Section 2: Clustering, Matrix Review

Adhyyan Narang

April 8, 2022

Introduction

The Scientific Revolution

A couple examples of the growth of human power in the last 500 years:

1. In 1500, there were 500 million Homo Sapiens. Today there are 7 Billion.
2. Value of goods and services in 1500 was 250 Billion in today's dollars. Now, 60 Trillion.
3. In 1500, humans consumed 13 Trillion calories/day. Now we consume 1500 trillion.

Short timeline of discoveries

1543: Heliocentric model of Astronomy
1545: Complex numbers
1637: Rene Descartes' discovered the scientific method.
1675: Anton van Leeuwenhoek discovered micro-organisms in pond water
1675: Calculus
1676: The first measurement of the speed of light
July 16, 1945: Atomic Bomb
1969: Landed on the moon

"The unreasonable effectiveness of mathematics in the Natural Sciences (Eugene Wigner)"

> *The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve.*

Where ML fits in

"Sapiens (YN Harari)"
"The unreasonable effectiveness of data (Alon Halevy et.al)"

> *However, scholars who attempted to reduce biology, economics and psychology to neat Newtonian equations discovered that these fields have a level of complexity that makes such an aspiration futile.*

Machine Learning Need a new toolset of mathematics that can work with and utilize BIG data to understand the world and create technologies.

ML workflow: Role of linear algebra

Real world task involving data $\longrightarrow$ Optimization Problem $\longrightarrow$ ML Model

Canonical Optimization Problem

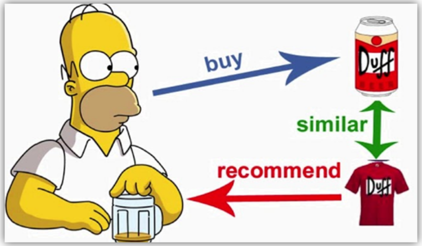
$$\min_{x \in \mathcal{X}} f(x).$$

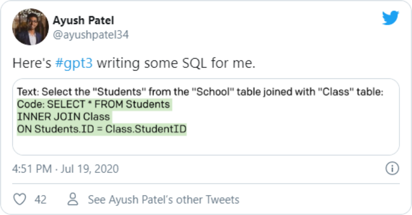
Terminology:

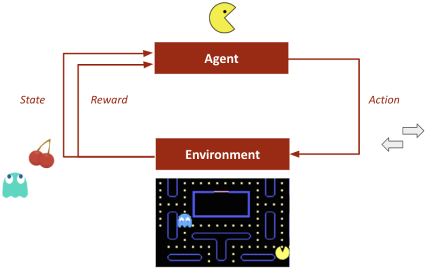
- x: Decision/choice/optimization variable
- $f(x)$: Objective/Loss function; in ML will often depend on data Z; write as $f(x; Z)$.
- $\mathcal{X}$: Constraint set

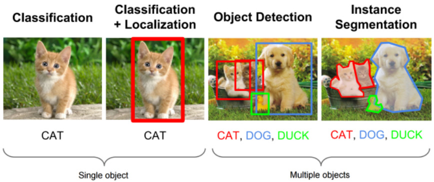
Central goal of this course Show that many interesting opt problems are framed using language from LA and solved using techniques from LA.

Applications of ML









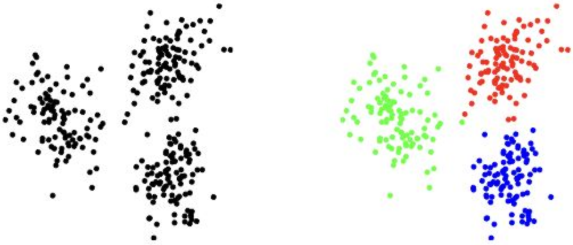
History of Machine Learning

1763: Bayes' Theorem
1805: Least Squares
1936: The Universal Turing Machine
1943: Artificial Neuron
1952: Arthur Samuel's Perceptron plays checkers
1967: Nearest neighbor algorithm
1969: Minsky&Papert write "Perceptrons" on limitations of neural nets
1970s: AI Winter
1986: Backpropagation is used
1989: Reinforcement learning
1995: Random Forests, Support Vector Machines
2009: ImageNet is created
2012: AlexNet CNN for vision
2016: AlphaGo
2020: GPT3 for language models
2021: AlphaFold 2 for Protein Structure Prediction

Part 1: Jupyter Notebook

Clustering: Our first ML Example

The goal Suppose we have N n-vectors $x_1 \dots x_N$. The goal of clustering is to group or partition the vectors (if possible) into k groups or clusters, with the vectors in each group close to each other.



Clustering objective

Notation:

- $G_j \subset \{1, \ldots, N\}$ is group j, for $j = 1, \ldots k$
- Cluster assignment c_i is group that x_i is in: $i \in G_{c_i}$
- Cluster Centers: $z_1, \ldots, z_k$

Optimization Problem:

$$\min_{z_1 \ldots z_k} \min_{c_1 \ldots c_N} \frac{1}{N} \sum_{i=1}^{N} \|x_i - z_{c_i}\|^2 \tag{1}$$

Sanity check question: If we increase k from 3 to 6 would the objective value reduce or increase?

