Machine Learning

Additional material

2022-02-02

Machine Learning

- What happens if you want to teach a computer to do a task, but you're not entirely sure how to do it yourself? Or the problem is so complex that it's impossible for you to encode all the rules and knowledge upfront?
- Machine learning is the field of computer science that enables computers to learn without being explicitly programmed and builds on top of computational statistics and data mining.

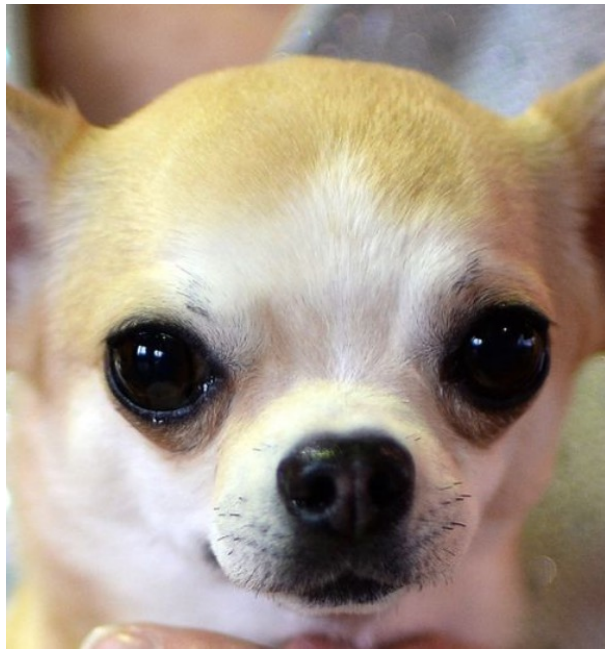
Supervised learning

- **Supervised learning** is when the computer is presented with input and output pairs, such as an image with a label (i.e. “cat”) and learns general rules to map the input to the output. Common tasks:
- If you are trying to predict whether an image is of a cat or a dog, this is a **classification** problem with discrete classes.
- If you are trying to predict the numeric price of a stock or other asset, this is a continuous output and can be framed as a **regression** problem.

Regression: What is the market value of the house?

900,000 USD vs. 100,000 USD

Classification: Is it a dog or a muffin?



source

Unsupervised learning

- **Unsupervised learning** occurs when computers are given unstructured rather than labeled data, i.e. no input-output pairs, and asked to discover inherent structure and patterns that lie within the data.
- One common application of unsupervised learning is clustering, where input data is divided into different groups based on a measure of “similarity”.
- For example, you may want to cluster your LinkedIn or Facebook friends into social groups based on how interconnected they are with each other.
- Unlike with supervised learning, the groups are not known in advance, and different measures of similarity will produce different results.

Semi-supervised learning

- **Semi-supervised learning** lies between supervised and unsupervised learning, where the input-output pairs are incomplete.
- Many real-world data sets are missing labels or have noisy, incorrect labels. Active learning, a special case of semi-supervised learning, occurs when an algorithm actively queries a user to discover the right output or label for a new input.
- Active learning is used to optimize recommender systems like the ones used to recommend new movies on Netflix or new products on Amazon.

Reinforcement learning

- **Reinforcement learning** is applied when computer programs are instructed to achieve a goal in a dynamic environment.
- The program learns by repeatedly taking actions, measuring the feedback from those actions, and improving its behavioral policy iteratively.
- Reinforcement learning is applied successfully in game-playing, robotic control, and other well-defined and contained problems, but is less effective with complex, ambiguous problems where rewards and environments are not well understood and quantified.

