Data visualization

COSC 480B

Reyan Ahmed

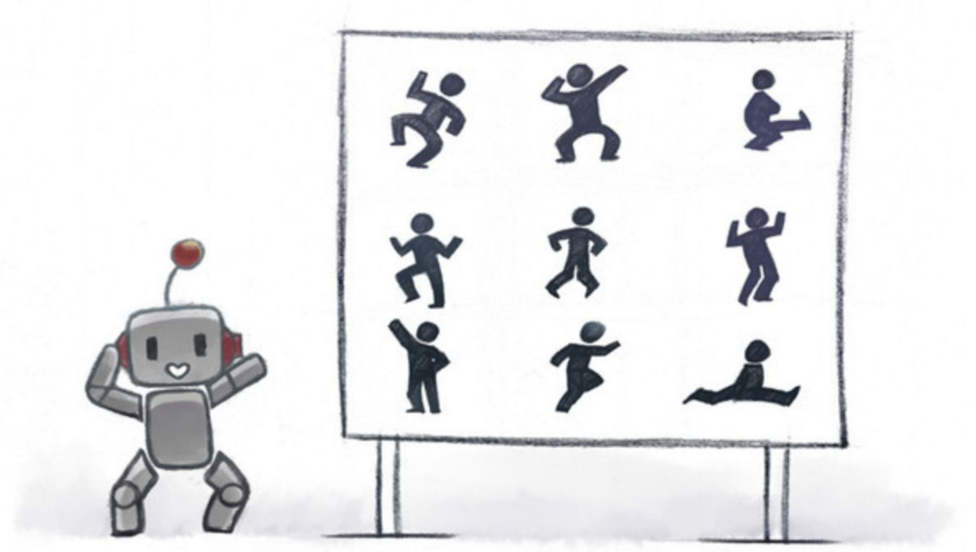
rahmed1@colgate.edu

Lecture 16

A machine-learning odyssey

Overview

- Machine-learning fundamentals
- Data representation, features, and vector norms
- Why TensorFlow



Machine learning fundamentals

Each pair of integers, when summed, results in an even or odd number. The input and output correspondences listed are called the ground-truth dataset.

Input		Output
$x_1 = (2, 2)$	$\rightarrow$	y_1 = Even
$x_2 = (3, 2)$	$\rightarrow$	y_2 = Odd
$x_3 = (2, 3)$	$\rightarrow$	y_3 = Odd
$x_4 = (3, 3)$	$\rightarrow$	y_4 = Even
...		...

Machine learning fundamentals

This table reveals the inner logic behind how the output response corresponds to the input pairs.

m \ n	Even	Odd
Even	$2m + 2n =$ $2(m + n)$ Even	$2m + (2n + 1) =$ $2m + 2n + 1$ Odd
Odd	$(2m + 1) + 2n =$ $2m + 2n + 1$ Odd	$(2m + 1) + (2n + 1) =$ $2(m + n + 1)$ Even

Machine learning fundamentals

An ML approach to solving problems can be thought of as tuning the parameters of a black box until it produces satisfactory results.

Machine learning fundamentals

Exercise 1

Suppose you've collected three months' worth of stock market prices. You'd like to predict future trends to outsmart the system for monetary gains. Without using ML, how would you go about solving this problem? (As you'll see in following lectures, this problem becomes approachable with ML techniques.)

Machine learning fundamentals

ANSWER

Believe it or not, hard-designed rules are a common way to define stock market trading strategies. For example, an algorithm as simple as “if the price drops 5%, buy some stocks” is often used. Notice that there’s no machine learning involved, just traditional logic.

Machine learning fundamentals

Suppose you're trying to bake desserts in an oven. If you're new to the kitchen, it can take days to come up with both the right combination and perfect ratio of ingredients to make something that tastes great. By recording recipes, you can remember how to quickly repeat the dessert if you happen to discover the ultimate tasty meal.

