Machine learning: Classification

The problem

Classification

- Let us say I show you a picture

- Now you have to solve a multiple choice question:
  - [ ] Kitten
  - [x] Bunny
  - [ ] Puppy

What makes it a classification problem?

- **Inputs**:
    - A single input value: e.g. body temperature
    - A set of values: e.g. a series of medical tests
    - A picture, a video or a sound file
- **Outputs**: a discrete set of choices (mutually exclusive)
    - yes / no
    - kitten / bunny / puppy
    - COVID / not-COVID

How do I know that it works?

- **accuracy:** percentage of correctly classified inputs
  - Out of 100 examples I got right 90 --> 90% accuracy
- **false negatives**: the number of "Yes" answers that were classified as "No"
  - Important in cases where most of the answers are negative
  - How many cases where it was COVID, my classifier said No?
- **false positives**:
  - How many cases where my classifier said it is COVID, but turns out it wasn't?
- **other performance metrics** a number of other metrics that are combinations of these ones

The solution

Classification in one dimension

- Let us say that we have one information point: the temperature of a person T
  - We call this a **feature**
  - As this is a number, we call it a numerical feature
  - Other alternatives: boolean values (True/False), discrete values (Red/Green/Blue)
- We want to classify into **fever** / **not fever**
  - We can use a **threshold**
    - $T > 100.4$ fever
    - $T \leq 100.4$ not fever
    TODO: picture with number line here.

Building a classifier with knowledge engineering

- Possibly, we only need a threshold value

- We can obtain this by consulting an **expert**

- This process is called knowledge engineering or knowledge elicitation

Classification in two dimensions

- We want to classify houses into **family home** and **mansion**

- We have two features:
  - Price P in dollars
  - Area A in sq ft.

- We can agree that a house, no matter how large, is not a mansion if it costs less than 100,000. And even a $2M apartment in New York is not a mansion if it is less than 1000 sqft.

- A classifier must divide the plane defined by the two features into two spaces.
  - The boundary between the two classes will be a line (or lines)
    TODO: picture with the 2D classification

Classification in many dimensions

- How many features are in the kitten/bunny/puppy classification?

- Let us say 1000x1000 picture: 1M pixels, each with 3 RGB values

- The number of features is very large: 3 million

- There is no problem of **thinking mathematically** about a 3 million dimensional space
    - Although we cannot **draw** it, or **visualize** it
    - And our intuitions, which are anchored in 3D space, do not work very well.

- We will show how to do classification in high dimensional spaces when we talk about computer vision.

Building a classifier with learning

- Instead of creating some rules, we start with a series of examples where we know the answer
  - House A: 4000 sqft, $300k -> FamilyHome
  - House B: 1500 sqft, $200k -> FamilyHome
  - House C: 10000 sqft, $10M -> Mansion
  - House D: 5500sqft, $500k -> FamilyHome
  - House E: 3500sqft, $480k -> Mansion

Training data

- A collection of such examples is called **training data**

- In particular, it is **supervised** training data
  - It is supervised because we know the answers which we call **labels**

- If we have examples of houses, but not the labels, we call it **unsupervised**

- We will see that such data can be also useful, in other contexts.

Building a classifier with a linear separation

- The idea: use a straight separator between the different classes
    - If the data has 1-dimension, it is a point
    - If the data has 2-dimensions, it is a line
    - If the data has 3-dimensions, it is plane
    - With more dimensions, eg. n, the separator will be some kind of n-dimensional hyperplane… We cannot visualize it, but we can think about it.

Applications

Applications of classification

- There are many practical examples that involve classification:
  - Banking: grant a loan or not?
  - Consumers: should I buy or not this product?
  - Self driving car: should I break or not?
  - Medicine: COVID or not COVID?
  - ...

How do we solve classification with a program?

- Input: information

- Output: class

Pitfalls and dangers

- A classifier can only as good as its training data.

- The classifier would not find some kind of underlying truth: it will only replicate what is in the training data.

- If the training data has biases, the trained classifier will have the same biases!

Try it out + Homework

- 02-ML-01-Classification.ipynb

- Homework:
    - Modify the training set by changing the existing data and adding new data points. Check how the result of the training changes.

    - In one paragraph: let us assume that you are the first person who invented a learning based classification model. Propose an idea for a business.