CS5339 Machine Learning

Introduction

Lee Wee Sun
School of Computing
National University of Singapore
leews@comp.nus.edu.sg

Semester 2, 2019/20

We would like machines to *learn* from data, instead of being explicitly programmed.

The analytical engine has no pretensions to *originate* anything. It can do *whatever we know how to order it* to perform.

Ada Lovelace, 1842.

Objectives

- Provide broad theoretical understanding of machine learning, and how the theory guides the development of algorithms and applications.
- Covers
  - **Representation**: What are the common representations used in machine learning? How do we design appropriate representations?
  - **Estimation**: How much data is required to learn from a class of functions? How do we use the data appropriately to learn?
  - **Optimization**: How much computation is required to do learning? How do we do the learning?

Textbooks

Understanding Machine Learning: From Theory to Algorithms
Shai Shalev-Shwartz and Shai Ben-David
Cambridge University Press.

Online copy from NUS library or
`http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/copy.html`

Useful Online Books

Deep Learning
Ian Goodfellow and Yoshua Bengio and Aaron Courville
Online copy from: `http://www.deeplearningbook.org/`

Boosting: Foundations and Algorithms
Robert Schapire and Yoav Freund
Online copy from: `https://mitpress.mit.edu/sites/default/files/titles/content/boosting_foundations_algorithms/titlepage.html`

Foundations of Data Science
Avrim Blum and John Hopcroft and Ravindran Kannan
Online copy from: `https://www.cs.cornell.edu/jeh/book%20no%20so;utions%20March%202019.pdf`

Pre-requisites

- Knowledge of basic machine learning (CS3244).
- Knowledge of the following is assumed
  - Calculus
  - Linear algebra (see background lecture video)
  - Statistics
  - Algorithms
- This is mostly a theory course – mathematical maturity is assumed.
- You will be required to **read theorems and understand/do proofs**.
  - The first two weeks require less of that.
  - To get a sense of the level of mathematical maturity expected (and whether the course is suitable for you), have a look the linear algebra background lecture.

Teaching Staff

Lecturer: Lee Wee Sun (leews@comp.nus.edu.sg)
Teaching Assistants:

- Dixant Mittal (e0210462@u.nus.edu)
- Huang Hengguan (e0409767@u.nus.edu)

Software

- For class exercises, we will use Scikit Learn `http://scikit-learn.org/`.
  - Install instructions `http://scikit-learn.org/stable/install.html`. Usually install through distributions such as Anaconda or Canopy.
  - To figure out how to use, see tutorials `http://scikit-learn.org/stable/documentation.html`.
  - We will also use Jupyter Notebook, which comes with Anaconda or Canopy in class exercises.
  - Download the dataset used from IVLE and put it in the directory where you will put the IPython/Jupyter notebooks.
- For some deep learning exercises, we will use Keras `https://keras.io/`.
  - See the Keras page for installation.

Graded Components

- Project (30%)
- Three problem sets (30%)
- Final exam (40%)

Exercises

- There is no separate tutorial for this course.
- Exercises are folded into the lecture.

Project

- Done in teams of 2 students.
- Focus on understanding.
- Most projects should study a topic from a recent research paper.
  - Review the paper. Give your own interpretation of the results.
  - Give additional examples or counterexamples.
  - Look at limitations.
  - Suggest possible future work.

Common Learning Problems

Common learning problems include:

- Supervised learning
  - Batch and Online learning
  - Active learning
- Unsupervised learning
- Reinforcement learning

Supervised Learning

Given a training set $S = ((x_1, y_1), \ldots, (x_m, y_m))$ drawn from $\mathcal{X} \times \mathcal{Y}$, the learning algorithm outputs a predictor $h : \mathcal{X} \to \mathcal{Y}$ that gives accurate prediction of y given x.

- When y is categorical, we are doing **classification** and h is often called a classifier.

Object Recognition [7]

- In object recognition, we want a classifier that takes in an image and outputs the class of the object shown in the image.
- The classifier is often learned using supervised learning.
  - Deep convolutional neural networks has been very successful recently.
  - ImageNet competition: 1000 classes, more than 1 million training images
  - 2010 to 2015 error rates: 28.2, 25.8, 16.4, 11.7, 6.7, 3.6
  - Around human level performance on the dataset now.

Spam Filtering

In spam filtering, we want a classifier that takes in an email and output whether it is *spam* or *ham* (non-spam).

- Often created by learning.
- Widely used on most people's email account.

