FOUNDATIONS OF DATA SCIENCE

Juan Soto
juan.soto@tu-berlin.de
5th May 2017

Technische Universität Berlin

ANNOUNCEMENT

- Be sure to register with QISPOS, the TUB examination management system by the end of May 2017.
- PhD Positions/Opportunities
- Next week's lecture will be with Christoph Boden on the topic of **Network Analysis**

SURVEY RESULTS

	A	E	F	L	M	N	P	Q	R	S	T	U	V	W
1	Student	Java	SQL	Hadoop	Flink	Spark	LA	PROB	STAT	ML	NLA	NA	NO	DA
2		3	3	5	4	5	3	3	3	3	3	3	4	4
3		2	5	5	5	5	2	3	2	2	4	4	4	2
4		2	1	5	4	5	3	3	2	4	3	4	3	2
5		2	2	5	5	5	2	2	2	2	1	2	2	1
6		1	1	5	5	5	2	2	3	3	5	5	5	5
7		2	3	4	2	2	2	2	2	2	2	3	4	2
8		1	2	4	3	4	2	2	2	2	2	3	3	3
9		3	2	2	3	2	3	3	3	5	3	3	5	3
10		2	2	3	2	3	2	2	2	2	2	3	3	2
11		3	3	4	4	4	1	2	2	3	1	2	3	3
12		3	3	5	4	5	3	3	4	5	2	3	3	4
13		3	3	5	4	4	4	4	4	4	5	5	5	5
14		2	2	5	4	5	2	3	3	3	3	3	4	3
15		2	2	5	5	5	4	4	4	4	4	4	4	2
16		1	1	3	3	4	1	1	1	3	1	2	3	1
17		1	2	3	2	1	3	3	3	2	3	4	5	2
18		3	2	4	5	5	1	2	2	3	1	1	2	2
19		2	3	4	3	2	3	2	3	2	4	4	3	2
20		3	4	5	5	5	3	3	3	3	4	4	4	4
21		2	2	5	5	5	3	4	3	4	5	5	5	5
22		3	1	4	3	4	3	3	4	4	3	3	4	3
23		3	2	5	5	5	2	1	1	1	2	2	2	1
24		2	2	4	4	4	2	3	3	3	3	3	3	3
25		3	2	4	5	5	3	2	4	3	4	4	4	4
26		2	2	5	5	5	2	2	2	4	3	4	4	4
27		3	2	5	5	5	2	2	2	3	3	2	2	2
28		2	3	5	5	5	3	3	3	3	3	4	4	4
29		1	2	5	4	5	3	2	2	3	3	3	3	4
30		1	3	5	5	5	2	3	3	3	2	3	2	2
31														
32	AVG	2,143	2,286	4,4137931	4,06897	4,27586	2,45	2,5517	2,6552	3,03	2,9	3,28	3,52	2,9

Survey Findings

1. On average strong Java/SQL
2. A few already have systems skills
3. Majority have math/stats foundation
4. Many possess numerics background
5. A few have weak math/stat skills
6. Overall looks promising!

Proficiency Level Scale

1 expert
2 strong
3 good
4 novice
5 !experience

TODAY'S LECTURE

- Cursory review of foundations
  - definitions and concepts referenced in upcoming lectures
  - mathematical analysis, linear algebra, probability, statistics
  - detailed information will be left to the reader
  - ML foundations not fully addressed, instead handled in an ML course
- Conduct in-class exercises

MOTIVATION

1. To refresh key definitions and principles
2. To acknowledge the role that mathematical / statistical libraries play in data analytics
3. To understand that potential numerical issues may arise
4. To underscore the importance of validating the correctness of data analytics

TOPICS TO BE COVERED

- mathematical analysis
  - distance measures (clustering)
  - similarity measures (collaborative filtering)
  - Pearson Correlation Coefficient (recommender systems)
- (numerical) linear algebra
  - solving linear systems
  - matrix factorization
  - eigenvectors, eigenvalues
  - diagonalisation
  - singular value decomposition (recommender systems)
- (applied) probability/statistics
  - random variables
  - probability distributions (marginal, conditional, joint)
  - rules (chain, product, Bayes)
  - (conditional) independence
  - parameter estimation, MLE (probabilistic/statistical models)
- (numerical) optimization
  - convex functions, gradients

