning rate is most important

- Most important hyperparameter is learning rate
- It controls model capacity in a more complicated way than other hyperparameters
- Effective capacity is highest when learning rate is correct, not when it is large or small
- Learning rate vs training error has a U-curve



Effect of hyperparameters on capacity

- Hyperparameters are set based on whether they increase/decrease capacity

Hyperparameter	Increases capacity when...	Reason	Caveats
Number of hidden units	increased	Increasing the number of hidden units increases the representational capacity of the model.	Increasing the number of hidden units increases both the time and memory cost of essentially every operation on the model.
Learning rate	tuned optimally	An improper learning rate, whether too high or too low, results in a model with low effective capacity due to optimization failure	
Convolution kernel width	increased	Increasing the kernel width increases the number of parameters in the model	A wider kernel results in a narrower output dimension, reducing model capacity unless you use implicit zero padding to reduce this effect. Wider kernels require more memory for parameter storage and increase runtime, but a narrower output reduces memory cost.

Effect of hyperparameters on capacity

- Table continued

Implicit zero padding	increased	Adding implicit zeros before convolution keeps the representation size large	Increased time and memory cost of most operations.
Weight decay coefficient	decreased	Decreasing the weight decay coefficient frees the model parameters to become larger	
Dropout rate	decreased	Dropping units less often gives the units more opportunities to "conspire" with each other to fit the training set	

Automatic hyperparameter optimization

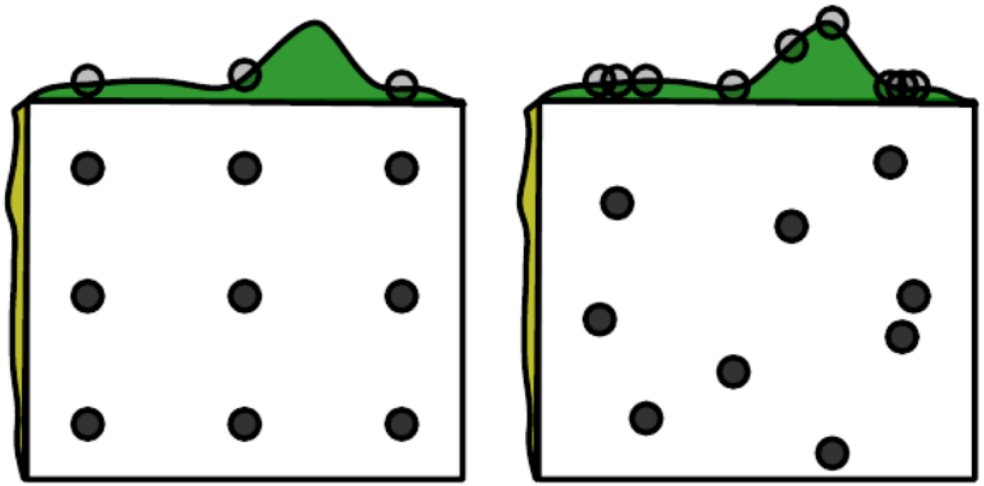
- In principle it is possible to develop hyperparameter optimization algorithms that wrap a learning algorithm and choose its hyperparameters
  - Thus hiding hyperparameters from the user
- But hyperparameter learning algorithms have their own hyperparameters such as range of values to be explored for hyperparameters

Grid Search

- When there are three or fewer parameters it is common to do grid search
- User selects a small set of values to be explored for each hyperparameter
- Then trains model for every joint specification of parameter values

Grid search vs Random search

- For grid search: provide a set of values for each hyperparameter
- For random search: provide a probability distribution over joint configurations
  - Hyperparameters are usually independent



Grid Random