First Lecture

October 20, 2017

What is Data?

- For this course, each data point is a point in $\mathbf{R}^d$, where d is large.

What is Data?

- For this course, each data point is a point in $\mathbf{R}^d$, where d is large.
- Why ? In many modern applications, there are many “features” (d features) and each data point has one component per feature which is the “importance” of that feature. [Eg. Image with d pixels: each component is the intensity of a pixel. Data here: Collection of images.]

What is Data?

- For this course, each data point is a point in $\mathbf{R}^d$, where d is large.
- Why ? In many modern applications, there are many “features” (d features) and each data point has one component per feature which is the “importance” of that feature. [Eg. Image with d pixels: each component is the intensity of a pixel. Data here: Collection of images.]
- Two broad areas:

What is Data?

- For this course, each data point is a point in $\mathbf{R}^d$, where d is large.
- Why ? In many modern applications, there are many “features” (d features) and each data point has one component per feature which is the “importance” of that feature. [Eg. Image with d pixels: each component is the intensity of a pixel. Data here: Collection of images.]
- Two broad areas:
  - **Modeling** Deciding what the features should be.

What is Data?

- For this course, each data point is a point in $\mathbf{R}^d$, where d is large.
- Why ? In many modern applications, there are many "features" (d features) and each data point has one component per feature which is the "importance" of that feature. [Eg. Image with d pixels: each component is the intensity of a pixel. Data here: Collection of images.]
- Two broad areas:
  - **Modeling** Deciding what the features should be.
  - **Understanding and Processing Data** Mathematical structure, properties of data, algorithms.

What is Data?

- For this course, each data point is a point in $\mathbf{R}^d$, where d is large.
- Why ? In many modern applications, there are many “features” (d features) and each data point has one component per feature which is the “importance” of that feature. [Eg. Image with d pixels: each component is the intensity of a pixel. Data here: Collection of images.]
- Two broad areas:
  - **Modeling** Deciding what the features should be.
  - **Understanding and Processing Data** Mathematical structure, properties of data, algorithms.
  - The course mainly deals with second area with occasional examples (as presently) of modeling.

Salton's Vector Space Model

- Suppose we have a collection of documents.

Salton's Vector Space Model

- Suppose we have a collection of documents.
- The "vocabulary" of the documents has *d* words - terms.

Salton's Vector Space Model

- Suppose we have a collection of documents.
- The "vocabulary" of the documents has d words - terms.
- Represent each document as a $d-$vector, listing the

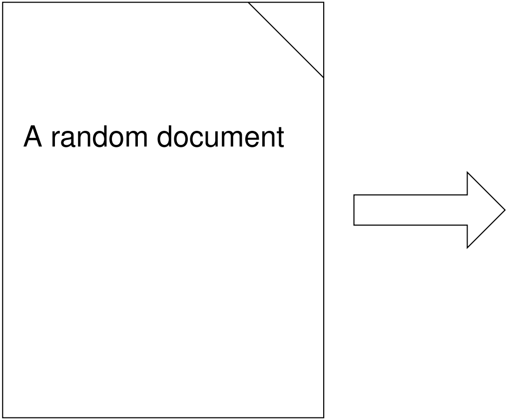
Salton's Vector Space Model

- Suppose we have a collection of **documents**.
- The "vocabulary" of the documents has d words - **terms**.
- Represent each document as a $d-$vector, listing the
  - frequency of each term in the document or

Salton's Vector Space Model

- Suppose we have a collection of documents.
- The "vocabulary" of the documents has d words - terms.
- Represent each document as a $d-$vector, listing the
  - frequency of each term in the document or
  - function of frequency.

Document turned into a vector



aardvark	0
abacus	0
⋮	
antitrust	42
⋮	
CEO	17
⋮	
Microsoft	61
⋮	
windows	14

URL's as Vectors

- Collection of d URL's.

URL's as Vectors

- Collection of d URL's.
- Each URL becomes a d vector with 0-1 coordinates.

