Machine Learning in Practice: a Crash Course

Lecture 1: Introduction to ML

Conan Hu

Course Information

- Prerequisite:
  - Basic math is preferred: Linear algebra, analysis, probability theory

- Course textbook: No textbook is required. (Other materials are available at the course web page)

- Objective:
  - Basic understandings of fundamental knowledge of ML
  - Basic ability to use some ML techniques to solve real world problems.

Why Take This Course?

- This is different from many courses in the web.
- This is about fundamental knowledge, basic concepts, best practices according to my experience, etc.
- Not much about theoretical parts

What is machine learning?

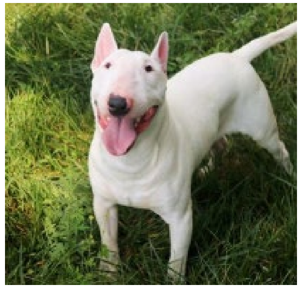
- Machine learning is the study of computer systems that improve their performance through experience (mostly, data)

- In machine learning, we study two types of problems:
  - Supervised learning
  - Unsupervised learning

The first kind of problems

Cat

Dog



?

The first kind of problems

30

28

18

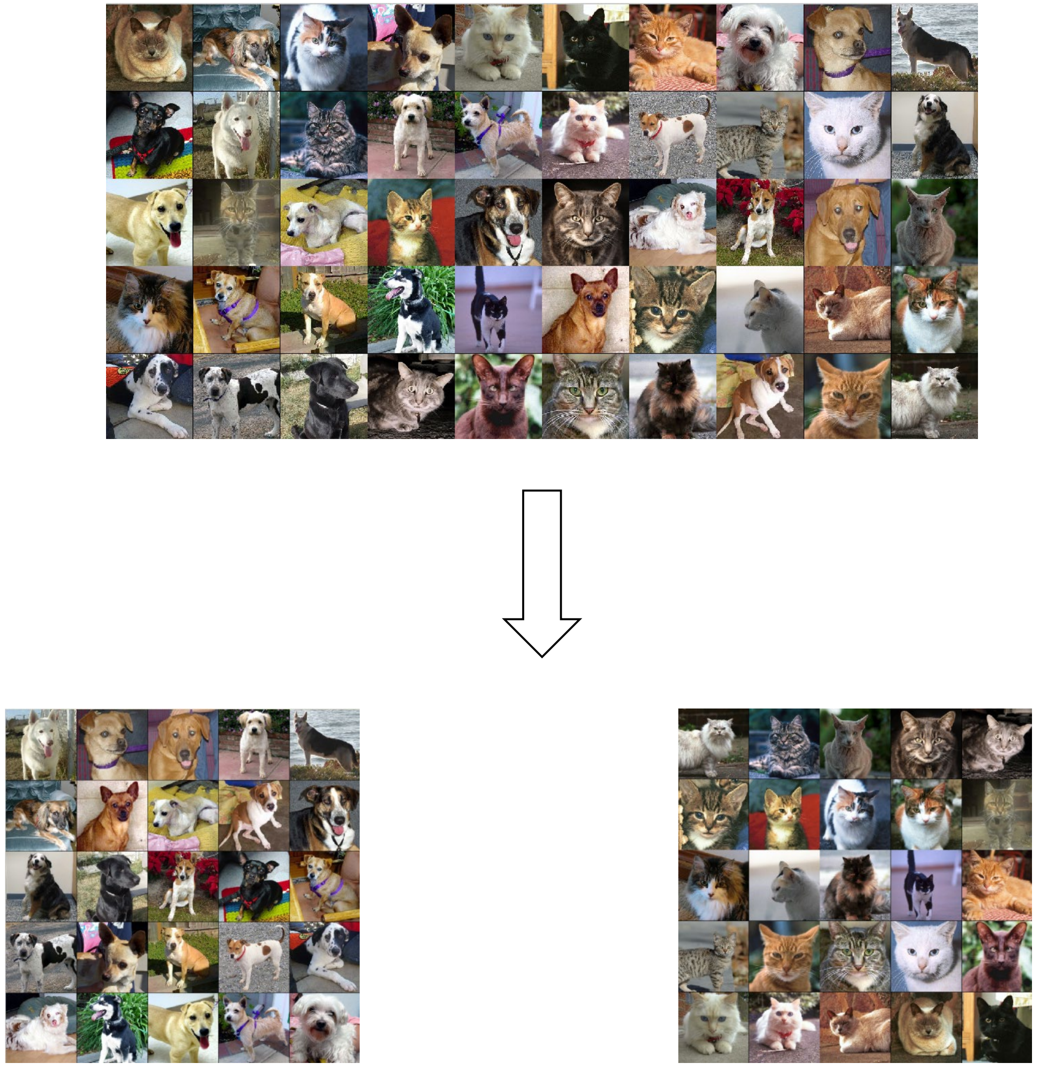
14

57

… …

?

