Machine Learning Basics

Advanced Topics in High-Performance Computing

Faisal Qureshi

UNIVERSITY OF ONTARIO INSTITUTE OF TECHNOLOGY

Learning algorithms

A machine learning algorithm is an algorithm that is able to learn from data.

> *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)

Task T

- The process of learning is *not* itself a task
- Learning allows us to attain the ability to perform the task

Kinds of tasks

- Classification
- Classification with missing inputs
- Regression
- Transcription
- Machine Translation
- Density estimation
- Structured output
- Anomaly detection
- Synthesis and sampling
- Imputation of missing values
- Denoising

Performance measure P

- We are often interested in how well a machine learning system *performs* on the data that it hasn't seen before.
- Deciding upon an appropriate *performance measure* is not a simple, straightforward tasks.
  - Consider, for example, transcription. How should we measure performance? Should we measure the accuracy of the system at transcribing the entire sequences only?
  - For regression, is it better to make a small error for many examples or one large error for a single example?
- Commonly used performance measures
  - Accuracy
  - Error rate

Experience E

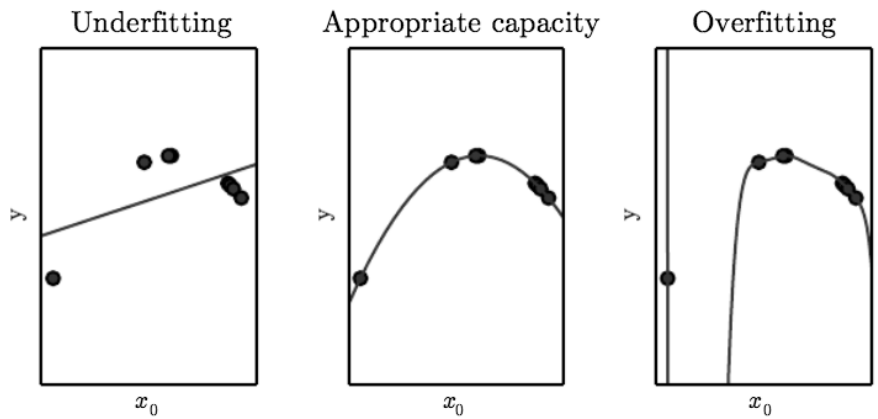
- Broadly speaking, we classify machine learning algorithms as *supervised* or *unsupervised*
- Another class of machine learning algorithms, called *reinforcement learning* do not learn from a fixed data set. These algorithms interact with the environment.

Generalization

- The ability to perform well on new, previously unseen inputs is call *generalization*
- Machine learning algorithms that fail to generalize are typically of little use.
- Error measure computed over the *training set* is called *training error*.
- Error measure computed over the *test set* is called *test error* or *generalization error*.
- *Generalization error* is defined as the expected value of the error on a new, previously unseen input.
- How well a machine learning algorithm behaves depends upon its ability to minimize the training error and reduce the gap between training and test error.

