Southeast Asia Machine Learning School

Introduction to Machine Learning

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

July 2019

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.

Common Learning Problems

Common learning problems include:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Outline

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 [6]

- 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.
  - 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 ImageNet now.

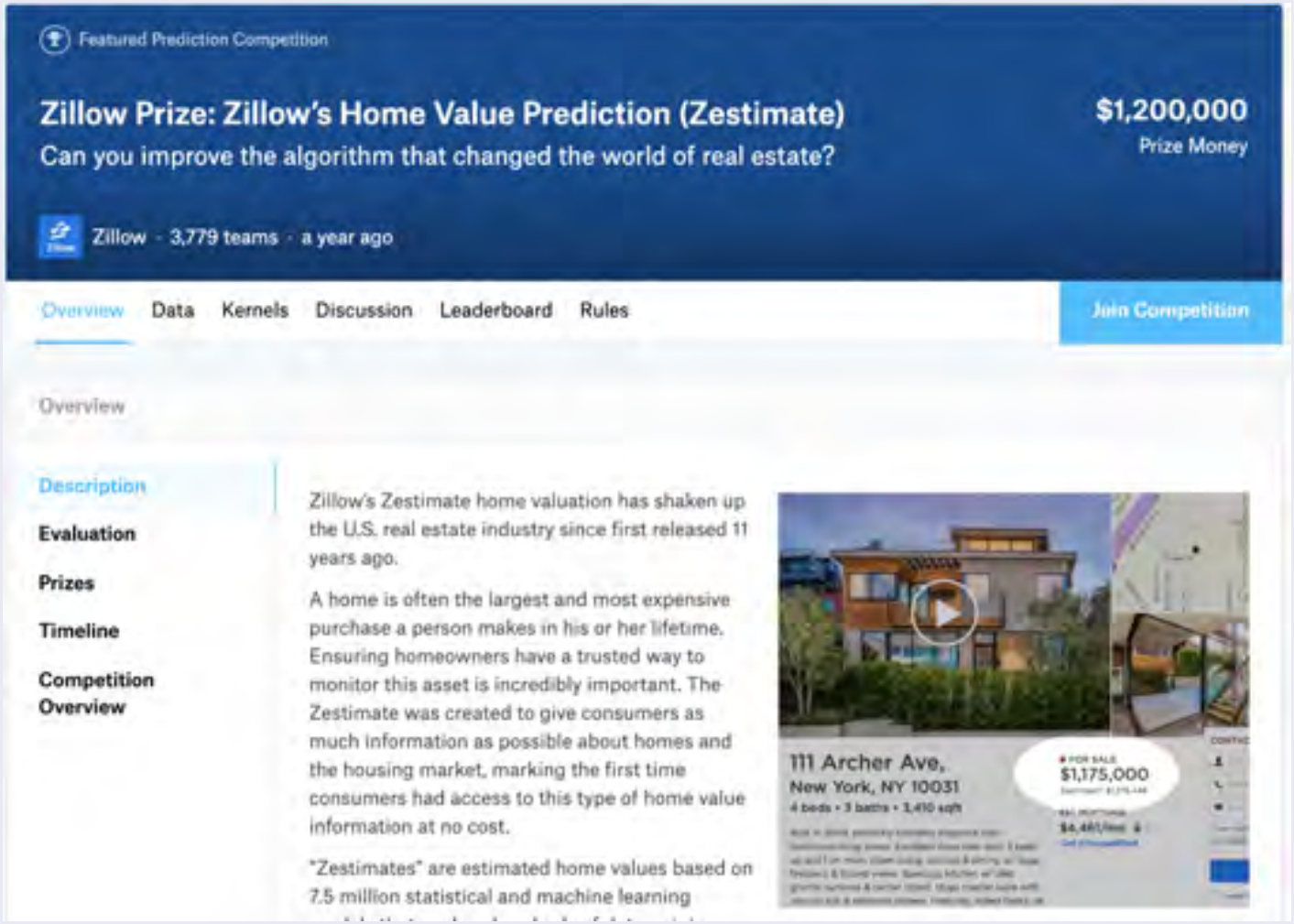
Spam Filtering

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.