The second kind of problems



Two kinds of problems

- Supervised learning **vs.** Unsupervised learning

- What are the differences?

- Supervised learning
  - Goal: learn a mapping from inputs $\boldsymbol{x}$ to outputs y
  - Training data: a labeled set of input-output pairs

  - Classification (Categorization, Decision making…)
    - y is a categorical variable
  - Regression
    - y is real-valued

Two kinds of problems

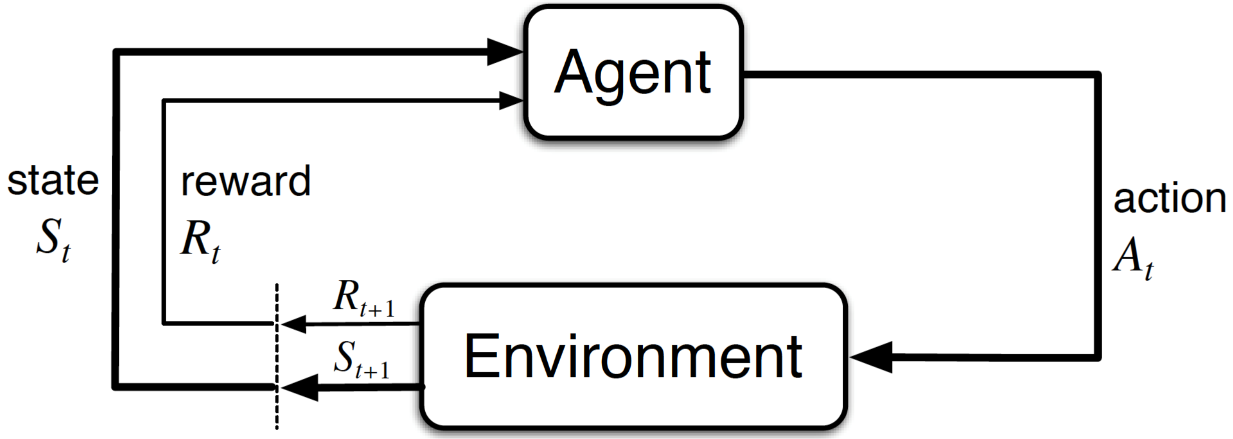
- What are the differences?
- Supervised learning **vs.** Unsupervised learning

- Unsupervised learning
  - We are only given inputs $\boldsymbol{x}$
  - Goal: find “interesting patterns”
  - Much less well-defined problem

  - Discovering clusters, Clustering
  - Discovering latent factors
    - Dimensionality reduction, Topic modeling

Two kinds of problems

- What are the differences?
- Another popular problem: reinforcement learning
  - It is a supervised learning scenario
  - No desired classification/regression label is given
  - The only teaching feedback is that the result is right or wrong.
  - This is useful for learning how to act or behave when given occasional reward or punishment signals.



Question?

Focus of This Course

- What are the typical ML **problems**?
  - Supervised Learning
    - Regression
    - Classification (decision making)
  - Unsupervised Learning
    - Clustering
    - Dimension reduction
  - Maybe a little bit about Reinforcement Learning
- What are the basic ML **methods/algorithms**?
- Practical topics about ML:
  - Typical pipeline for ML
  - Popular Python tools for ML: pandas, sklearn, matplotlib, etc

Basic Concepts of Supervised Learning

- Sample, example, instance

- Features, representations, predictors
  - $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_n$
- labels, targets, pattern class, class
  - $y_1, y_2, \cdots y_c$
- Training data
  - $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots (\boldsymbol{x}_n, y_n)$
- Model, classifier, regressor
  - f
- Test data
  - $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots (\boldsymbol{x}_n, y_n)$
- Training error & test error

Question?

When should we use ML?