Partitioning vectors given representatives

Simpler problem Suppose representatives $z_1, \ldots, z_k$ are given. Then, how do we assign vectors to groups, i.e., choose $c_1, \ldots, c_N$?

$$\min_{c_1 \ldots c_N} \frac{1}{N} \sum_{i=1}^{N} \|x_i - z_{c_i}\|^2$$

Easy solution
To minimize, choose c_i so that $\|x_i - z_{c_i}\|^2 = \min_j \|x_i - z_j\|^2$
i.e., assign each vector to its *nearest representative*

Choosing representatives given partition

Simpler problem 2 Given partition $G_1, \ldots, G_k$, how to choose representatives $z_1, \ldots, z_k$ to minimize J?

Solution:

1. Decompose J:

$$\min_{z_1 \ldots z_k} J = \min_{z_1 \ldots z_k} J_1 + \ldots + J_k =$$

   where $\qquad J_j = 1/N \sum_{i \in G_j} \|x_i - z_j\|^2$

2. so we choose z_j to minimize mean square distance to points in its partition
3. this is the mean (or centroid) of the points in the partition:

$$z_j = \frac{1}{|G_j|} \sum_{i \in G_j} x_i$$

4. alternating between these two steps gives the famous k-means algorithm!

K-means algorithm

Algorithm 4.1 k-MEANS ALGORITHM

given a list of N vectors $x_1, \ldots, x_N$, and an initial list of k group representative vectors $z_1, \ldots, z_k$

repeat until convergence

1. *Partition the vectors into k groups.* For each vector $i = 1, \ldots, N$, assign x_i to the group associated with the nearest representative.
2. *Update representatives.* For each group $j = 1, \ldots, k$, set z_j to be the mean of the vectors in group j.

Does it solve the Opt Problem (1) exactly?
No, it might find a local minima.
But everyone uses it anyway because it's really easy and often enjoys good empirical perf

Coding exercise: Parts (a) and (b)

Synthetic Data Often used as a good way to sanity check your algorithm and provide statistical guarantees on the algorithm.

Part 2: Matrix Review Problems

Matrix Vector Multiplication

Question Previously in K-means, we wanted to compute quantity $\bar{x}_j = \frac{1}{|G_j|} \sum_{i \in G_j} x_i$.
Given a matrix $A \in \mathbb{R}^{n \times N}$

$$A = \begin{bmatrix} | & | & \cdots & | \\ x_1 & x_2 & \cdots & x_N \\ | & | & \cdots & | \end{bmatrix}$$

find a vector w such that $Aw = \bar{x}_j$.

Match the columns

Consider a matrix $A \in \mathbb{R}^{m \times n}$.

Concept

- Null space
- Column space
- Rank
- dim(Null(A))

Description

- The dimension of the column space
- $\{b : Ax = b$ has a solution$\}$
- Number of LI rows in A
- The eigenspace corresponding to eigenvalue 0.
- Number of LI columns in A
- $n-$ Number of LI columns in A

VMLS 10.41: Kmeans as approx matrix factorization

Suppose we run the k-means algorithm on the N n-vectors $x_1, \ldots, x_N$, to obtain the group representatives $z_1, \ldots, z_k$. Define the matrices

$$X = \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix}, \quad Z = \begin{bmatrix} z_1 & \cdots & z_k \end{bmatrix}.$$

X has size $n \times N$ and Z has size $n \times k$. We encode the assignment of vectors to groups by the $k \times N$ clustering matrix C, with $C_{ij} = 1$ if x_j is assigned to group i, and $C_{ij} = 0$ otherwise. Each column of C is a unit vector; its transpose is a selector matrix.

(a) Give an interpretation of the columns of the matrix $X - ZC$, and the squared norm (matrix) norm $\|X - ZC\|^2$.
(b) Justify the following statement: The goal of the k-means algorithm is to find an $n \times k$ matrix Z, and a $k \times N$ matrix C, which is the transpose of a selector matrix, so that $\|X - ZC\|$ is small, i.e., $X \approx ZC$.

VMLS 11.3: Matrix cancellation

Suppose the scalars a, x, and y satisfy $ax = ay$. When $a \neq 0$ we can conclude that $x = y$; that is, we can cancel the a on the left of the equation. In this exercise we explore the matrix analog of cancellation, specifically, what properties of A are needed to conclude $X = Y$ from $AX = AY$, for matrices A, X, and Y ?

(a) Give an example showing that $A \neq 0$ is not enough to conclude that $X = Y$.
(b) Show that if A is left-invertible, we can conclude from $AX = AY$ that $X = Y$.
(c) Show that if A is not left-invertible, there are matrices X and Y with $X \neq Y$, and $AX = AY$.

VMLS 11.9 Push through identity

Suppose A is $m \times n, B$ is $n \times m$, and the $m \times m$ matrix $I + AB$ is invertible.

(a) Show that the $n \times n$ matrix $I + BA$ is invertible. Hint. Show that $(I + BA)x = 0$ implies $(I + AB)y = 0$, where $y = Ax$.

(b) Establish the identity

$$B(I + AB)^{-1} = (I + BA)^{-1}B.$$

This is sometimes called the push-through identity since the matrix B appearing on the left 'moves' into the inverse, and 'pushes' the B in the inverse out to the right side. Hint. Start with the identity

$$B(I + AB) = (I + BA)B,$$

and multiply on the right by $(I + AB)^{-1}$, and on the left by $(I + BA)^{-1}$.