Lecture 10: Tensors and low rank tensor decomposition.

In data analysis, matrices often represent correlation structure between 2 features.

To represent higher order structure, you need tensors.

e.g. k-grams. Given a document, and a list of possible words $w_1, \dots, w_n$, there are many statistics one cares about:

1). bag-of-words: how many times each word appears.

$v \in \mathbb{R}^n$

2). 2-grams: how many times any pair of words appear consecutively.

$M \in \mathbb{R}^{n \times n}$. e.g. w_i = "hello", w_j = "world"

M_{ij} = # times we see "hello world"

3). k-grams: how many times any sequence of k words appear consecutively.

e.g. Moment tensors. In 1D: k^{th} moment of a distribution $\mathbb{E}[X^k]$.

In high dimensions:

$k=1$: mean $\mathbb{E}[X]$.

$k=2$: covariance $\mathbb{E}[XX^T]$ ← need to capture information about $\mathbb{E}[x_i x_j]$

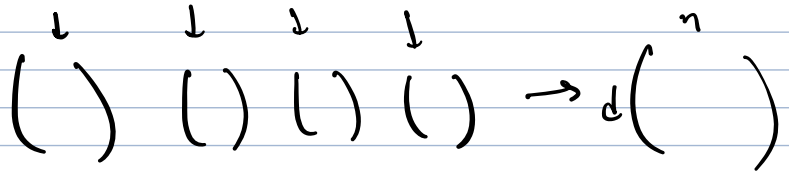
$k=3$? $\mathbb{E}[x_i x_j x_u] = ?$

e.g. batches of data. Often seen in real world ML.

single data point = sequence of n tokens

each token is represented as a d-dim embedding

"homework is so free"



so a batch of b data points is a collection of b $n \times d$ matrices.

Def: A k-tensor in dimensions $n_1 \times n_2 \times \cdots \times n_k$ T is a set of $n_1 \cdot n_2 \cdots n_k$ numbers arranged into a hypercube.

$T_{i_1 i_2 \cdots i_k}$ = element in position $(i_1, i_2, \cdots, i_k)$.

e.g. $k = 2$ → matrices.

e.g. 3-grams.

T_{ijk} = # of times $w_i w_j w_k$ appear consecutively

e.g. moment tensors

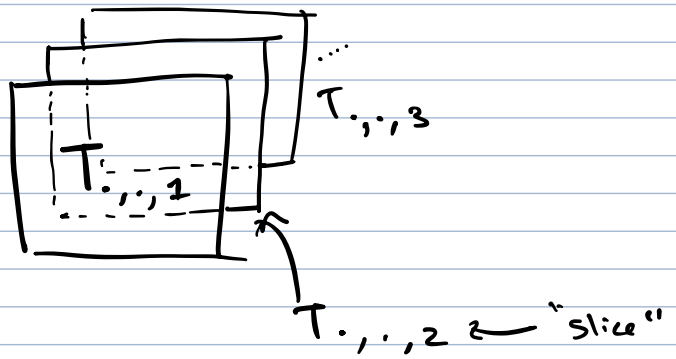
$$T_{ijk} = \mathbb{E}[x_i x_j x_k].$$

e.g. batches of data

k^{th} slice is the k^{th} data point.

denoted $T_{\cdot,\cdot,k}$ or $T(\cdot,\cdot,k)$.

$k = 3$



Alternative Definitions

Just like how it's often not good to think about matrices as blocks of n^2 numbers, this definition is often not useful for tensors.

Some equivalent notions that might be useful:

1. Tensors as linear functions on tensors.

e.g. 3-tensors are maps from $\mathbb{R}^{n_3}$ to $\mathbb{R}^{n_1 \times n_2}$

$$T(\cdot,\cdot,u) = T(\cdot,\cdot,1) \cdot u_1 + T(\cdot,\cdot,2) \cdot u_2 + \cdots + T(\cdot,\cdot,n_3)\, u_{n_3}$$

(can also switch n_1, n_2, n_3).

more generally, a k-tensor is a map from vectors to $(k-1)$-tensors.

2. Tensors as multilinear forms

$T(\cdot, \cdot, u)$ is a matrix. (2-tensor)

$T(\cdot, v, u)$ is a vector. (1-tensor)

↳ $= T(\cdot, \cdot, u) \cdot v$.

$T(w, v, u)$ is a number (0-tensor).

So $T: \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathbb{R}^{n_3} \to \mathbb{R}$ as a function.

T is <u>multilinear</u> i.e.

$$T(x+y, v, w) = T(x, v, w) + T(y, v, w)$$

and this holds in all positions.

3. Tensors as polynomials

This map has a nice form as a multilinear polynomial

e.g. $k = 1$ $\quad u \in \mathbb{R}^n$ is a tensor.

$$u(x) = \langle u, x \rangle = \sum u_i x_i$$ is a degree 1 poly.

$k = 2$. $\quad M \in \mathbb{R}^{n_1 \times n_2}$

$$M(x, y) = x^T M y = \sum_{ij} M_{ij} x_i y_j$$ is deg 2.