- Instead of using only *spam* or *ham* as output, we can output a real value representing the probability of *spam*.
- The output can be thresholded using different thresholds to minimize false positive.
- Problems requiring real-valued outputs are often referred to as **regression**, solved e.g. using logistic regression.

Collaborative Filtering

In collaborative filtering, we are given data on the item ratings of many users, usually described as a matrix X where $x_{i,j}$ is the rating of user i on movie j.
In 2006, Netflix offered a million dollar prize for any team that can improve the video rating prediction algorithm it was using by 10%. The prize was won in 2009 using a combination of methods.

- One of the main techniques used for collaborative filtering is matrix factorization, approximating X with a low rank approximation $X_k = U_k V_k^T$.
- Closely related to SVD.
- Matrix X is partially (very sparsely) filled as most users have only watched a small number of movies.
- The low rank approximation fills in the missing entries.

Can treat this as a **regression** problem, predicting the real-valued missing rating given a user-movie pair.

Batch and Online Learning

Consider the spam filter problem.

- If we are given a training set $S = ((x_1, y_1), \ldots, (x_m, y_m))$ and asked to output a hypothesis h, we have a **batch learning** problem.
  - If you are designing a general spam filter to be used by many people, you may do batch learning.
- When the spam filter is being used by a particular user, it would be desirable to update the filter for that user each time the user provides a new label (spam or ham) to the spam filter.
  - This is an **online learning** problem, where we have to update the hypothesis after each new label is received.

Active Learning

Consider building a personalized spam filter for a new user.

- Can ask the user to label a large sample of his randomly selected emails: batch learning. The number of emails to be labeled may be inconveniently large.
- Get the person to label emails as ham or spam interactively.
  1. Machine selects an email to be labeled with the objective of learning quickly.
  2. On receiving the label, the machine updates its model and repeat Step 1.
- Called **active learning**. Often reduces the amount of labeling required substantially.

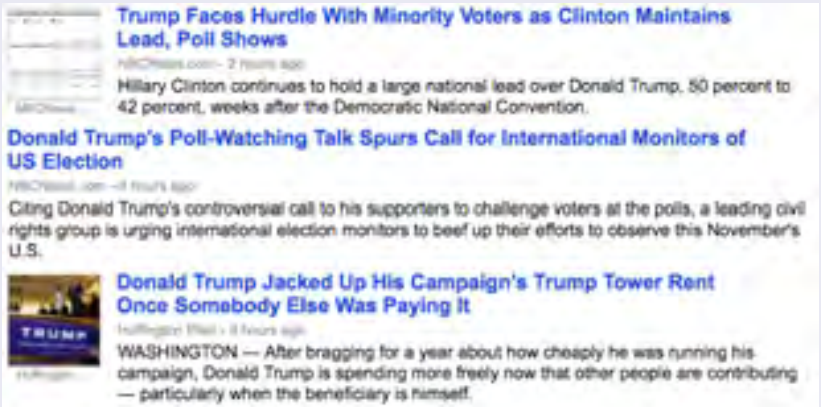
Unsupervised Learning

Given a training set $S = (x_1, \ldots, x_m)$, without a labeled output, construct a "good" model/description of the data.

- Look at the model/description, and find "interesting structure", e.g. clustering.
- Can be used for dimension reduction to find the essential parts of the data and remove noise, e.g. PCA.
- Often used for representation learning to learn a representation or embedding that is useful for other tasks (representation often used as features for supervised learning).
- Unsupervised learning often minimizes description length of data: useful for efficient data transmission/storage.
- Model of the data can also be used scoring how likely the data is, and for generating similar data.

Clustering

Organizing News



Google News group all the articles about the same topic together into clusters to organize the articles for the readers.

- Articles within the same cluster are similar to each other.
- Articles in different clusters are different compared to articles from the same clusters.

Clustering is often used to organize data into groups that are (hopefully) meaningful to users.

PCA and Dimension Reduction

Eigenfaces [4]

- Eigenfaces was a face recognition method developed in the early 1990s.
  - Compute PCA and represent each image with its k largest principle components (unsupervised dimension reduction).
  - Can do face recognition using the nearest neighbour on the principle components (in the low dimensional subspace). Substantial improvement over using nearest neighbour with raw images. Can be considered as representation learning.