MATHEMATICAL ANALYSIS

DEFINITION

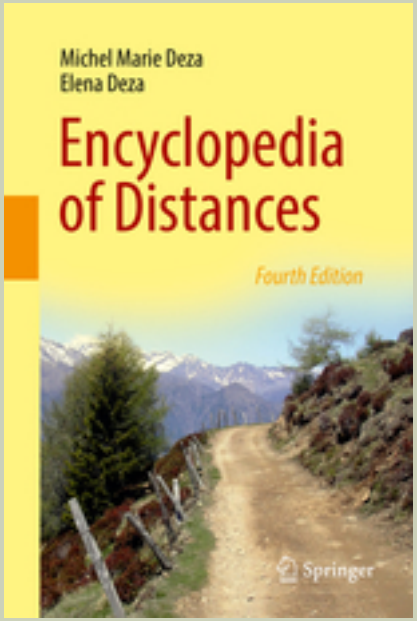
- **argmax** (the argument of the maximum)
  - set of points of the given argument for which the given function attains its maximum value
  - refers to the inputs which create maximum outputs
  - $\underset{x}{\operatorname{argmax}} f(x) = \{x \mid \forall y : f(y) \leq f(x)\}$
- Example
  - $\underset{x}{\operatorname{argmax}}(1 - |x|) = \{0\}$

DISTANCE MEASURES/METRICS

- Given a set of points in space, a distance measure on this space is a function d(x,y) that takes two points in the space and produces a real number.

- A distance measure/metric d(x,y) satisfies these four properties
  - $d(x,y) \geq 0$ (non-negative)
  - $d(x,y) = 0 \Leftrightarrow x = y$ (coincidence)
  - $d(x,y) = d(y,x)$ (symmetry)
  - $d(x,y) \leq d(x,z) + d(z,y)$ (triangle inequality)

- Examples of distance measures include
  - Euclidean distance, Minkowski distance,
  - Manhattan distance, Supremum distance,
  - Jaccard distance, Cosine distance,
  - Edit distance, Hamming distance

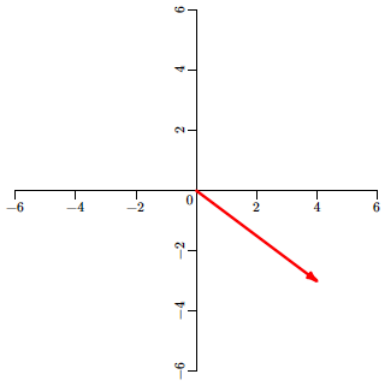


VECTOR DEFINITIONS

A **vector** is

- A 1D array of numbers
- A geometric entity with magnitude and direction
- A matrix with exactly one row or column
  - Called row vector and column vector, resp.
  - **Transpose** $\boldsymbol{v}^T$ transposes a row vector into a column vector and vice versa
- A (latent) object or attribute

	Year
Stockholm	9.95
Minsk	10.77
London	14.85
Budapest	14.91
Paris	15.46
Bucharests	16.44
Barcelona	19.90
Rome	20.44
Lisbon	21.36
Athens	22.31
Valencia	22.36
Malta	23.35



$$\begin{pmatrix} 4 \\ -3 \end{pmatrix}$$

	Jan	*Apr*	*Jul*	*Oct*	*Year*
Stockholm	−0.70	8.60	21.90	9.90	10.00

Ref.: dws.informatik.uni-mannheim.de/en/people/professors/prof-dr-rainer-gemulla/

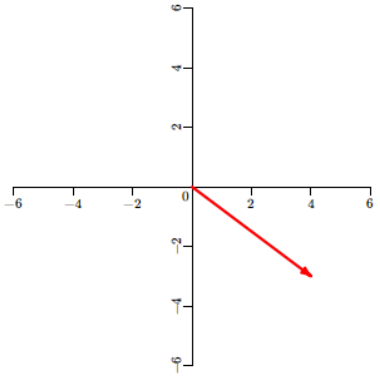
VECTOR NORM

The **norm** of vector defines its magnitude. Let $\boldsymbol{v} = \begin{pmatrix} v_1 & v_2 & \cdots & v_n \end{pmatrix}^T$.

- **Euclidean norm**: $\|\boldsymbol{v}\| = \sqrt{\sum_{i=1}^{n} v_i^2}$
  - Corresponds to intuitive notion of length in Euclidean space
- L_p **norm** for $1 \leq p \leq \infty$: $\|\boldsymbol{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p\right)^{1/p}$
  - L_1 norm = sum of absolute values (Manhattan distance from origin)
  - L_2 norm = Euclidean norm (bird-fly distance from origin)
  - L_∞ norm = maximum absolute value
  - The L_p norms decrease as p increases, i.e.,

$$\|\boldsymbol{v}\|_{p+a} \leq \|\boldsymbol{v}\|_p \qquad \text{for } a \geq 0$$