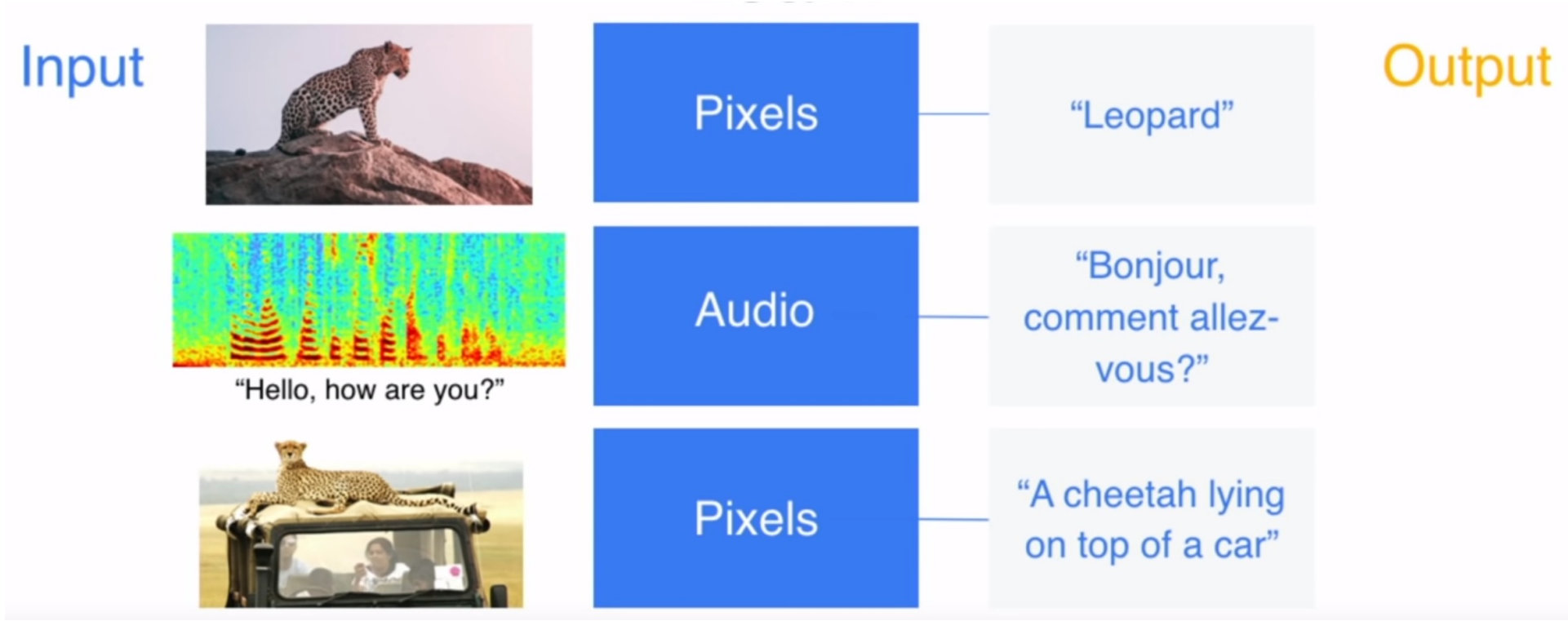
Ensemble methods

- **Ensemble methods** combine different machine learning models to produce superior results to any single model. Most successful applications of ML to real-world problems use ensemble approaches.
- Four broad categories:
  - **Bagging** entails training the same algorithm on different subsets of the data, e.g. Random Forest algorithm.
  - **Boosting** involves training a sequence of models, where each model prioritizes learning from the examples that the previous model failed on, e.g. XGBoost algorithm.
  - In **bucketing**, or “buckets of models”, you train multiple models for a given problem and dynamically choose the best one for each specific input.
  - In **stacking**, you directly combine the output of many models, using a combiner algorithm.

Deep Learning

- **Deep learning** is part of machine learning that builds algorithms using multi-layered artificial neural networks (ANNs).
- ANNs are only loosely representing how human brain works, as no one knows yet how it really works.
- Invented in the 1950s, ANNs gained popularity only in the last 5-10 years with significant advances in amounts of available data and computational power.
- The biggest successes in image classification and object detection and more recently in other tasks like speech recognition, machine translation, text classification, text summarization, etc.
- E.g. promising startups like Clarifai employ deep learning to achieve state-of-the-art results in recognizing objects in images and video for Fortune 500 brands.

Deep Learning task examples



Deep Learning vs Machine Learning

- Is Deep Learning (ANNs) better than a more traditional Machine Learning algorithms?
- Deep Learning gives state-of-the-art results in many NLP/NLU, voice, and Image (medical/satellite/natural) or video related tasks.
- People successfully used ANNs in other tasks like Stock price prediction, building recommendation systems, etc.
- However, for many real-world tasks traditional ML still outperforms Deep Learning due to limited amount of data or relatively simple data generating processes.

Kaggle vs Typical data science problems

Kaggle competitions

- By nature, competitions (with prize pools) must meet several criteria.
- Problems must be difficult. Competitions shouldn't be solvable in a single afternoon. To get the best return on investment, host companies will submit their biggest, hairiest problems.
- Solutions must be new. To win the latest competitions, you'll usually need to perform extended research, customize algorithms, train advanced models, etc.
- Performance must be relative. Competitions must crown a winner, so your solution will be scored against others'.

"Typical" data science

- In contrast, day-to-day data science doesn't need to meet those same criteria.
- Problems can be easy. In fact, data scientists should try to identify low-hanging fruit: impactful projects that can be solved quickly.
- Solutions can be mature. Most common tasks (e.g. exploratory analysis, data cleaning, A/B testing, classic algorithms) already have proven frameworks. There's need to reinvent the wheel.
- Performance can be absolute. A solution can be very valuable even if it simply beats a previous benchmark.
- Kaggle competitions encourage you to squeeze out every last drop of performance, while typical data science encourages efficiency and maximizing business impact.

What is an insight?

- Job of executive is to make decisions. Decisions based on data tend to be more accurate and less risky.
- The purpose of analytics is to provide insights.
- To be an insight and not a mere observation:
  - The information has to be new, relevant and non-trivial.
- Good insights should:
  - Typically focus on consumer behavior
  - Quantify causality
  - Provide competitive advantage
  - Generate financial implication
  - Be action-able