MLISP: Machine Learning in Signal Processing

Lecture 1

Prof. V. I. Morgenshtern

Scribe: E. Sperschneider

Agenda:

1. Examples of machine learning problems and basic terminology
2. The common task framework
3. Other machine learning applications

1 Examples of machine learning problems and basic terminology

Machine learning (ML) is a collection of methods that allow us to automatically learn from data. Typically we have a **model** (statistical model) that describes our prior knowledge about the problem at hand. The model is parameterized by some parameters; these are variables, whose values are unknown or uncertain. An **algorithm** (statistical procedure) is used to process the **data** and estimate the unknown parameters the best we can. At this point, the calibrated model becomes useful for making **predictions** (inference, decisions) in the real world.

Machine learning is profoundly changing the landscape of science and technology. Every one of us uses machine learning algorithms every day.

1.1 Example: spam filter

Predict `spam` or `email` based on the statistics of words in the message.

- **Dataset**: 4601 email documents, each marked as `spam` or `email`
- **Model**: Certain individual words are very common in spam emails, other words are very uncommon in spam emails.
- **Algorithm**: Collect statistics on relative frequencies of 57 most common words in emails:

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

Based on this aggregate statistics we can come up with the following reasonable rule:

$$\text{if } \%\text{george} < 0.6 \text{ and } \%\text{you} > 1.5 \text{ then spam, else email.}$$

Another rule might be:

$$(0.2 \cdot \%\text{you} - 0.3 \cdot \%\text{george}) > 0 \text{ then spam, else email.}$$

- **Prediction**: For a new message collect the relative statistics of the 57 selected words. Use one of the rules above and make the decision if the message is spam or email.

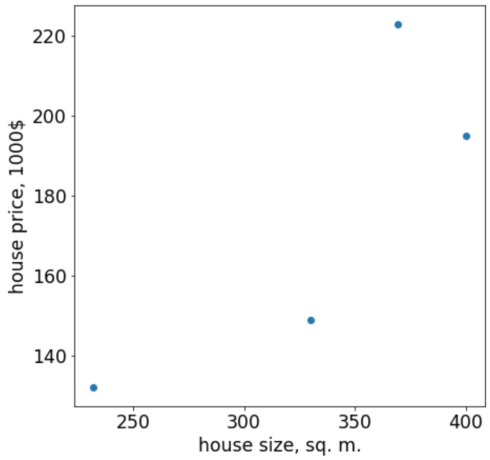
1.2 Example: house price

The goal is to predict the price of a house based on its size.

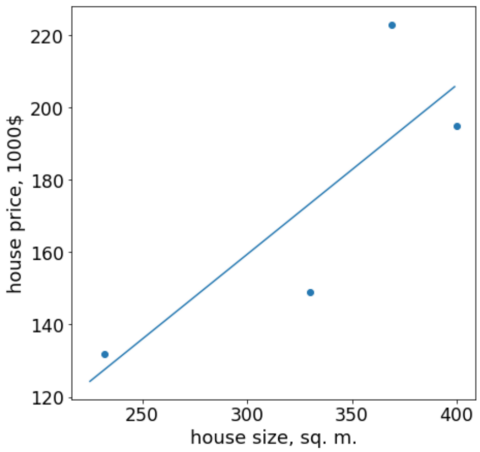
- **Dataset**: Table of house prices and house sizes

Living area m^2	Price(1000$)
195	400
149	330
223	369
132	232
...	...

we can conveniently plot this dataset on a scatter plot



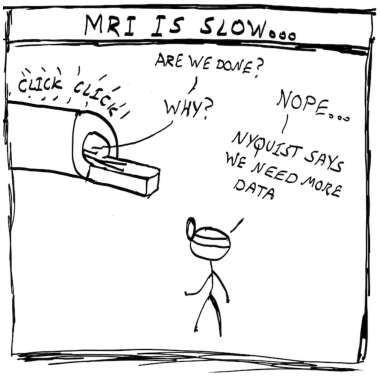
- **Model**: observe that the points on the scatter plot can be reasonably accurately approximated by a straight line. We can use linear regression to fit the line:



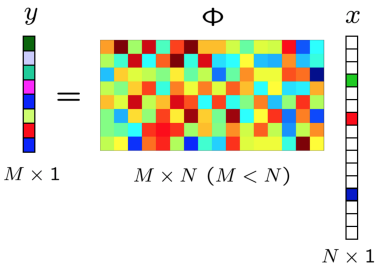
- **Algorithm**: Linear regression.
- **Prediction**: For a new house of a certain size we can now use the regression line to make the prediction.