- Properties of vector norms
  - $\|\boldsymbol{v}\| > 0$ when $\boldsymbol{v} \neq 0$ and $\|\boldsymbol{v}\| = 0$ iff $\boldsymbol{v} = \boldsymbol{0}$
  - $\|a\boldsymbol{v}_1\| = |a|\,\|\boldsymbol{v}\|$ (absolute scalability)
  - $\|\boldsymbol{v}_1 + \boldsymbol{v}_2\| \leq \|\boldsymbol{v}_1\| + \|\boldsymbol{v}_2\|$ (triangle inequality)



$$v = \begin{pmatrix} 4 \\ -3 \end{pmatrix}$$

$\|\boldsymbol{v}\|_1 = 7$

$\|\boldsymbol{v}\| = 5$

$\|\boldsymbol{v}\|_\infty = 4$

Ref.: dws.informatik.uni-mannheim.de/en/people/professors/prof-dr-rainer-gemulla/

NORMS AND DISTANCES

The **distance** between two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ can be quantified with norm $\|\boldsymbol{u} - \boldsymbol{v}\|$.

	Jan	*Apr*	*Jul*	*Oct*	*Year*
• Stockholm, $\boldsymbol{s} =$	$(-0.70$	8.60	21.90	9.90	$10.00)$
• Minsk, $\boldsymbol{m} =$	$(-2.10$	12.20	23.60	10.20	$10.60)$
• Athens, $\boldsymbol{a} =$	$(12.90$	20.30	32.60	23.10	$22.30)$

L_1	$\boldsymbol{s}$	$\boldsymbol{m}$	$\boldsymbol{a}$
$\boldsymbol{s}$	0.00	7.60	61.50
$\boldsymbol{m}$	7.60	0.00	56.70
$\boldsymbol{a}$	61.50	56.70	0.00

L_2	$\boldsymbol{s}$	$\boldsymbol{m}$	$\boldsymbol{a}$
$\boldsymbol{s}$	0.00	4.27	27.60
$\boldsymbol{m}$	4.27	0.00	25.98
$\boldsymbol{a}$	27.60	25.98	0.00

L_∞	$\boldsymbol{s}$	$\boldsymbol{m}$	$\boldsymbol{a}$
$\boldsymbol{s}$	0.00	3.60	13.60
$\boldsymbol{m}$	3.60	0.00	15.00
$\boldsymbol{a}$	13.60	15.00	0.00

Ref.: dws.informatik.uni-mannheim.de/en/people/professors/prof-dr-rainer-gemulla/

DOT PRODUCT (ALGEBRAIC DEFINITION)

The **dot product** of two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ is given by

$$\boldsymbol{u} \cdot \boldsymbol{v} = \sum_{i=1}^{n} u_i v_i.$$

- Also known as **scalar product** or **inner product**
- We'll often use matrix product notation and write $\boldsymbol{u}^T \boldsymbol{v}$
- Properties (with $a, b \in \mathbb{R}$)
  - $\boldsymbol{u} \cdot \boldsymbol{v} = \boldsymbol{v} \cdot \boldsymbol{u}$
  - $(a\boldsymbol{u}) \cdot \boldsymbol{v} = a(\boldsymbol{u} \cdot \boldsymbol{v})$
  - $(a\boldsymbol{u} + b\boldsymbol{v}) \cdot \boldsymbol{w} = (a\boldsymbol{u}) \cdot \boldsymbol{w} + (b\boldsymbol{v}) \cdot \boldsymbol{w}$
- Many uses, many interpretations

Ref.: dws.informatik.uni-mannheim.de/en/people/professors/prof-dr-rainer-gemulla/

WITH DOT PRODUCTS WE CAN …

- Compute the (squared) Euclidean norm

$$\boldsymbol{v} \cdot \boldsymbol{v} = \sum_{i=1}^{n} v_i^2 = \|\boldsymbol{v}\|^2$$

- Normalize a vector to length 1 (then a **unit vector**)

$$\hat{\boldsymbol{v}} = \boldsymbol{v}/\|\boldsymbol{v}\|$$

Ref.: dws.informatik.uni-mannheim.de/en/people/professors/prof-dr-rainer-gemulla/

DOT PRODUCTS: WEIGHTED SUMS

The elements of one vector are interpreted as weights for the elements of the other vector.

Example: Anna goes shopping

Item	Bread	Butter	Pizza
Price/piece	1 €	0.50 €	3 €
Quantity bought	1	2	5

- How much does Anna pay?
- Prices can be interpreted as "weights": $\boldsymbol{p} = \begin{pmatrix} 1 & 0.5 & 3 \end{pmatrix}^T$
- Quantities are $\boldsymbol{n} = \begin{pmatrix} 1 & 2 & 5 \end{pmatrix}^T$
- Total is $\boldsymbol{p} \cdot \boldsymbol{n} = 1 \cdot 1 + 0.5 \cdot 2 + 3 \cdot 5 = 17$
- Similarly: Can interpret quantities as weights for prices

Ref.: dws.informatik.uni-mannheim.de/en/people/professors/prof-dr-rainer-gemulla/

DOT PRODUCT: EXPECTED VALUE

One vector corresponds to probabilities, the other one to a random variable.