Machine learning fundamentals

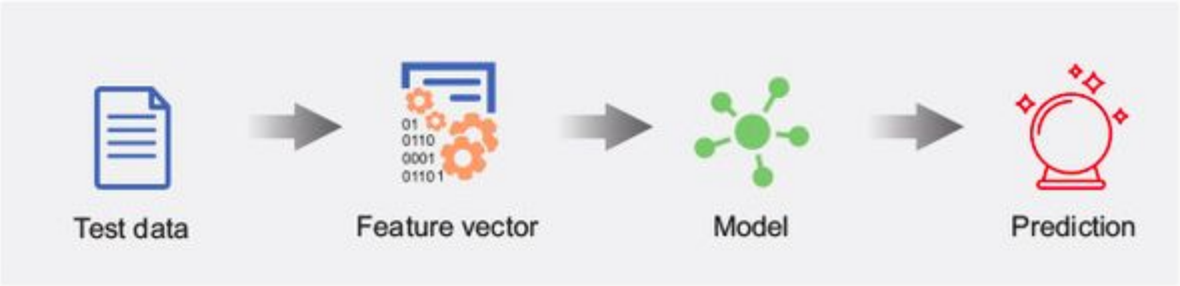
Similarly, machine learning shares this idea of recipes. Typically, we examine an algorithm in two stages: learning and inference. The objective of the learning stage is to describe the data, which is called the feature vector, and summarize it in a model. The model is our recipe. In effect, the model is a program with a couple of open interpretations, and the data helps disambiguate it.

Machine learning fundamentals

A feature vector is a practical simplification of data. You can think of it as a sufficient summary of real-world objects into a list of attributes. The learning and inference steps rely on the feature vector instead of the data directly.

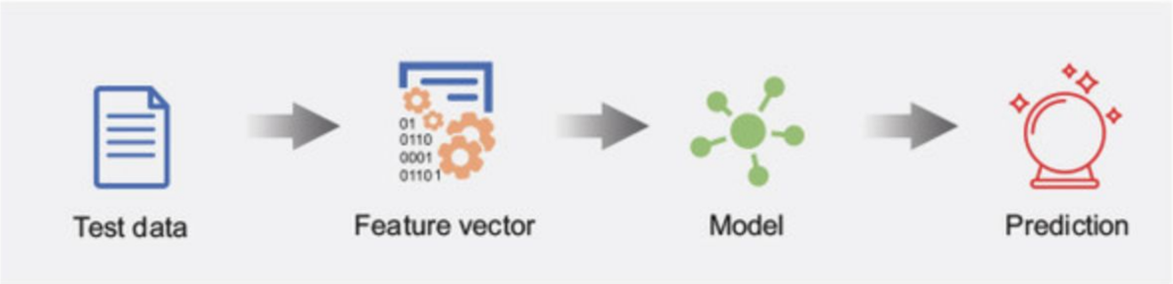
Machine learning fundamentals

The inference approach generally uses a model that has already been either learned or given. After converting data into a usable representation, such as a feature vector, it uses the model to produce intended output.



Machine learning fundamentals

The inference approach generally uses a model that has already been either learned or given. After converting data into a usable representation, such as a feature vector, it uses the model to produce intended output.