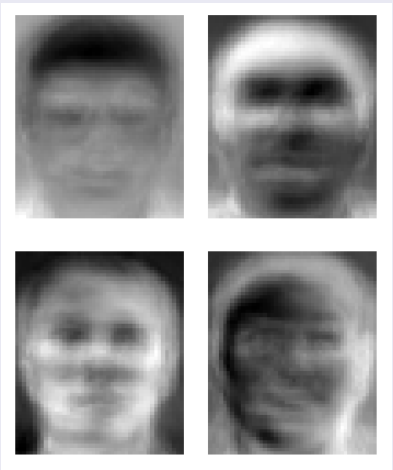


Figure: First few eigenvectors

Data Compression

- In **lossy compression**, we seek to trade off code length with reconstruction error. Often used when small error can be tolerated, e.g. for audio, images and video.
- In **vector quantization**, we seek a small set of vectors $\{\mathbf{z}_i\}$ to describe a large dataset of vectors $\{\mathbf{x}_i\}$ such that we can represent each $\mathbf{x}_i$ with its closest approximation in $\{\mathbf{z}_i\}$ with small error. This is a clustering problem, and algorithms for vector quantization are often equivalent to clustering algorithms, e.g. k-means.

- In **transform coding**, we transform the data, usually using a linear transformation. We then quantize the data in the transformed domain, usually discarding the small coefficients, corresponding to removing some of the dimensions.
- The optimal transform in terms of giving the best approximation with a small number of dimensions is the Karhunen-Loeve transform, or equivalently the principal component analysis.
- Image and video compression standards usually use cheaper to compute transformations such as wavelet or discrete cosine transforms.

Generative Models

- A probabilistic model $p(x_1, \ldots, x_n)$ can be used to score how likely data $x_1, \ldots, x_n$ is.
  - For example, useful for machine translation for selecting sentence to generate, given input sentence in another language.
- These models often generate realistic data.

< S > With even more new technologies coming onto the market quickly during the past three years , an increasing number of companies now must tackle the ever-changing and ever-changing environmental challenges online . < S > Check back for updates on this breaking news story . < S > About 800 people gathered at Hever Castle on Long Beach from noon to 2pm , three to four times that of the funeral cortège . < S > We are aware of written instructions from the copyright holder not to , in any way , mention Rosenberg 's negative comments if they are relevant as indicated in the documents , " eBay said in a statement . < S > It is now known that coffee and cacao products can do no harm on the body . < S > Yuri Zhirkov was in attendance at the Stamford Bridge at the start of the second half but neither Drogba nor Malouda was able to push on through the Barcelona defence .

Generated text from [5].

Can also generate images.

From [9].

Generated images. Top from [6], bottom from [3].

Reinforcement Learning

Autonomous Helicoptor Flight

Figure: `http://heli.stanford.edu/`

Given the current state of the helicopter (position, orientation, velocity, angular velocity), what control action should the helicoptor take to complete its goal (e.g. stunt flying)?

In **reinforcement learning**:

- Each action in a state has an associated cost and a probability distribution of the next state.
- Goal is to learn a policy (mapping from state to action) that minimizes the sum of expected current and future costs.

AlphaGo

Image from [2].

AlphaGo (neural networks plus Monte Carlo tree search) defeated 18-time World Champion Lee Sedol in 2016.

- Board position is the state.
- Learned the policy first by supervised learning (from expert games), then by self-play using reinforcement learning.

References I

[1] *Ada Lovelace Biography*. [Online: http://www.biography.com/people/ada-lovelace-20825323, accessed July 2017].

[2] *AlphaGo AI Defeats Sedol Again, With 'Near Perfect Game'*. [Online: http://www.tomshardware.com/news/alphago-defeats-sedol-second-time,31377.html, accessed July 2017].

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. "Large scale gan training for high fidelity natural image synthesis". In: *arXiv preprint arXiv:1809.11096* (2018).

[4] *Eigenfaces*. [Online: https://en.wikipedia.org/wiki/Eigenface#/media/File:Eigenfaces.png, accessed July 2017].

References II

[5] Rafal Jozefowicz et al. "Exploring the limits of language modeling". In: *arXiv preprint arXiv:1602.02410* (2016).

[6] Tero Karras et al. "Progressive growing of gans for improved quality, stability, and variation". In: *arXiv preprint arXiv:1710.10196* (2017).

[7] Olga Russakovsky et al. "Imagenet large scale visual recognition challenge". In: *arXiv preprint arXiv:1409.0575* (2014).

[8] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 815–823.

References III

[9] Jun-Yan Zhu et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *arXiv preprint* (2017).