Introduction

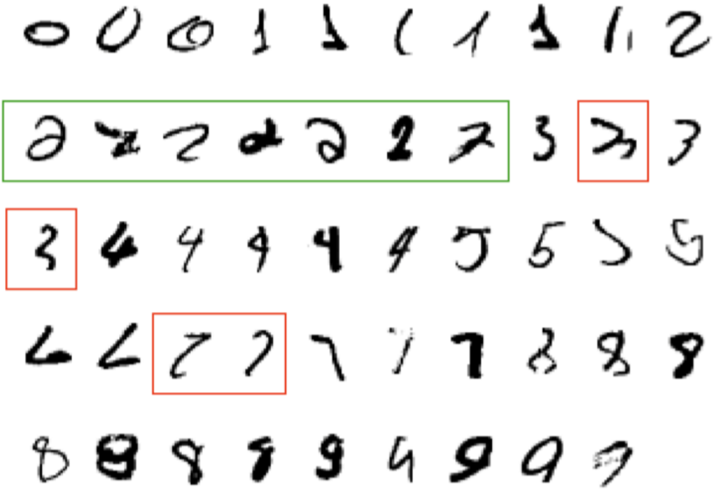
Machine Learning (CSCI 5770G)

Faisal Z. Qureshi

http://vclab.science.ontariotechu.ca

Part I

What Makes a "2?"



Courtesy R. Urtasun

What Makes a “2?”

- It is very hard to write programs that solve problems like recognizing a handwritten digit
  - What distinguishes a 2 from a 7?
  - How does our brain do it?
- Instead of writing a program by hand, we collect examples that specify the correct output for a given input
- A machine learning algorithm then takes these examples and produces a program that does the job
  - The program produced by the learning algorithm may look very different from a typical hand-written program. It may contain millions of numbers.
  - If we do it right, the program works for new cases as well as the ones we trained it on.

Applications of Machine learning?

When to Apply Machine Learning?

- Human expertise is absent
- Humans are unable to explain their expertise
- Solution changes with time
- Solution needs to be adapted to particular cases
- The problem size is vast for our limited reasoning capabilities

Machine Learning and Statistics

- Machine learning deals with inference in the presence of uncertainty
  - Statistical theory to build models
- Machine learning can be seen as applying computational techniques to statistical problems

Machine Learning and Statistics (R. Tibshirani)

	Machine Learning	**Statistics**
	network, graphs	models
	weights	parameters
	learning	fitting
	generalization	test performance
	supervised learning	regression, classification
	unsupervised learning	density estimation, clustering
large grant	$1,000,000	$50,000
conference location	French Alps	Las Vegas in August

Data and Machine Learning

- Machine learning assumes access to large corpus of data for training
- “Large” text dataset
  - 1,000,000 words in 1967
  - 1,000,000,000,000 words in 2006
  - 1,00,000,000,000,000 words in 2023 (the Internet)

Data and Machine Learning

- Large Synoptic Survey Telescope
  - 3,200,000,000 pixels, almost 1500 times the resolution of HDTV (2,073,600)
- Rubin Observatory's Legacy Survey of Space and Time
  - 20,000,000,000,000 bytes per night
  - 60,000,000,000,000,000 over 10 years

LSST CAMERA
World's Largest Camera
for
Astronomy

Courtesy https://youtu.be/eq5fopwwW3M

Types of Learning

	Supervised	**Unsupervised**
Discrete	Classification	Clustering
Continuous	Regression	Dimensionality reduction

$+$ Reinforcement learning

Supervised Learning

Classification

- Outputs are categorical (1 to N)
- Inputs can be anything
- Goal: select correct class for new inputs
- Examples: object recognition, medical diagonisis, fault detection, etc.

Regression

- Outputs are continuous
- Inputs can be anything (typical continuous)
- Goal: predict outputs for new inputs
- Examples: customer ratings, house prices, object distances, etc.