Example: Bob is gambling

Outcome	Jackpot	Win	Loss
Probability	0.1	0.2	0.7
Amount won	5 €	1 €	-2 €

- How much does Bob win in expectation? (Should he play?)
- Probabilities $\boldsymbol{p} = \begin{pmatrix} 0.1 & 0.2 & 0.7 \end{pmatrix}^T$
  - A non-negative vector that sums to one ($\|\boldsymbol{p}\|_1 = 1$) is called a **probability vector**
  - Corresponds to a probability distribution over a finite set of outcomes
- Amounts won $\boldsymbol{x} = \begin{pmatrix} 5 & 1 & -2 \end{pmatrix}^T$
  - Corresponds to a random variable; associates a real value with each outcome
- Expected value $\boldsymbol{p} \cdot \boldsymbol{x} = 0.1 \cdot 5 + 0.2 \cdot 1 + 0.7 \cdot (-2) = -0.7$

Ref.: dws.informatik.uni-mannheim.de/en/people/professors/prof-dr-rainer-gemulla/

MATRIX NORMS

- The ℓ_1-*norm*:

$$\|A\|_1 = \max_{\|\mathbf{x}\|_1=1} \|A\mathbf{x}\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}|.$$

That is, the ℓ_1-norm of a matrix is its maximum column sum.

- The ℓ_∞-*norm*:

$$\|A\|_\infty = \max_{\|\mathbf{x}\|_\infty=1} \|A\mathbf{x}\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|.$$

That is, the ℓ_∞-norm of a matrix is its maximum *row* sum.

- The ℓ_2-*norm*:

$$\|A\|_2 = \max_{\|\mathbf{x}\|_2=1} \|A\mathbf{x}\|_2.$$

Ref.: http://www.math.usm.edu/lambers/mat610/sum10/lecture2.pdf

EXERCISE 1: VECTOR NORMS

- Let x = (1 2 3 … 10)^T, compute the following vector norms
  - L_2 norm: $\|x\|_2 = \ldots$
  - L_1 norm: $\|x\|_1 = \ldots$
  - L_∞ norm: $\|x\|_\infty = \ldots$

SOLUTION 1

- Let x = (1 2 3 … 10)^T, compute the following vector norms
  - L_2 norm: $\|x\|_2 = \sqrt{385} \approx 19.62$
  - L_1 norm: $\|x\|_1 = 55$
  - L_∞ norm: $\|x\|_\infty = 10$

EXERCISE 2: VECTOR NORMS

- Let x = (1 2 3 … n)^T, compute the following vector norms
  - L_2-norm: $\|x\|_2 = \ldots$
  - L_1-norm: $\|x\|_1 = \ldots$
  - L_∞-norm: $\|x\|_\infty = \ldots$

SOLUTION 2

- Let x = (1 2 3 … n)^T, compute the following vector norms
  - L_2-norm: $\|x\|_2 = \sqrt{\frac{1}{6}n(n+1)(2n+1)}$
  - L_1-norm: $\|x\|_1 = \frac{1}{2}n(n+1)$
  - L_∞-norm: $\|x\|_\infty = n$

COSINE SIMILARITY MEASURE

- **cosine similarity measure**
  - computes the **cosine distance** between two vectors, A and B

$$similarity = \cos\theta = \frac{A\ B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

  - θ is a measure of orientation (e.g., measures cohesion in clusters)
  - cos(0°) = 1 and cos(θ°) < 1, for θ ≠0
  - two vectors with the same orientation have a similarity = 1
  - two vectors at 90° have a similarity = 0
  - two vectors diametrically opposed have a similarity = -1

COSINE SIMILARITY: SIDE NOTES

- Similarity ranges from −1 (exactly opposite) to 1 (exactly the same)
  - with 0 (indicating orthogonality/decorrelation), and
  - in-between values (indicating intermediate similarity or dissimilarity)
- For *text matching*, the attribute vectors *A* and *B* are usually the term frequency vectors of documents
  - cosine similarity can be seen as a method of normalizing document length during comparison

COSINE SIMILARITY: SIDE NOTES

- Similarity ranges from −1 (exactly opposite) to 1 (exactly the same)
  - with 0 (indicating orthogonality/decorrelation), and
  - in-between values (indicating intermediate similarity or dissimilarity)
- For *text matching*, the attribute vectors *A* and *B* are usually the term frequency vectors of documents
  - cosine similarity can be seen as a method of normalizing document length during comparison
- In *information retrieval*, the cosine similarity of two documents will range [0,1] since term frequencies (TF-IDF weights) cannot be negative
  - the angle between two term frequency vectors cannot be greater than 90°
- If the attribute vectors are normalized by subtracting the vector means (e.g., $A - \overline{A}$), the measure is the "centered cosine similarity" and is equivalent to the Pearson correlation coefficient.