1.3 Example: compressed sensing

The goal is to reconstruct the image of the human organs based on as few as possible linear measurements.

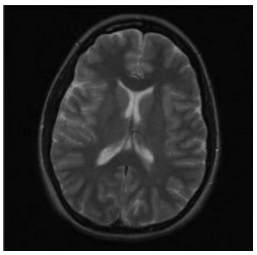


- **Dataset**: A set of linear measurement about an image:



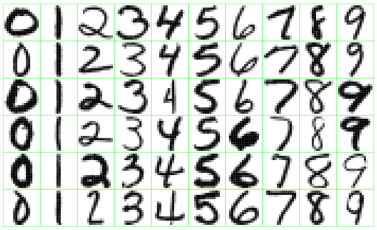
- **Model**: All natural images are very sparse in the wavelet basis.

- **Algorithm**: Linear regression with l1 penalty term.
- **Prediction**: Reconstruction of the image



1.4 Example: hand-written digit recognition (MNIST)

The goal is to automatically recognize hand-written digits



- **Dataset**: A large collection of pictures of hand-written digits together with correct labels.
- **Model**: Many different models can be used here. Consider the simplest. Assume that we just need to distinguish between images of 0's and 1's, i.e. no other digits. We may attempt to fit the logistic regression:

$$\text{if } \sigma(\langle \mathbf{x}, \alpha \rangle) > 1/2 \text{ then } 0, \text{ else } 1.$$

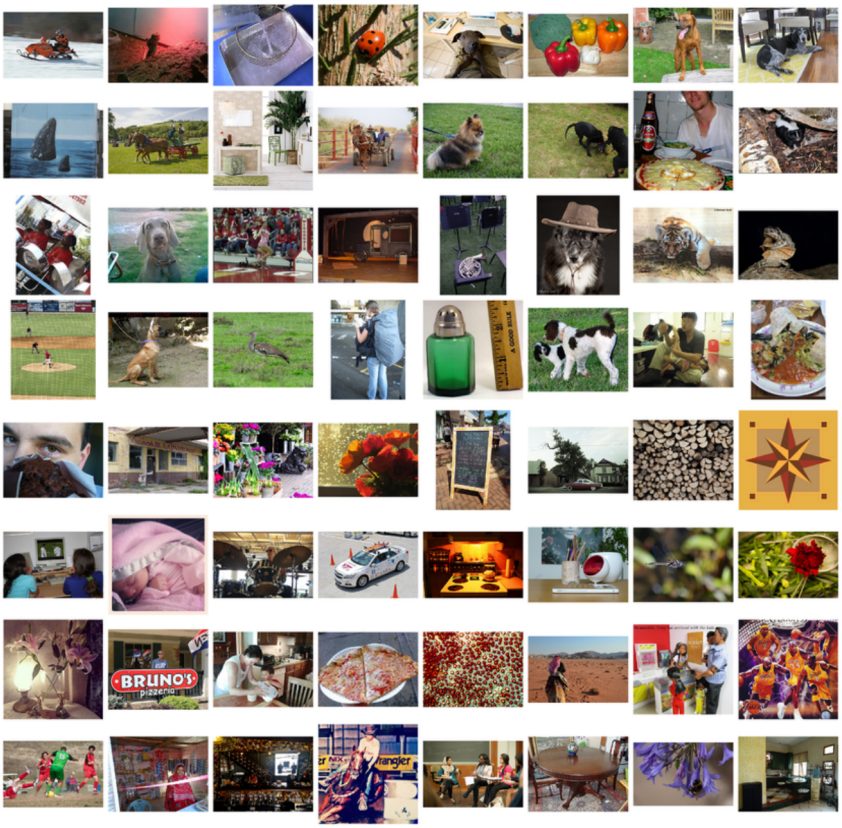
  Above, $\mathbf{x}$ is the vectorized image to be classified, α is the filter to be learned, and $\sigma = 1/(1+\exp(-x))$ is the logit function.
- **Algorithm**: Adjust α so that the error on the training set is minimized.
- **Prediction**: Classify the new image.