URL's as Vectors

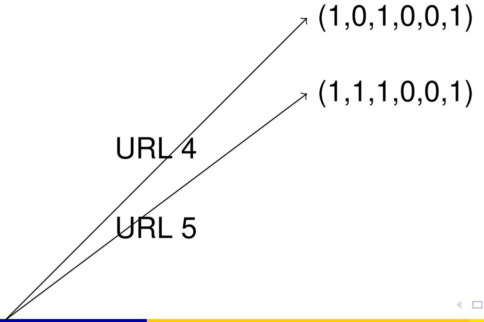
- Collection of d URL's.
- Each URL becomes a d vector with 0-1 coordinates.
  - 1 in position i if there is a hypertext link from our URL to the i th one in the collection; 0 otherwise.

URL's as Vectors

- Collection of *d* URL's.
- Each URL becomes a *d* vector with 0-1 coordinates.
  - 1 in position *i* if there is a hypertext link from our URL to the *i* th one in the collection; 0 otherwise.
- Detail : Very Sparse, so linked list instead of array representation. (Don't worry about this now.)

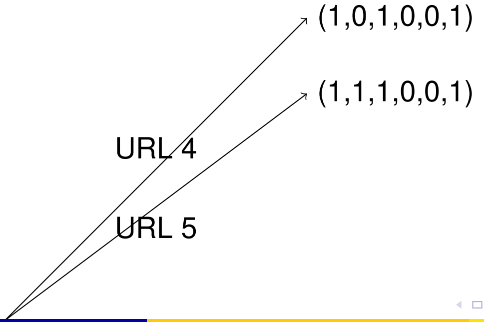
Is vector representation just a book keeping device ?

- No. Correlation between a pair of URL's maybe defined as their dot product.



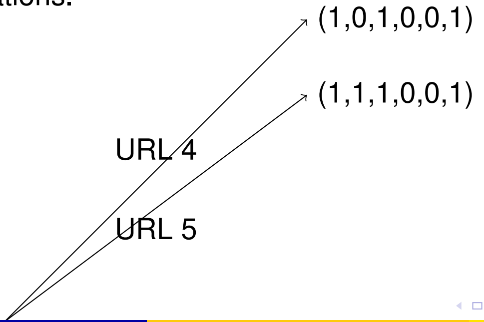
Is vector representation just a book keeping device ?

- No. Correlation between a pair of URL's maybe defined as their dot product.
  - Greater the number of common hypertext links (of two URL's), the higher their dot product or correlation. [We get a 1 in the dot product for each common hypertext link.]



Is vector representation just a book keeping device ?

- No. Correlation between a pair of URL's maybe defined as their dot product.
  - Greater the number of common hypertext links (of two URL's), the higher their dot product or correlation. [We get a 1 in the dot product for each common hypertext link.]
- Dot Products, Angles, Linear Algebra quantities all have significance in Information Retrieval, Web and many other applications.



High Dimensional Geometry

- Since data consists of points in high dim space, important to understand properties of high dim space which are quite different from our usual 2-d or 3-d intuition.

High Dimensional Geoemetry

- Since data consists of points in high dim space, important to understand properties of high dim space which are quite different from our usual 2-d or 3-d intuition.
- First **Volumes Surface Areas** and **integrals**. Volume of cube of side 1 in 3-d is 1. In fact the volume of a cube of side 1 in $\mathbf{R}^d$ (see below) is still 1.

High Dimensional Geometry

- Since data consists of points in high dim space, important to understand properties of high dim space which are quite different from our usual 2-d or 3-d intuition.
- First **Volumes Surface Areas** and **integrals**. Volume of cube of side 1 in 3-d is 1. In fact the volume of a cube of side 1 in $\mathbf{R}^d$ (see below) is still 1.
  - $\{\mathbf{x} = (x_1, x_2, \ldots, x_d) : 0 \leq x_i \leq 1\}$

High Dimensional Geometry

- Since data consists of points in high dim space, important to understand properties of high dim space which are quite different from our usual 2-d or 3-d intuition.
- First **Volumes Surface Areas** and **integrals**. Volume of cube of side 1 in 3-d is 1. In fact the volume of a cube of side 1 in $\mathbf{R}^d$ (see below) is still 1.
  - $\{\mathbf{x} = (x_1, x_2, \ldots, x_d) : 0 \leq x_i \leq 1\}$
- What is the volume of $d-$dim cube of side 2?
  - Since each of d sides has doubled, volume goes up by a factor of 2^d.