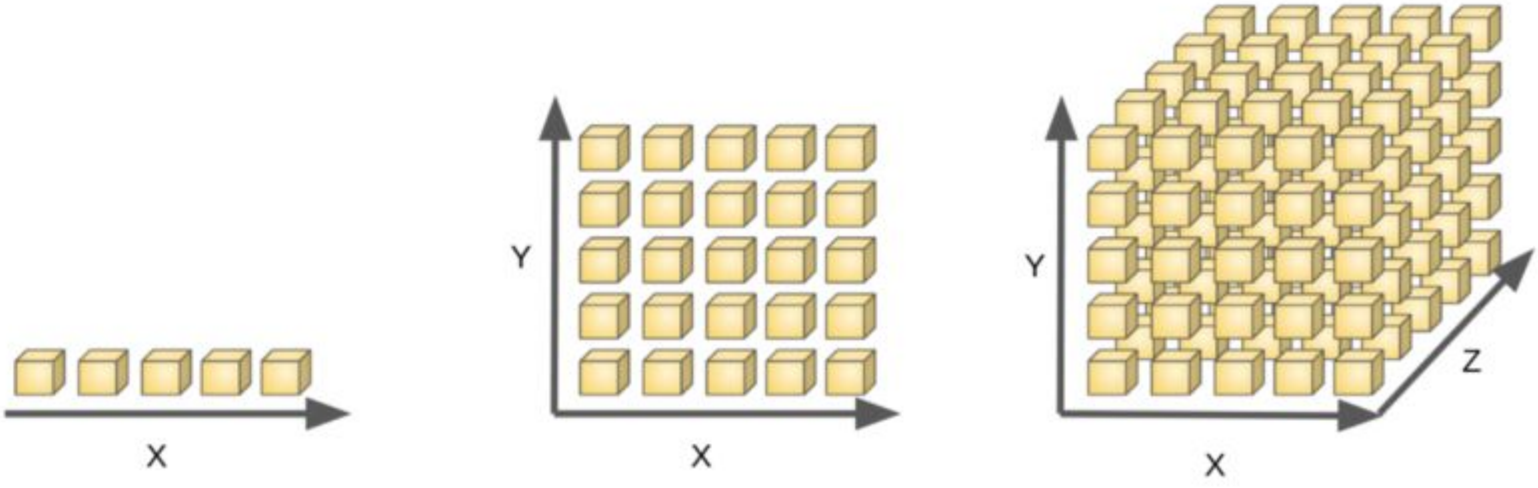


Data representation and features

Feature engineering is the process of selecting relevant features for the task.

Data representation and features

Curse of dimensionality



Data representation and features

- If the size of feature vector is too large then training will be costly
- If it is too small, then it may lack necessary information
- For example, in the car detection problem, color of the image is not important
- Only the shape and texture matters
- Hence, we can consider grayscale image so the machine does not spent time learning colors

Data representation and features

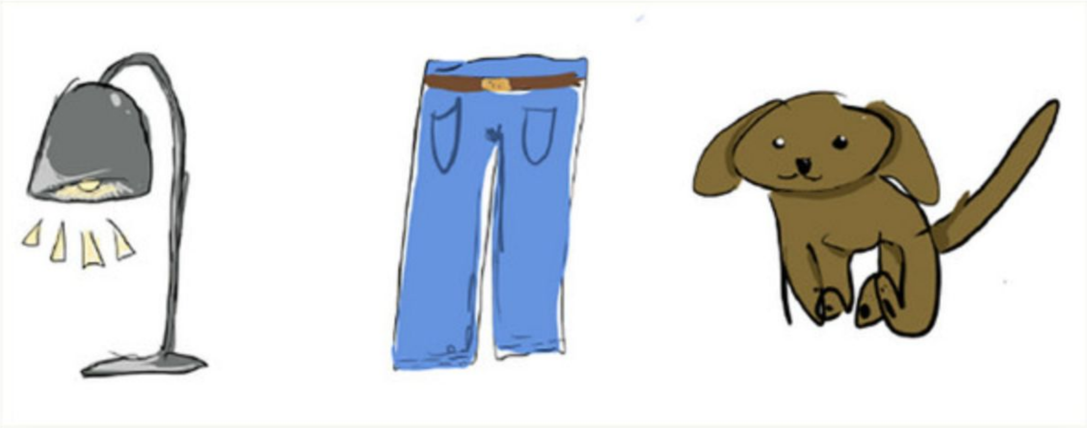
Exercise 2: A robot is trying to fold a shirt. What are good features of the shirt to track?

Data representation and features

Answer: The width, height, x-symmetry score, y-symmetry score, and flatness are good features to observe when folding clothes. Color, cloth texture, and material are mostly irrelevant.