2 The common task framework

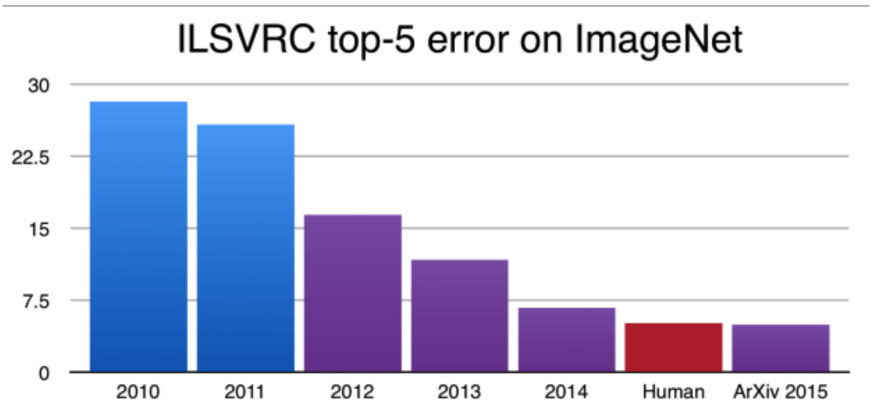
MNIST problem is the prototypical machine learning problem. Significant progress in machine learning has been accomplished through the common task framework. Here is how it works. First, a small team of researchers creates a sufficiently large and carefully organized dataset. In case of MNIST this dataset contains 70000 examples of hand-written digits. The dataset is divided into two parts: 60000 examples for the training set and 10000 examples for the test set. The dataset is published and the competition is organized. The competing researchers are allowed to use the

training dataset to create new algorithms. The performance of the algorithms is evaluated on the test set. Over the years better and better solutions are invented and significant and reproducible progress is being made. Many machine learning techniques have been tried on the MNIST problem: logistic regression (as discussed), support vector machines (SVMs), regression trees and others.

Currently the best results on most image recognition problems are obtained via Deep Learning. This technology allows to solve image recognition problems that are much more complex than MNIST with super-human precision. Consider the ImageNet competition.



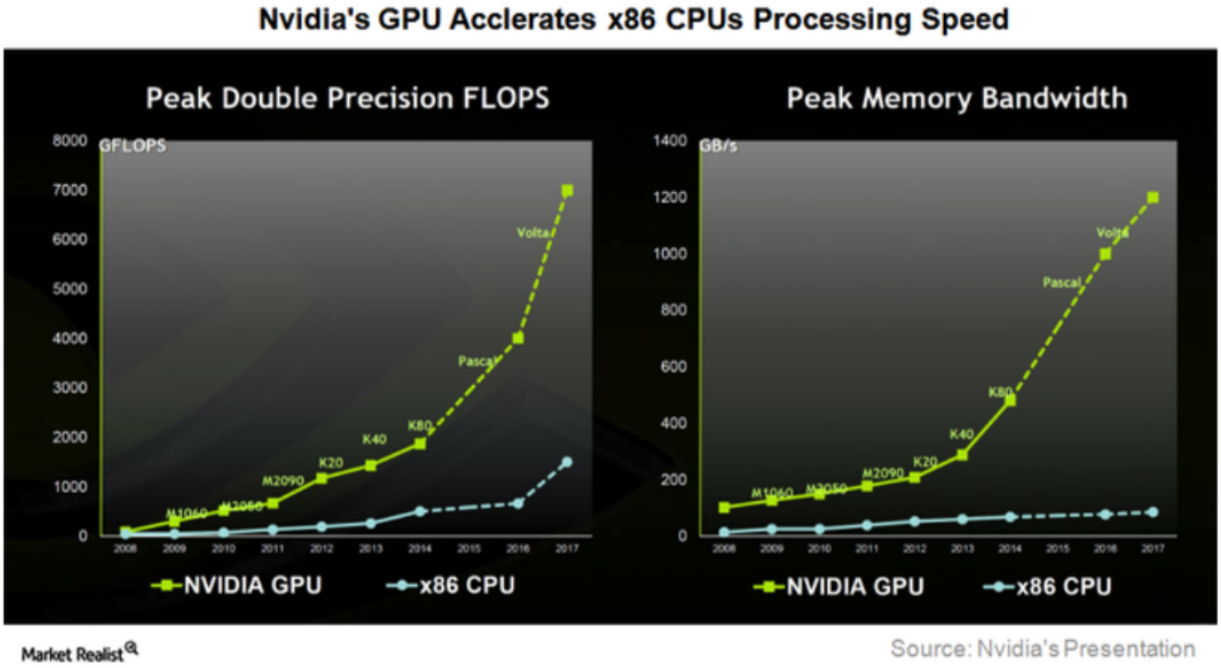
The goal is to list the objects that are depicted in the image. Here is the progress on this tasks over the years:



What are the ingredients for this spectacular success?

- The common task framework:
  - talent
  - large and clean datasets
  - reproducible research
- Cheap computational resources: GPUs

To support the last point, here is the progress in personal computing over the last decade:



3 Other machine learning applications

These include:

- Netflix prise
- Google's PageRank algorithm
- Google translate
- Siri: speech understanding and speech generation
- Shazam song recognition algorithm
- AlphaGo