Home Price Prediction



- Predict home price given location, size, number of rooms, etc.
- This is a regression problem.

Outline

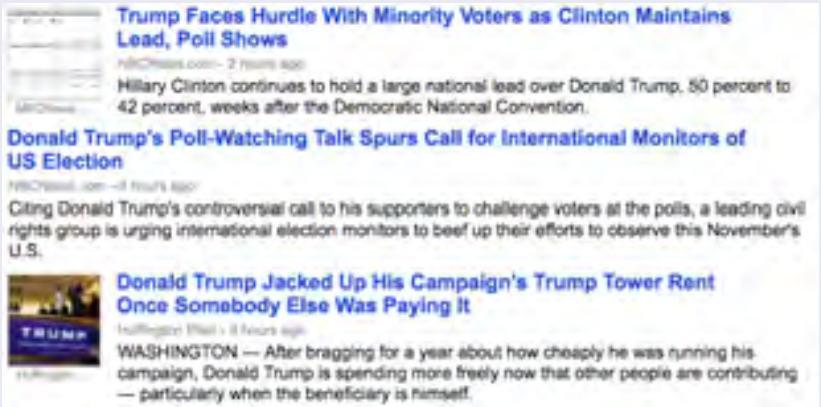
Unsupervised Learning

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.
- 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 groups all the articles about the same topic together into clusters to organize the articles for the readers.

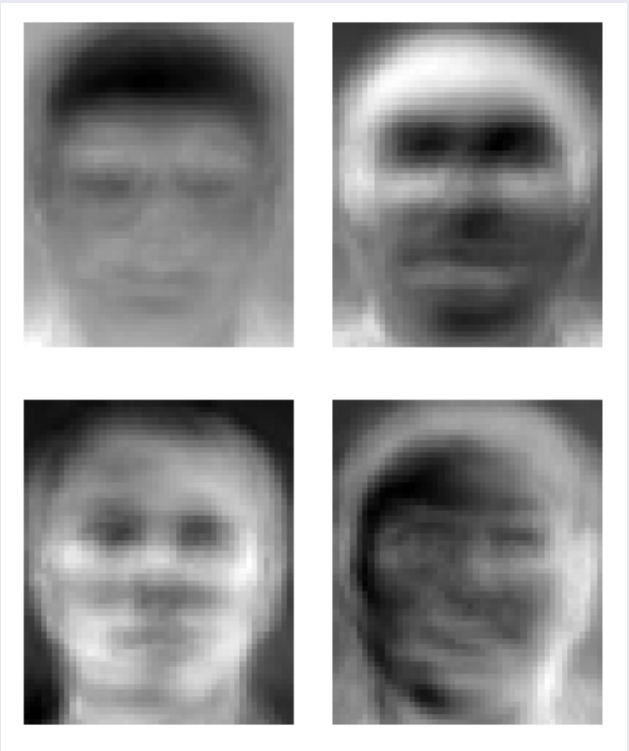
- Articles within the same cluster are similar to each other.
- Articles in different clusters are different compared to articles from the same clusters.

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.



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.