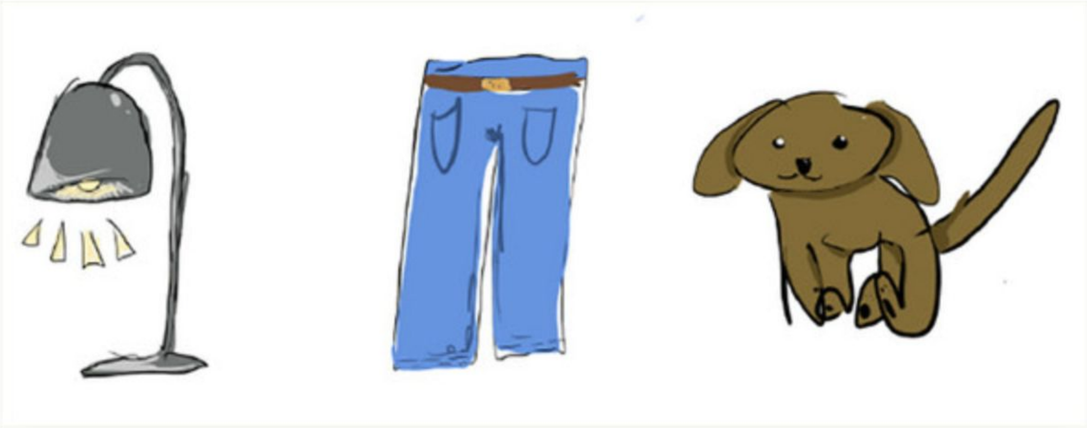
Data representation and features

Exercise 3: Now, instead of detecting clothes, you ambitiously decide to detect arbitrary objects; the following figure shows some examples. What are some salient features that can easily differentiate objects? Here are images of three objects: a lamp, a pair of pants, and a dog. What are some good features that you should record to compare and differentiate objects?



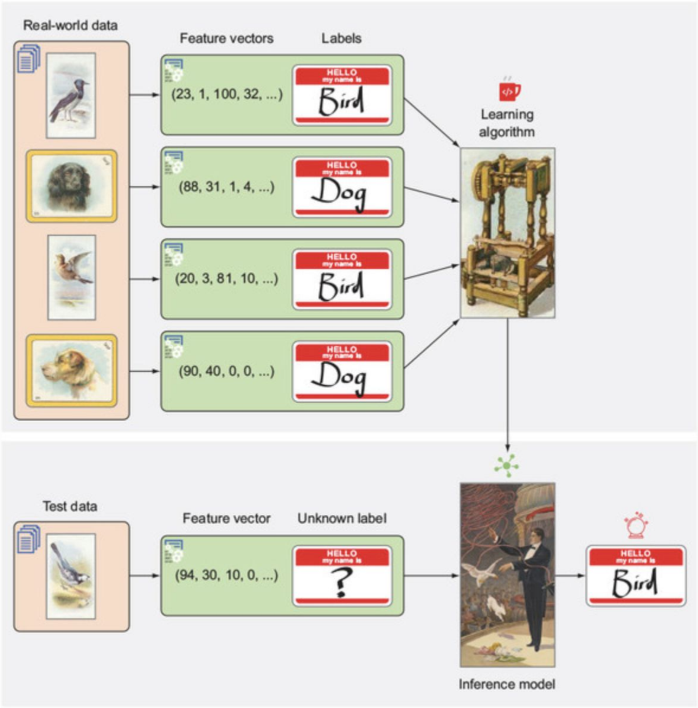
Data representation and features

Answer: Observing brightness and reflection may help differentiate the lamp from the other two objects. The shape of pants often follows a predictable template, so shape would be another good feature to track. Lastly, texture may be a salient feature to differentiate the picture of a dog from the other two classes.