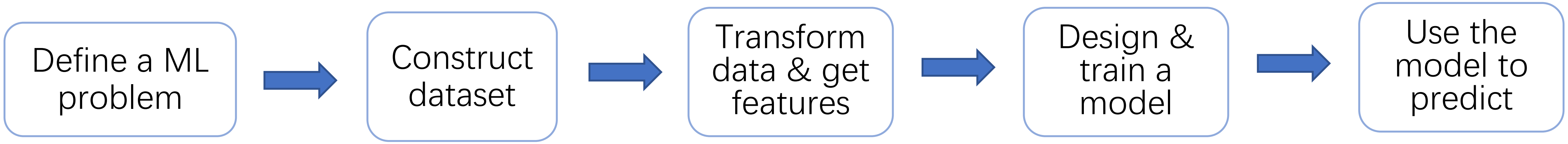
- First, we need to articulate the problem.
    - What is the problem we are facing?
    - Is this a typical ML problem? If it is, then what kind?

- We should be relatively easy to get the data.

- There should be some “patterns” in the data.

- Some other existing approaches are not good enough.
    - E.g. use machine learning to sort arrays may not be a good idea

What is a typical pipeline of ML

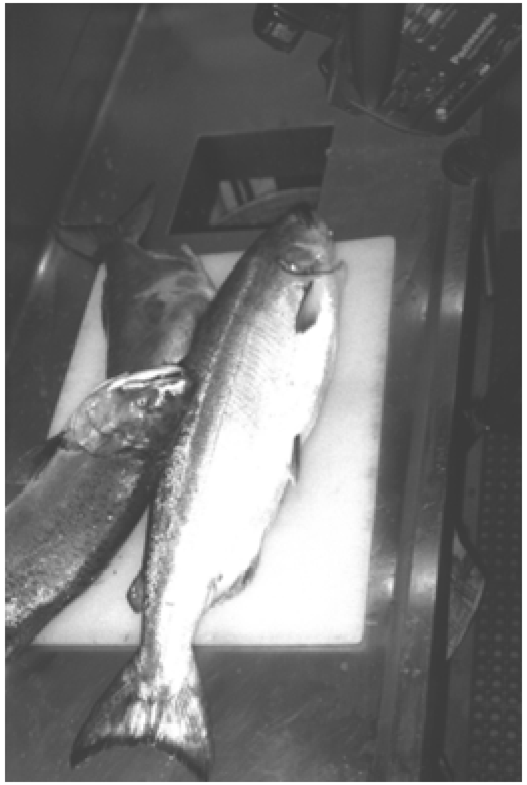


Define a ML problem → Construct dataset → Transform data & get features → Design & train a model → Use the model to predict

- It is often an iterative process
- Which step(s) is the most important step(s)?
  - Actually it depends. For performance, data & feature is often the most important one.

A toy example

- Fish Classification: Salmon v. Sea Bass

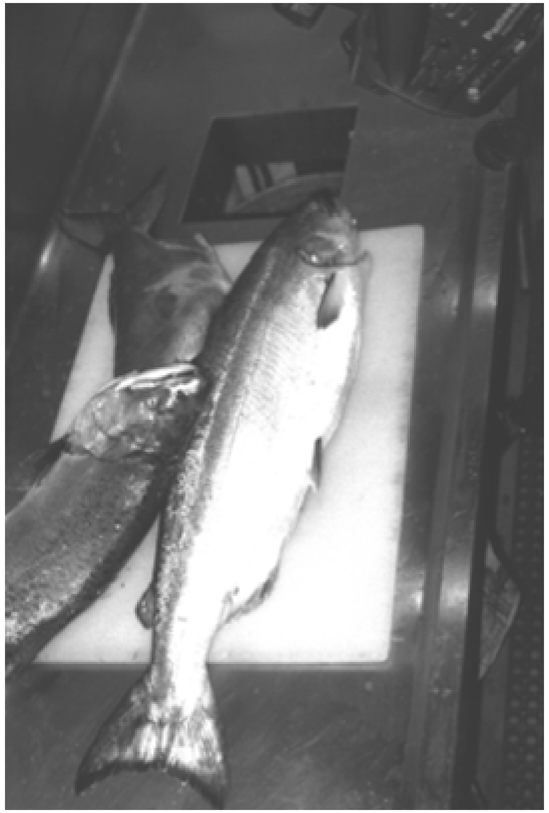


Fish Classification: Salmon v. Sea Bass

- 1. Define a ML problem
  - Articulate your problem
    - The goal is to classify an image of fish
    - This is a supervised and classification problem
    - There should be only one fish in the image
  - What are the labels and where are they from?
    - Simple: one fish is salmon or sea bass
    - It can be labelled by some experienced fishers.
  - Determine obtainable inputs
    - The photos of fish
    - Design features: will talk about this later
  - What is the metric, i.e. the evaluation of the goodness of a model?
    - Accuracy: the percentage of correctly classified samples.