High Dimensional Geometry

- Since data consists of points in high dim space, important to understand properties of high dim space which are quite different from our usual 2-d or 3-d intuition.
- First **Volumes Surface Areas** and **integrals**. Volume of cube of side 1 in 3-d is 1. In fact the volume of a cube of side 1 in $\mathbf{R}^d$ (see below) is still 1.
    - $\{\mathbf{x} = (x_1, x_2, \ldots, x_d) : 0 \leq x_i \leq 1\}$
- What is the volume of $d-$dim cube of side 2?
    - Since each of d sides has doubled, volume goes up by a factor of 2^d.
    - Similarly, the volume of a d dimensional sphere of radius 2 is 2^d times the volume of a d dim sphere of radius 1.

High Dimensional Geometry

- Since data consists of points in high dim space, important to understand properties of high dim space which are quite different from our usual 2-d or 3-d intuition.
- First **Volumes Surface Areas** and **integrals**. Volume of cube of side 1 in 3-d is 1. In fact the volume of a cube of side 1 in $\mathbf{R}^d$ (see below) is still 1.
  - $\{\mathbf{x} = (x_1, x_2, \ldots, x_d) : 0 \leq x_i \leq 1\}$
- What is the volume of $d-$dim cube of side 2?
  - Since each of d sides has doubled, volume goes up by a factor of 2^d.
  - Similarly, the volume of a d dimensional sphere of radius 2 is 2^d times the volume of a d dim sphere of radius 1.
    - Follows by integration since each infinitesimal cube has its sides doubled.

“Surprises”

The volume of a d dim hypersphere of radius 1 goes to 0 as d goes to infinity. The course will prove such statements properly. Try this one at home. **In any case, review your multivariate Calculus immediately.**

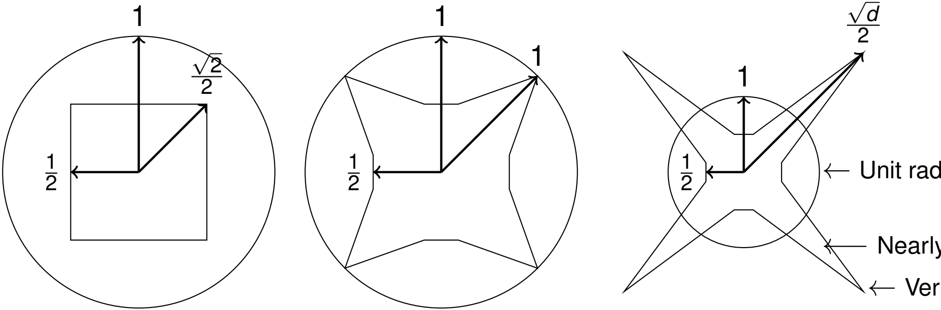
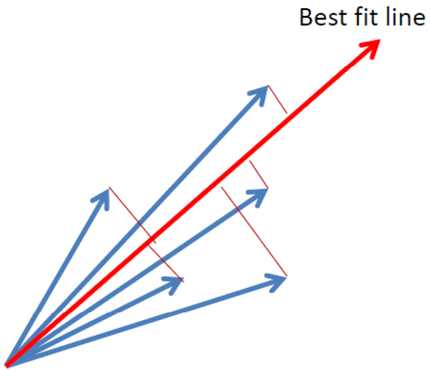


Figure: Illustration of the relationship between the sphere (radius 1) and the cube (of side 1) in 2, 4, and d dimensions.

The Best-Fit Document-direction

Best-fit direction for a set of vectors minimizes the sum of squared perpendicular distances to all documents (best-fit line).



Principal Component Analysis

- Given a set of vectors, find the best-fit direction $\mathbf{v_1}$.

Principal Component Analysis

- Given a set of vectors, find the best-fit direction $\mathbf{v_1}$.
- Find the best fit direction $\mathbf{v_2}$ perpendicular to $\mathbf{v_1}$.

