Practical Deep Neural Networks

GPU computing perspective

Machine Learning Basics

Yuhuang Hu Chu Kiong Loo

Advanced Robotic Lab
Department of Artificial Intelligence
Faculty of Computer Science & IT
University of Malaya

Outline

Outline

Assumed prerequisites

- ☆ Basic Linear Algebra (DL book chapter 2)
- ☆ Basic Probability and Information Theory (DL book chapter 3)
- ☆ Basic Numerical Computation (DL book chapter 4)

Suggest Readings

- ✍ CS229 lecture notes 1 and 4:
  http://cs229.stanford.edu/notes/cs229-notes1.pdf
  http://cs229.stanford.edu/notes/cs229-notes4.pdf
- ✍ Pattern Recognition and Machine Learning: Chapter 1 and 2.
- ✍ Machine Learning: A probabilistic perspective: Chapter 1 and 2.
- ✍ CS231n: Image Classification: Data-driven Approach, k-Nearest Neighbor, train/val/test splits
  http://cs231n.github.io/classification/

Outline

Definition of Learning Algorithm

"A computer program is said to learn from

- ✓ experience E with respect to
- ✓ some class of tasks T and
- ✓ performance measure P,

if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell, 1997)

The task T

Classification specify which k categories some input belongs to. $(f : \mathbb{R}^n \to \{1, \dots, k\})$

Regression predict a numerical value given some input. $(f : \mathbb{R}^n \to \mathbb{R})$

Transcription output a sequence of symbols, rather than a category code. (similar to classification, *e.g.* speech recognition, machine translation, image captioning)

Denoising predict *clean* samples $\boldsymbol{x}$ from *corrupted* samples $\tilde{\mathbf{x}}$ (estimate $P(\mathbf{x}|\tilde{\mathbf{x}})$).

Many more types are not listed here.

The performance measure P

- Measure P is usually specific to the task T. (*e.g.* accuracy to classification)
- Batches of unseen *test* data is introduced to measure performance.
- Design measure P can be very subtle. It should be effective.

The experience E

Experience is what learning algorithms are allowed to have during learning process.

- Experience is usually an *dataset*, a collection of *examples*.
- *Unsupervised learning algorithms* experience a dataset containing many features, learning useful structure of the dataset (estimate $p(\mathbf{x})$).
- *Supervised learning algorithms* experience a dataset containing features, but each example is also associated with a *label* or *target* (estimate $p(\mathbf{y}|\mathbf{x})$).

Outline

Problem Description

The *task* is to build a system that can take a vector $\mathbf{x} \in \mathbb{R}^n$ as input and predict the value of a scalar $y \in \mathbb{R}$ as its output. Let $\hat{y}$ be the value that out model predicts y should take on. We define the output to be

$$\hat{y} = \mathbf{w}^\top \mathbf{x} + b$$

where $\mathbf{w} \in \mathbb{R}^n$ is a vector of parameters and b is an intercept term.

Performance Measure and Experience

The *experience* contains a set of training examples where each sample is a pair of input and output $(\mathbf{x}, y)$.

One *performance measure* here can apply is *mean squared error* of the model on test set. Let test example as $\mathbf{x}^{(\text{test})}$ and regression targets as $\mathbf{y}^{(\text{test})}$, $\hat{\mathbf{y}}^{(\text{test})}$ is the predictions of the model on the test set, the mean square error is given by:

$$MSE_{\text{test}} = \frac{1}{m} \sum_i (\hat{\mathbf{y}}^{(\text{test})} - \mathbf{y}^{(\text{test})})_i^2.$$

Outline

Generalization, Capacity, Overfitting, Underfitting

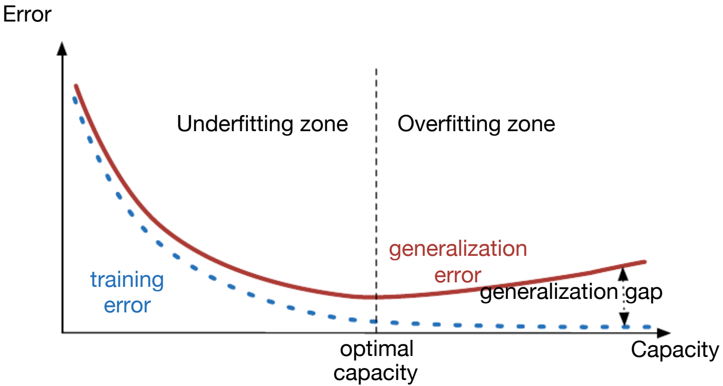
Generalization ability to perform well on previously unobserved inputs.

Capacity ability to fit a wide variety of functions.

Overfitting occurs when the gap between training error and test error is too large

Underfitting occurs when the model is not able to obtain a sufficiently low error value on the training set.

Generalization, Capacity, Overfitting, Underfitting



No Free Lunch Theorem

The no free lunch theorem for machine learning (Wolpert, 1996) states that, averaged over all possible data generating distributions, every classification algorithm has the same error rate when classifying previous unobserved points. In some sense, no ML algorithm is universally any better than any other.

Seek solution for some relevant distributions, NOT universal distribution.

Hyperparameters, Validation Sets

Hyperparameters settings that control the behavior of the learning algorithm. Usually choose empirically.

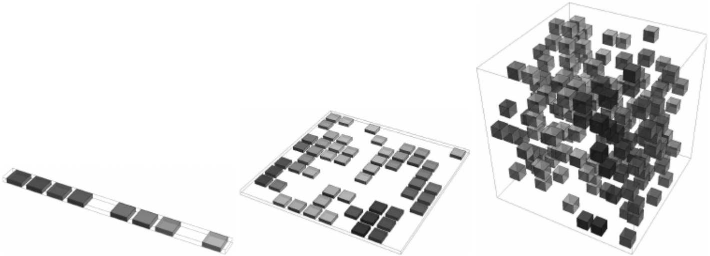
Validation Sets Subset of data used to guide the selection of hyperparameters. (split from training dataset, usually 20%)

Outline

Curse of Dimensionality

Many machine learning problem become exceedingly difficult when the number of dimensions in the data is high. The phenomenon is known as the *curse of dimensionality*. Of particular concern is that the number of possible distinct configurations of the variables of interest increases **exponentially** as the dimensionality increases.

Curse of Dimensionality



Q& A

References I

Mitchell, T. M. (1997). *Machien learning*. New York: McGraw-Hill.

Wolpert, D. H. (1996, October). The lack of a priori distinctions between learning algorithms. *Neural Comput.*, *8*(7), 1341–1390.