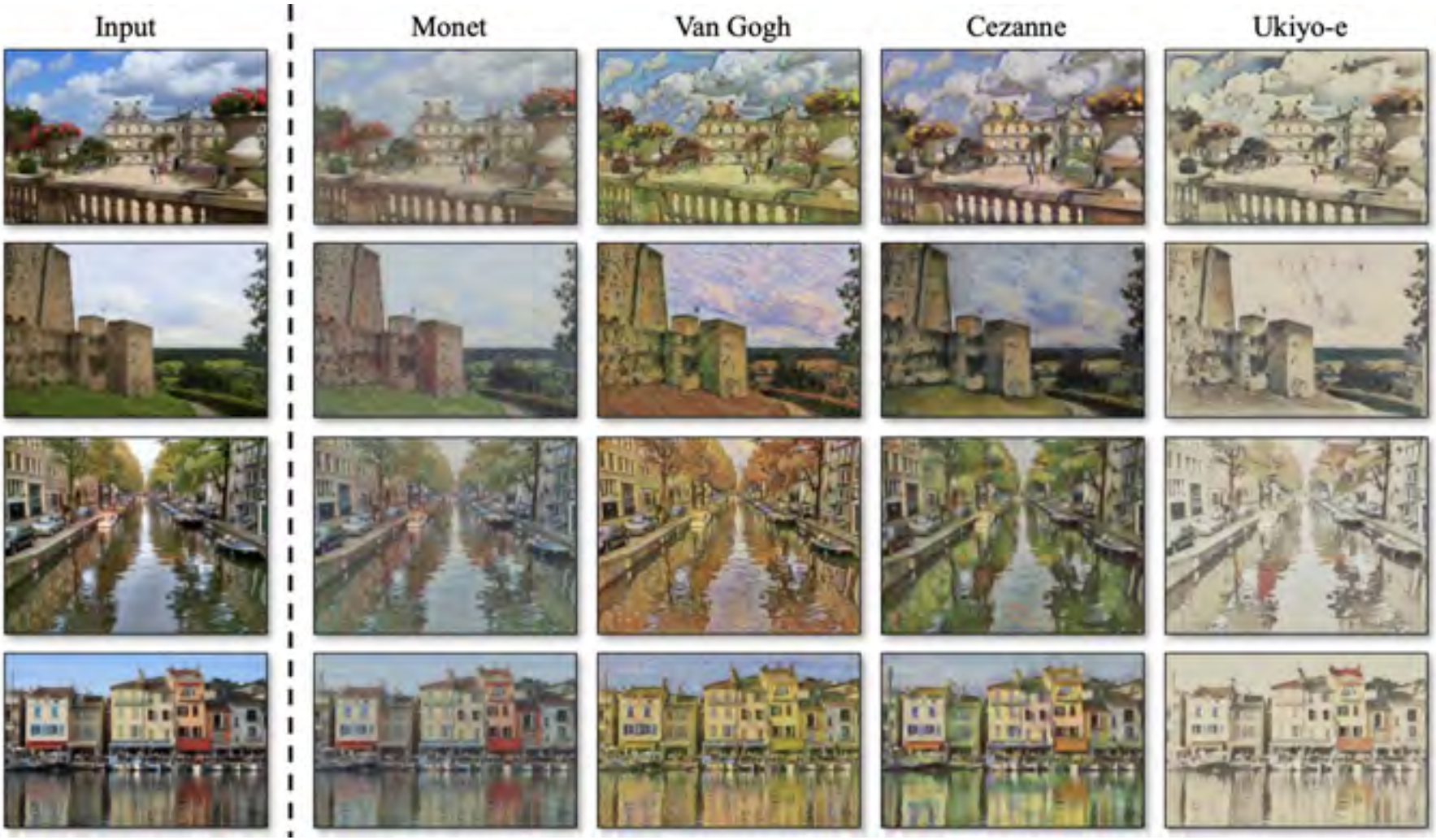
Generative Models

- 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.

"Machine learning has helped us move from a traditional computer language environment to one that utilizes natural languages. For example, if I speak Japanese to another Japanese person and they speak English, then the system can understand that English-to-Japanese sentence. We can also look at things like speech, but instead of trying to figure out what someone is saying, we use the linguistic rules that already exist. We can take a sentence like "I love you" and replace it with the sentence "I love you", and we can still understand the meaning of the sentence."

Text generated by OpenAI.

Can also generate images.



From [8].

Generated images. Top from [5], bottom from [3].

Outline

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

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

[7] 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.

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