Principal Component Analysis

- Given a set of vectors, find the best-fit direction $\mathbf{v_1}$.
- Find the best fit direction $\mathbf{v_2}$ perpendicular to $\mathbf{v_1}$.
- Find the best-fit direction $\mathbf{v_3}$ perpendicular to both $\mathbf{v_1}$ and $\mathbf{v_2}$.

Principal Component Analysis

- Given a set of vectors, find the best-fit direction $\mathbf{v_1}$.
- Find the best fit direction $\mathbf{v_2}$ perpendicular to $\mathbf{v_1}$.
- Find the best-fit direction $\mathbf{v_3}$ perpendicular to both $\mathbf{v_1}$ and $\mathbf{v_2}$.
- Find k such directions.

Principal Component Analysis

- Given a set of vectors, find the best-fit direction $\mathbf{v_1}$.
- Find the best fit direction $\mathbf{v_2}$ perpendicular to $\mathbf{v_1}$.
- Find the best-fit direction $\mathbf{v_3}$ perpendicular to both $\mathbf{v_1}$ and $\mathbf{v_2}$.
- Find k such directions.
- Project all data to space spanned by these k directions.

Principal Component Analysis

- Given a set of vectors, find the best-fit direction $\mathbf{v_1}$.
- Find the best fit direction $\mathbf{v_2}$ perpendicular to $\mathbf{v_1}$.
- Find the best-fit direction $\mathbf{v_3}$ perpendicular to both $\mathbf{v_1}$ and $\mathbf{v_2}$.
- Find k such directions.
- Project all data to space spanned by these k directions.
- Do something in the projection.

Principal Component Analysis

- Given a set of vectors, find the best-fit direction $\mathbf{v_1}$.
- Find the best fit direction $\mathbf{v_2}$ perpendicular to $\mathbf{v_1}$.
- Find the best-fit direction $\mathbf{v_3}$ perpendicular to both $\mathbf{v_1}$ and $\mathbf{v_2}$.
- Find k such directions.
- Project all data to space spanned by these k directions.
- Do something in the projection.
- Very widely used Algorithm. Course will see properties/algorithm for finding these directions.

Probability and high dimensional geometry

- Review / be current on your basic Probability : Random Variables, Mean, Variance, Independence, Conditional expectations, Central Limit Theorem (statement). Variance-Covarince matrix. Multi-variate Gaussian density.

Probability and high dimensional geometry

- Review / be current on your basic Probability : Random Variables, Mean, Variance, Independence, Conditional expectations, Central Limit Theorem (statement). Variance-Covarince matrix. Multi-variate Gaussian density.
- If a data quantity is the average of many (independent) quantities, it behaves like a Gaussian random variable.

Probability and high dimensional geometry

- Review / be current on your basic Probability : Random Variables, Mean, Variance, Independence, Conditional expectations, Central Limit Theorem (statement). Variance-Covarince matrix. Multi-variate Gaussian density.
- If a data quantity is the average of many (independent) quantities, it behaves like a Gaussian random variable.
- Many analogies between vectors of d independent random variables and points from d dimensional hypersphere.

Probability and high dimensional geometry

- Review / be current on your basic Probability : Random Variables, Mean, Variance, Independence, Conditional expectations, Central Limit Theorem (statement). Variance-Covarince matrix. Multi-variate Gaussian density.
- If a data quantity is the average of many (independent) quantities, it behaves like a Gaussian random variable.
- Many analogies between vectors of d independent random variables and points from d dimensional hypersphere.
- Example: If $x_1, x_2, \ldots x_d$ are independent mean 0 Gaussians, their sum is close to 0.

Probability and high dimensional geometry

- Review / be current on your basic Probability : Random Variables, Mean, Variance, Independence, Conditional expectations, Central Limit Theorem (statement). Variance-Covarince matrix. Multi-variate Gaussian density.
- If a data quantity is the average of many (independent) quantities, it behaves like a Gaussian random variable.
- Many analogies between vectors of d independent random variables and points from d dimensional hypersphere.
- Example: If $x_1, x_2, \ldots x_d$ are independent mean 0 Gaussians, their sum is close to 0.
- If $\mathbf{x}$ is a random point from the d dim hypersphere centered at the origin, then $\sum_{i=1}^{d} x_i \approx 0$. Most of the mass of the hypersphere is close to the equator.