TD-IDF STATISTIC

TD-IDF

The term TD-IDF stands for (term frequency)×(inverse document frequency) and is a technique for weighting keywords from a document. Let

w_{ik} be the weight of keyword k in document i,
n_k be the number of documents containing keyword k,
tf_{ik} be the number of occurrences of keyword k in document i, and
ndoc be the total number of documents.

Then $w_{ik} = tf_{ik} \log \frac{\text{ndoc}}{n_k}$. Sometimes the term frequency tf is normalized by the number of words in the document to prevent a bias towards long documents.

TD-IDF measure reflects how important a word is to a document in a collection or corpus

EXERCISE 3: COSINE SIMILARITY

- Let x = (1 2 3 4 5 6 7 8 9 10)^T and y = (-1 -2 -3 -4 -5 -6 -7 -8 -9 -10)^T
- compute the vector cosine measure and interpret
  - $similarity = \cos(\theta) = \ldots$

SOLUTION 3

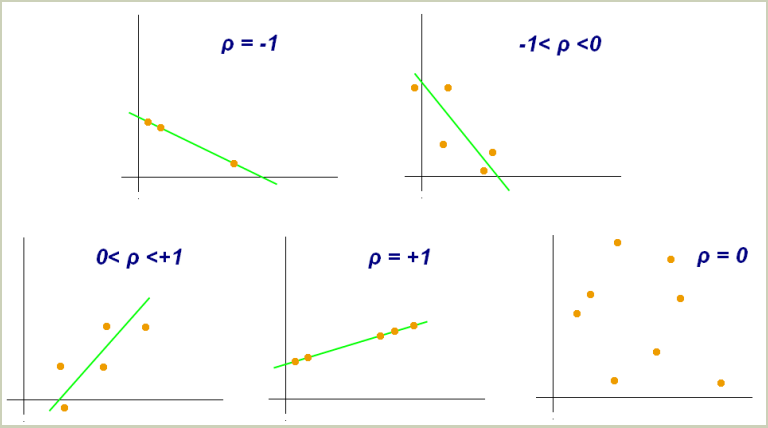
- Let x = (1 2 3 4 5 6 7 8 9 10)^T and y = (-1 -2 -3 -4 -5 -6 -7 -8 -9 -10)^T
- compute the vector cosine measure & interpret
  - $similarity = \cos(\theta) = \frac{-385}{385} = -1$
  - vectors x and y are clearly diametrically opposed

SIMILARITY MEASURE: PEARSON CORRELATION COEFFICIENT

- Pearson correlation coefficient, r is defined as follows:

$$r = \frac{\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^{n}(A_i - \bar{A})^2}\sqrt{\sum_{i=1}^{n}(B_i - \bar{B})^2}}$$

  - r measures linear correlation (dependence) between two random variables
  - r lies in the interval [-1, 1]
  - r = 0 ➔ no correlation
  - r = 1 ➔positively correlated
  - r = -1 ➔ negatively correlated



EXERCISE 4: METRICS

- Given ***A*** = *[1 2 0.5 -1]* and ***B*** = *[-2 1 -0.5 2]*, compute the
  - L_2 norm
  - Vector cosine measure
  - Pearson correlation coefficient

- For each case, interpret the meaning.

SOLUTION 4 (1 OF 3)

- Euclidean/L_2 norm = 4.4721

- **Interpretation**
  - The L_2 norm takes values in the range [0, + ∞).
  - When the L_2 norm is 0, this implies the vectors are *identical*.
  - As the measure approaches 0, the more similar the vectors.
  - As the measure increases, so to does the vector dissimilarity.
  - In this case, vectors A and B are clearly dissimilar.

SOLUTION 4 (2 OF 3)

- vcm = -0.2959

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

- Interpretation
  - vector cosine measure can attain values in [-1,1]
  - vcm = -1 ==> vectors are *exactly opposite*
  - vcm = +1 ==> vectors are *exactly identical*
  - vcm = 0 ==> vectors are *independent*
  - vcm in the range (-1,0) ==> *deg. of dissimilarity* as measure approaches -1
  - vcm in the range (0,1) ==> *deg. of similarity* as measure approaches +1
  - since the vcm = -0.2959, the vectors are slightly dissimilar.