It is multilinear because it doesn't contain x_i^2, x_i^3, etc.

$k = 3$ $\quad T \in \mathbb{R}^{n_1 \times n_2 \times n_3}$

$$T(u, v, w) = \sum_{ijk} T_{ijk} u_i v_j w_k.$$

still multilinear.

e.g. for $k=3$ moment tensors.

$$T_{ijk} = \mathop{\mathbb{E}}_{X \sim D} [x_i x_j x_k]$$

so $$T(u, v, w) = \sum_{ijk} \mathbb{E}[x_i x_j x_k] u_i v_j w_k$$

$$= \mathbb{E}\left[\left(\sum_{ijk} x_i u_i x_j v_j x_k w_k\right)\right]$$

$$= \mathbb{E}\left[\left(\sum_i x_i u_i\right)\left(\sum_j x_j v_j\right)\left(\sum_k x_k w_k\right)\right]$$

$$= \mathbb{E}[\langle x, u\rangle \langle x, v\rangle \langle x, w\rangle].$$

Tensor rank:

Def: given vectors $u_1 \in \mathbb{R}^{n_1}, \dots, u_k \in \mathbb{R}^{n_k}$, the tensor product of $u_1, \dots, u_k$, denoted $u_1 \otimes u_2 \otimes \cdots \otimes u_k \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_k}$ is a k tensor whose $(i_1, i_2, \dots, i_k)$-th element is

$$(u_1)_{i_1} \cdot (u_2)_{i_2} \cdot \cdots \cdot (u_k)_{i_k}.$$

Such a tensor is known as a rank-1 tensor.

The rank of a tensor T is the smallest r we can write T as a sum of r rank-1 tensors.

Important note: Many of the nice properties of matrices do not transfer to tensors.

e.g.

1. There is no SVD/spectral decomposition!

   For instance, if $u \otimes v \otimes w$ is best rank-1 approx of T, it is not the case that $\text{rank}(T - u \otimes v \otimes w) = \text{rank}(T) - 1$.

   ↑ can be bigger!

2. Computing rank of k-tensor ($k \geq 3$) is NP-hard.

   "Most Tensor Problems are NP-hard" [Hillar, Lim '09].

   - eigenvectors of tensors
   - largest eigenvalue of tensors
   - rank
   - finding best rank k approximation
   - ⋮

3. We know that a random 3-tensor in $n \times n \times n$ has rank n^2.

   We don't know how to explicitly construct a family of tensors w/ rank even $\to n^{1.1}$.

However, if the rank of the tensor is sufficiently small, there is an algo to recover the decomposition.

Jennrich's Algorithm [Leurgans, Ross, Abel '93].

$k=3$, given $T = \sum_{i=1}^{r} u_i \otimes v_i \otimes w_i$, goal: find u_i, v_i, w_i.

Assume: $\{u_i\}$, $\{v_i\}$, $\{w_i\}$ are all linearly independent

Note: for $k=2$, i.e. matrices, this decomposition is non-unique.

$$M = \sum u_i v_i^T$$

take matrix B s.t. $B B^T = I$.

$$= \sum u_i B B^T v_i^T$$
$$= \sum (u_i B)(v_i B)^T$$

But for $k \geq 3$, this is unique! (sometimes).

Idea: let z be a random vector, and look at

$$T(\cdot, \cdot, z) = \sum_{i=1}^{r} u_i \otimes v_i \cdot \langle w_i, z \rangle$$

$$= \underbrace{\begin{bmatrix} | & | & \\ u_1 & u_2 & \cdots \\ | & | & \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} \langle w_1, z \rangle & & \\ & \langle w_2, z \rangle & \\ & & \ddots \langle w_r, z \rangle \end{bmatrix}}_{D_z} \begin{bmatrix} — & v_1^T & — \\ — & v_2^T & — \\ & \vdots & \\ & V^T & \end{bmatrix}$$

so if we take z, z' both random,

$$T(\cdot, \cdot, z) = U D_z V^T = M_z$$
$$T(\cdot, \cdot, z') = U D_{z'} V^T = M_{z'}$$

Now look at ← pseudoinverse (just pretend it's inverse).

$$M_z \cdot M_{z'}^{\dagger}$$

$$= U D_z V^T (U D_{z'} V^T)^{\dagger}$$

$$= U D_z V^T (V^T)^{\dagger} D_{z'}^{\dagger} U^{\dagger}$$

$$= I$$

$$= U D_z D_{z'}^{-1} U^{\dagger}$$

diagonal matrix, $D_{ii} = \frac{\langle w_i, z \rangle}{\langle w_i, z' \rangle}$. with high prob, the entries are unique → this is why we need random.

$$M_z M_{z'}^t \approx U D U^{-1}$$

(≈ : pretty much)

So eigenvectors of $M_z M_{z'}^t$ are u_i! Note: not symmetric, but still works in this case.

Since $D_z D_{z'}^{-1}$ are all distinct, eigenvectors are unique.

When do low rank tensors come up?

e.g. hidden Markov models. set of T hidden states. An agent's hidden state changes at every timestep, and we see output based on the agent's state.