Temporal predictions

- Goal: classification or regression on new inputs given previous items in the sequence

Unsupervised Learning

Clustering

- Inputs are categorical (or vectors)
- Group data based upon some notion of “similarity”

Dimesionality reduction

- Goal: construct an encoder/decoder such that the size of encoder output is much smaller than the original. Encoder followed by a decoder returns values similar to the original input.

Anomaly detection

- Inputs can be anything
- Goal: select highly unusual cases

Machine Learning and Data Mining

Data mining

- In the past, using simple machine learning algorithms on very large datasets
- (Perhaps misuse) of statistical procedures to look for hidden relationships that may exist within data

Machine learning

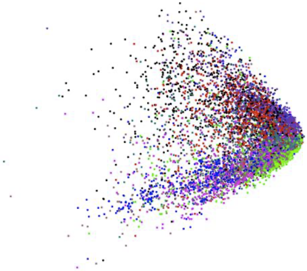
- More recently the lines between machine learning and data mining have been blurred.
- Machine learning is now in the ascendence. Machine learning is also been applied to problems in the domain of data mining.

Art of Machine Learning

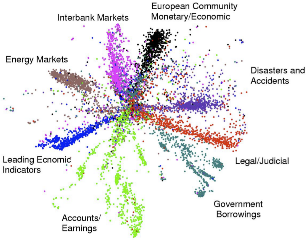
- Deciding how to represent inputs (and outputs)
- Selecting a hypothesis space (i.e., the model)
  - Must have the right complexity
    - Rich enough to capture the relevant relationship between inputs and outputs
    - Simple enough to be searched
    - Simple enough to not get bogged down by unnecessary details
- Data and compute considerations

Encoding Information to Reduce the *Semantic Gap*

- Each document is converted to a vector of word counts, which is mapped to a 2D space. The color represents hand-labeled classes. The 2D layout shows how good is the encoded information at separating document classes.



Latent Semantic Analysis



Neural Networks

Other Challenges

- Multi-task and transfer learning (generalization)
- Scaling and energy efficiency
- Ability to generate data (e.g., computer vision as inverse graphics)
- Architectures for artificial intelligence
- Ethical and social impact considerations

Part II

Let's be a little formal

Learning algorithms (Mitchell, 1997)

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.*

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
- Dimensionality reduction
- Structured output
- Anomaly detection
- Synthesis and sampling
- Imputation of missing values
- Denoising
- Clustering

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

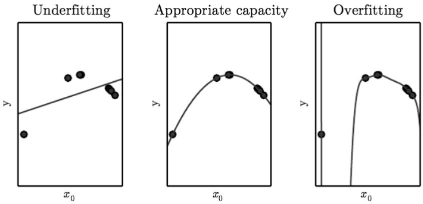
- The ability to perform well on new, previously unseen inputs is call generalization.
- Machine learning algorithms that fail to generalize are typically of little use.
- Error measure computed over the *training set* is called training error.
- Error measure computed over the *test set* is called test error or generalization error.
  - Generalization error is defined as the expected value of the error on a new, previously unseen input.

Training and Test errors

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.

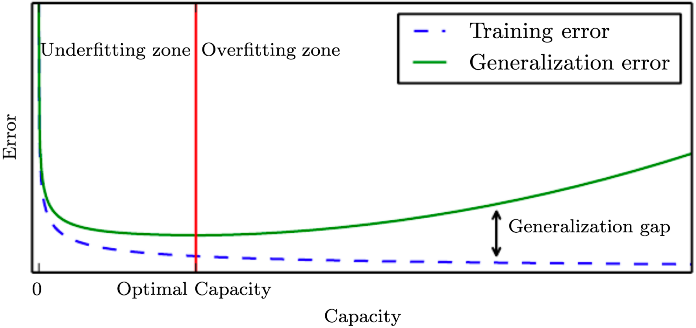
Underfitting and overfitting

- 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

Copyright and License