SOLUTION 4 (3 OF 3)

- r = -0.3905

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- Interpretation
  - correlation coefficient (r) measures linear correlation (dependence)
  - r lies [-1,1]
  - r = 1 ==> *all data points lie on a line for which Y increases as X increases (i.e., total positive correlation).*
  - r = -1 ==> *all data points lie on a line for which Y decreases as X increases (i.e., total negative correlation).*
  - r = 0 ==> *no linear correlation between variables (i.e., independence)*
  - since r = -0.3905 ==> variables are slightly negatively correlated

EXERCISE 5: EDIT DISTANCE

- **Edit distance** is the minimum no. of character insertions/ deletions required to turn one string into another
  - It arises in natural language processing and bioinformatics

- Compute edit distance between each string pair: **he**, **she**, **his**, **hers**

- Identify a true statement among the following:

  (a) There are 3 pairs at distance 4.
  (b) There are 4 pairs at distance 5.
  (c) There are 4 pairs at distance 1.
  (d) There is 1 pair at distance 4.

SOLUTION 5

- (d) There is 1 pair at distance 4
- i.e., she and his

JACCARD SIMILARITY & JACCARD DISTANCE

- Jaccard similarity: **SIM(A,B)** a.k.a. Jaccard index: **J(A,B)** compares the **similarity** and diversity of sample sets
  - $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$, if A and B are both empty, J(A,B) = 1
  - Clearly, 0 ≤ J(A,B) ≤ 1
  - Interpretation
    - 0 precisely dissimilar
    - 1 perfectly similar
    - (0,1) partly (dis-)similar
- Jaccard distance measures **dissimilarity** between sample sets
  - $d_J(A,B) = 1 - J(A,B)$

EXERCISE 6: JACCARD SIMILARITY

- Suppose we have two sets A = {1,2,3,4,7} and B = {1,4,5,7,9}
- Compute the Jaccard similarity
  - $J(A,B) = \ldots$

SOLUTION 6

- Suppose we have two sets A = {1,2,3,4,7} and B = {1,4,5,7,9}
- Compute the Jaccard similarity
  - $J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{3}{7} \approx 0.4286$
  - Thus, sets A and B are partially similar

EXERCISE 7: JACCARD DISTANCE

- Here are five vectors in a 10-dimensional space
  - 1111000000, 0100100101, 0000011110, 0111111111, 1011111111
- Compute the Jaccard distance between each pair of the vectors
- Then, identify one of these distances from the list below.
- **Jaccard similarity = (no. of 1-1 matches) / (no. of bits - no. of 0-0 matches)**

(a) 7/8
(b) 6/7
(c) 7/9
(d) 1/2

SOLUTION 7

- Here are five vectors in a 10-dimensional space
  - 1111000000, 0100100101, 0000011110, 0111111111, 1011111111
- Compute the Jaccard distance (not Jaccard measure) between each pair of the vectors. Then, identify one of these distances from the list below.
- **Jaccard similarity = no. of 1-1 matches /(no. of bits - no. of 0-0 matches)**

  (a) 7/8

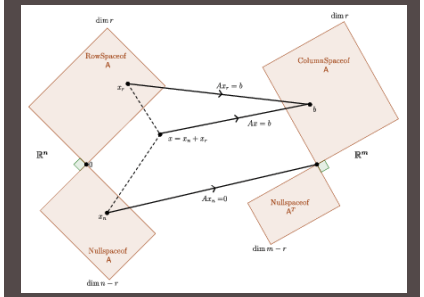
  (b) 6/7

  (c) 7/9

  (d) 1/2

```
1111000000 1-1/(10-3) = 6/7
0100100101

1111000000 1-0/(10-1) = 1
0000011110

1111000000 1-3/(10-0) = 7/10
0111111111

1111000000 1-3/(10-0)= = 7/10
1011111111

0100100101 1-1/(10-3) = 6/7
0000011110

0100100101 1-1/(10-3) = 6/7
0000011110

0100100101 1-2/(10-0) = 4/5
1011111111

0000011110 1-4/(10-1) = 5/9
0111111111

0000011110 1-4/(10-1) = 5/9
1011111111

0111111111 1-8/(10-0) = 1/5
1011111111
```

LINEAR ALGEBRA

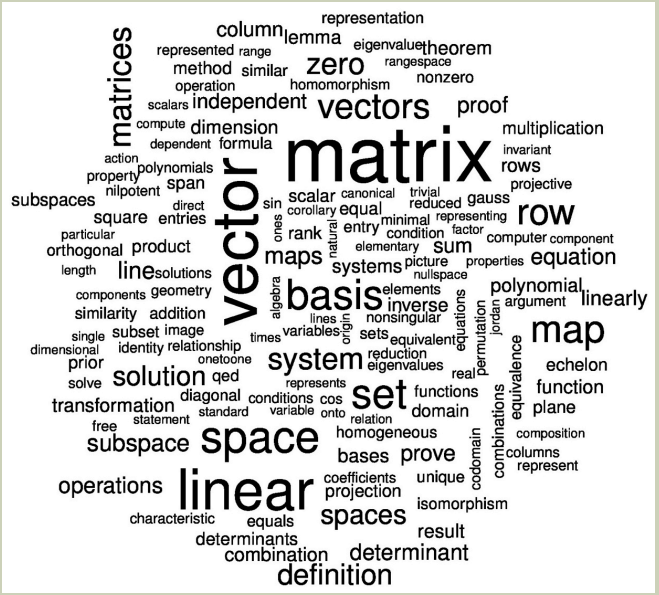


FAST FACTS

- "Central problem of linear algebra is the solution of linear equations." –**Strang**
- Linear equations of the form Ax = y