Data representation and features

Feature vectors are used in both learning and inference



Distance metrics

Let's say we have two feature vectors, $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$. The Euclidian distance $||x - y||$ is calculated by

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}$$

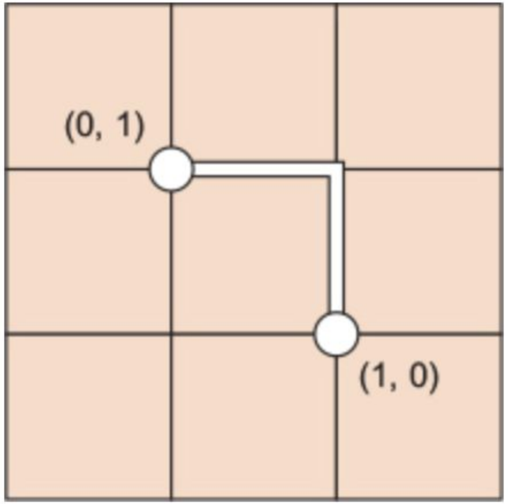
Distance metrics

For example, the Euclidean distance between (0, 1) and (1, 0) is

$$
\begin{aligned}
& \|(0,1) - (1,0)\| \\
= & \|(-1,1)\| \\
= & \sqrt{(-1)^2 + 1^2} \\
= & \sqrt{2} = 1.414\ldots
\end{aligned}
$$

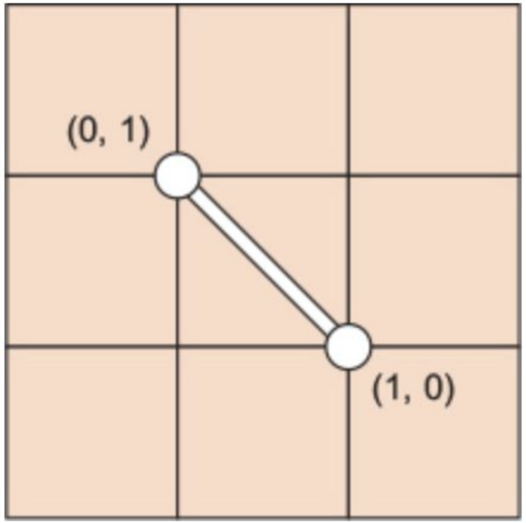
Distance metrics

The L1 distance is also called the Manhattan distance (also referred to as the taxicab metric), because it resembles the route of a car in a grid-like neighborhood such as Manhattan. If a car is traveling from point (0,1) to point (1,0), the shortest route requires a length of 2 units.



Distance metrics

The L2 norm between points (0,1) and (1,0) is the length of a single straight-line segment between both points.



Distance metrics

- L0 norm: counts the total number of nonzero elements of a vector
- L1 norm: summation of absolute values in each dimension, also known as Manhattan distance
- L2 norm: Euclidean distance
- LN norm: $(\Sigma|xn|^N)^{1/N}$
- L-infinity norm: largest magnitude among each element

Types of learning

- Supervised learning:
  - A set of dataset is given
  - We have to find a weight that minimizes the loss on that data

$$\theta^* = \arg\min_\theta Cost(\theta|X)$$

$$\text{where} \quad Cost(\theta|X) = \sum_{x \in X} \|g(x|\theta) - f(x)\|$$

Types of learning

- Unsupervised learning:
  - Clustering: groups together different data points based on the distance, each group represents a class
  - Dimensionality reduction: reduce the amount of features, specially the ones that are not that important

Types of learning

- Reinforcement learning
  - An agent takes actions, and gets rewards
  - Form the amount of rewards, the agent learns which actions are good
  - Exploration vs. exploitation: suppose the agent determined a set of good actions. Now there is a tradeoff between continuously apply that action or looking for even better alternatives
  - The agent interacts with environment, and the environment changes based on actions
  - Environments are represented by states, the agent maximizes the expected reward of each states

Tensorflow

- A machine learning framework
- Has automatic differentiation capabilities, hence easy to develop new models
- Has a visualization to analyze a model, known as tensor board
- Segments can be reused

Tensorflow

Example of TensorBoard in action