Fish Classification: Salmon v. Sea Bass

- 2. Construct dataset
  - Get some photos of different fish
  - Do some preprocessing:
    - Separate touching or occluding fishes
    - Abandon awful photos
    - …
  - Ask some fishers to classify them

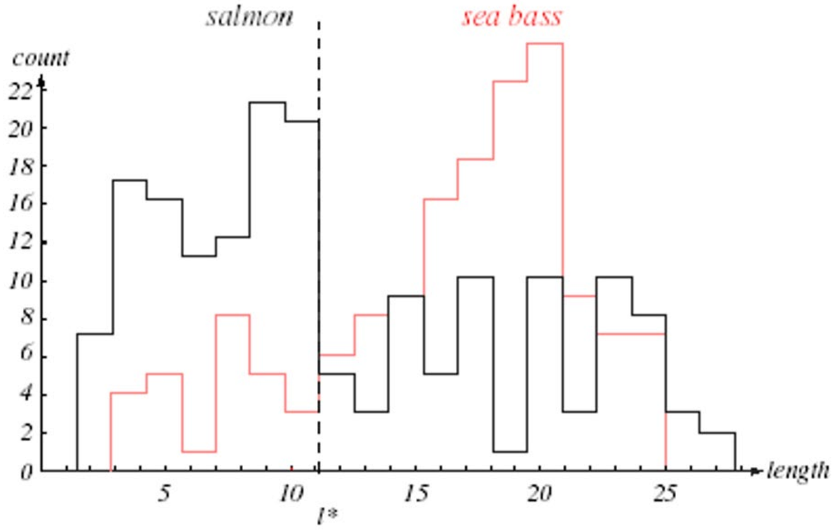


Fish Classification: Salmon v. Sea Bass

- 3. Transform data
  - First you need to determine the features
    - say the length of a fish in this problem
    - In practice we often use a lot more features
  - Then get the features from the data
    - Measure the length of the fish in photos
  - These features represent samples in machine learning models
  - Split data into train/test dataset

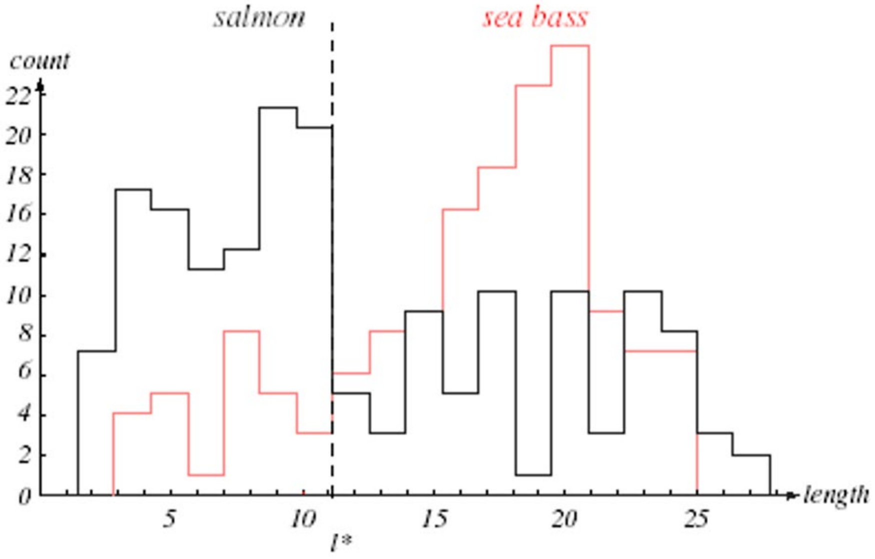
Fish Classification: Salmon v. Sea Bass

- 4. Design & train a model (Training)
  - If there is only one feature, what should be the classifier?
    - Histogram!
  - If there are more features available,
    we often use more complicated model



Fish Classification: Salmon v. Sea Bass

- 5. Use the model to predict
  - Evaluate the accuracy of this approach on some unseen photos in training to measure the performance of the ML system (Testing)
  - When there is a new image of fish coming, get the length from it and predict the label based on the model (Deploying & Serving)



Question?

Thanks and welcome to give me suggestions and feedbacks afterwards.