$$\begin{array}{ccccccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1j}x_j & + & \cdots & + & a_{1n}x_n & = & y_1 \\ a_{21}x_2 & + & a_{22}x_2 & + & \cdots & + & a_{2j}x_j & + & \cdots & + & a_{2n}x_n & = & y_2 \\ \cdots & & \cdots & & \cdots & & \cdots & & \cdots & & \cdots & & \cdots \\ a_{i1}x_1 & + & a_{i2}x_2 & + & \cdots & + & a_{ij}x_j & + & \cdots & + & a_{in}x_n & = & y_i \\ \cdots & & \cdots & & \cdots & & \cdots & & \cdots & & \cdots & & \cdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mj}x_j & + & \cdots & + & a_{mn}x_n & = & y_m \end{array}.$$

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_j \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_m \end{bmatrix}$$



- A few linear solvers …
  - Gauss Jordan Method
  - LU Decomposition Based Method
  - Jacobi (Iterative) Method
  - Gauss-Seidel (Iterative) Method

gams.nist.gov/cgi-bin/serve.cgi/Class/D2

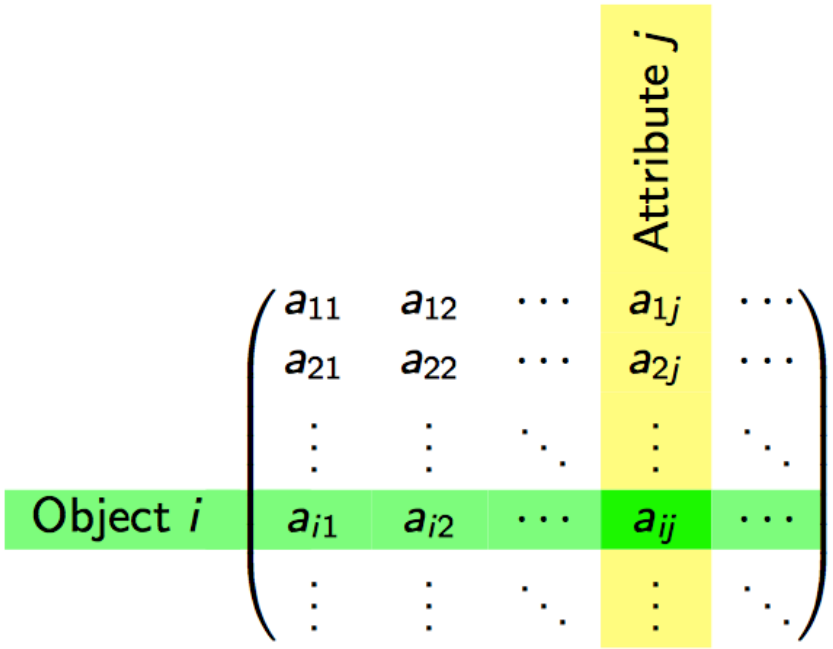
What is a matrix?

- A means to describe *computation*
  - Rotation
  - Rescaling
  - Permutation
  - Projection
  - ...