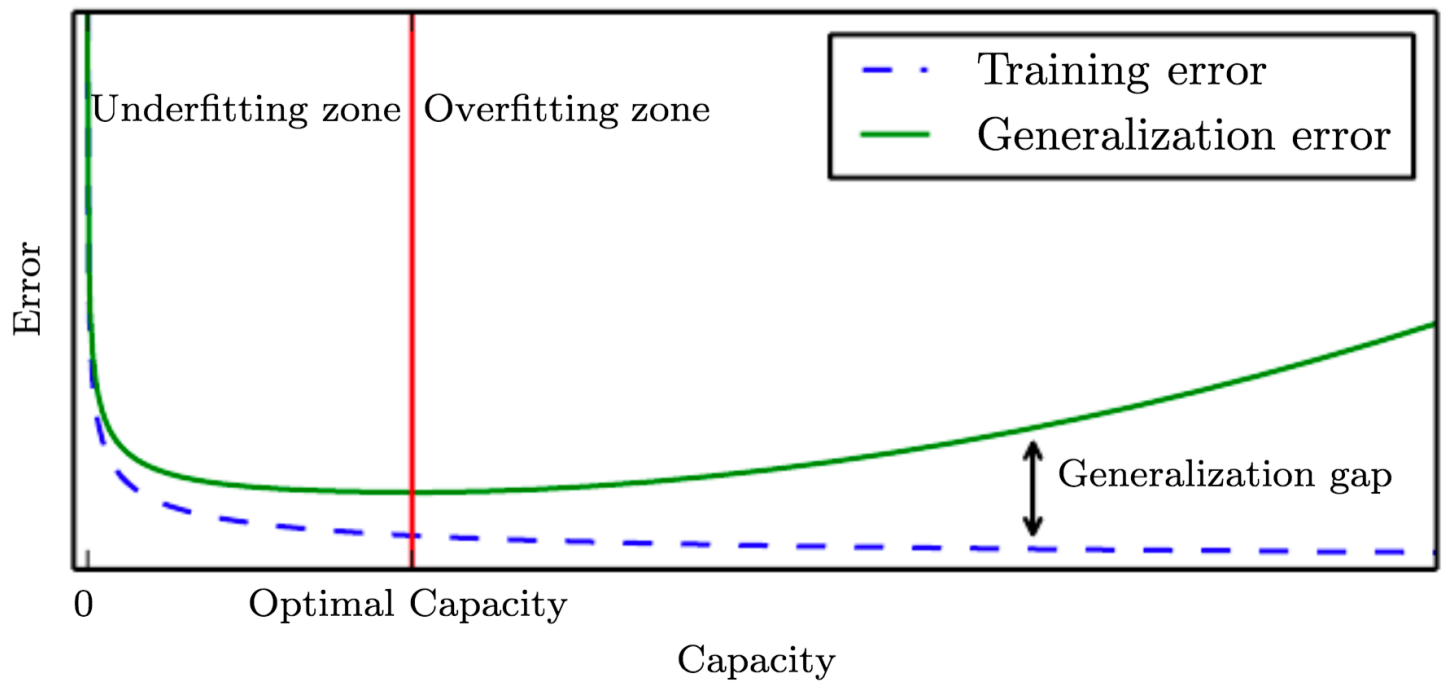
Underfitting and overfitting

- Underfitting occurs when the machine learning algorithm is not able to obtain a sufficiently low error value on the training data set.
- Overfitting occurs when the gap between the training and test error is too large.
- We can reduce both underfitting and overfitting by selecting appropriate *capacity* of the model.
  - Models that are too simple, often underfit.
  - Models that are too complex, often overfit.



Underfitting and overfitting (Goodfellow et al., 2017)

Underfitting and overfitting



Error vs. capacity (Goodfellow et al., 2017)

The No Free Lunch Theorem

> *Averaged over all possile data-generating distributions, every classification algorithm has same error rate when classifying previously unobserved points.*

(Wolpert, 1996)

In other words, no machine learning algorithm is universally better than any other. However, if we make assumptions about the kind of data-generation probability distributions, we can design a machine learning algorithm that will perform better on these distributions.

Regularization

- Any modification we make to the learning algorithm that is intended to reduce its generalization error but not its training error.
- Regularization is a form of expressing a preference over one function over an other.
- The No Free Lunch Theorem also implies that there is *no* best form of regularization.
- The philosophy of deep learning is that a wide range of tasks can be solved using very general-purpose form of regularization.

Bias and Variance

Bias is a measure of how well does a model perform on training data. A high bias suggests that model parameters are far from the true, unknown parameters that will reduce the training error. The model misses important features of the data, thus underfitting.

Variance is an error due to small fluctuations in the training set. Model is stuck in noise, unimportant features of the data, thus overfitting.

Summary

We have briefly discussed concepts that appear time and again when studying machine learning. We will revisit these concepts and discuss them in greater detail in the upcoming lectures.

Readings

- Ch. 5, Goodfellow, et. al., 2017