  } Linear operators

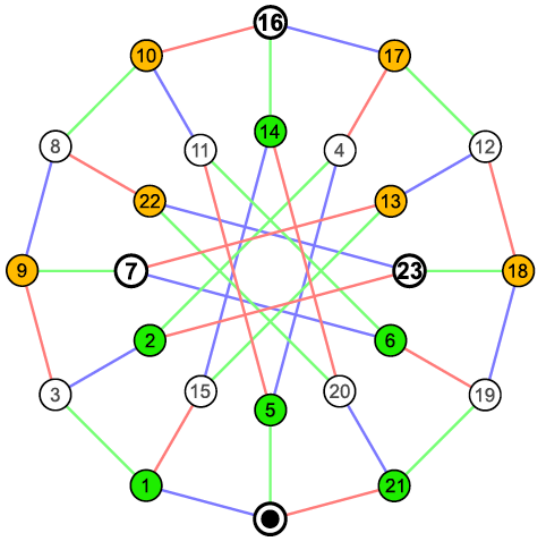
- A means to describe *data*

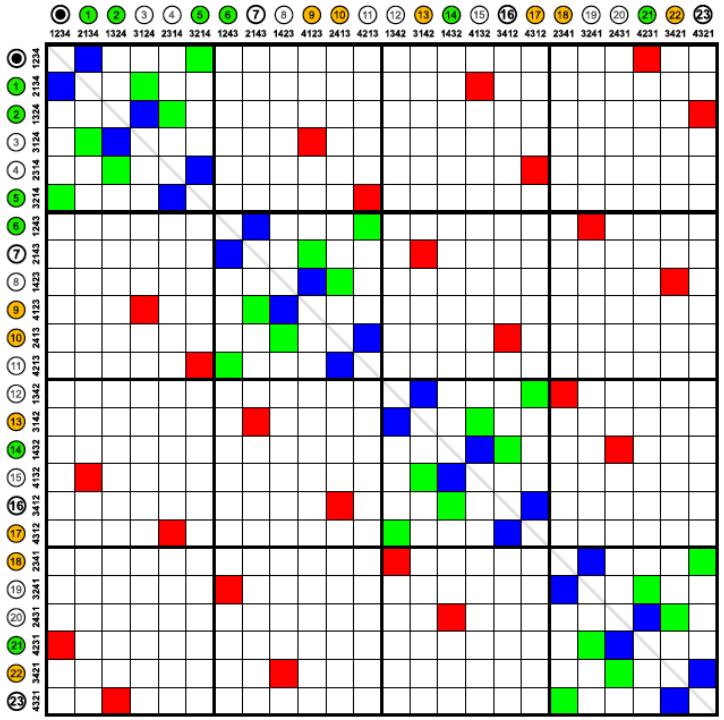
Rows	Columns	Entries
Objects	*Attributes*	*Values*
Equations	Variables	Coefficients
Data points	Axes	Coordinates
Vertices	Vertices	Edges
⋮	⋮	⋮



Ref.: dws.informatik.uni-mannheim.de/en/people/professors/prof-dr-rainer-gemulla/

A GRAPH AS AN ADJACENCY MATRIX





Adjacency matrix

Ref.: dws.informatik.uni-mannheim.de/en/people/professors/prof-dr-rainer-gemulla/

OBJECTS/ATTRIBUTES AS MATRICES

Anna, Bob, and Charlie went shopping

- Anna bought butter and bread
- Bob bought butter, bread, and beer
- Charlie bought bread and beer

	Bread	Butter	Beer
Anna	1	1	0
Bob	1	1	1
Charlie	0	1	1

Customer transactions

	Data	Matrix	Mining
Book 1	5	0	3
Book 2	0	0	7
Book 3	4	6	5

Document-term matrix

	Avatar	The Matrix	Up
Alice		4	2
Bob	3	2	
Charlie	5		3

Incomplete rating matrix

	Jan	Jun	Sep
Saarbrücken	−1	11	10
Helsinki	−6.5	10.9	8.7
Cape Town	15.7	7.8	8.7

Cities and monthly temperatures

Ref.: dws.informatik.uni-mannheim.de/en/people/professors/prof-dr-rainer-gemulla/

MATRIX RANK & LINEAR INDEPENDENCE

- A vector $\boldsymbol{u} \in \mathbb{R}^n$ is **linearly dependent** on set of vectors $V = \{\boldsymbol{v}_i\} \subset \mathbb{R}^n$ if $\boldsymbol{u}$ can be expressed as a linear combination of vectors in V
  - $\boldsymbol{u} = \sum_i a_i \boldsymbol{v}_i$ for some $a_1, \ldots, a_n \in \mathbb{R}$
  - Set V is linearly dependent if some $\boldsymbol{v}_i \in V$ is linearly dependent on $V \setminus \{\boldsymbol{v}_i\}$
  - If V is not linearly dependent, it is **linearly independent**
- The **column rank** of matrix $\boldsymbol{A}$ is the maximum number of linearly independent columns of $\boldsymbol{A}$
- The **row rank** of $\boldsymbol{A}$ is the maximum number of linearly independent rows of $\boldsymbol{A}$

Ref.: dws.informatik.uni-mannheim.de/en/people/professors/prof-dr-rainer-gemulla/

LINEAR INDEPENDENCE AND RANK

- A set of vectors $\{x_1, \ldots, x_n\}$ is linearly independent if $\nexists\{\alpha_1, \ldots, \alpha_n\}$: $\sum_{i=1}^{n} \alpha_i x_i = 0$
- Rank: $A \in \mathbb{R}^{m \times n}$, then $rank(A)$ is the maximum number of linearly independent columns (or equivalently, rows)
- Properties:
  - $rank(A) \leq \min\{m, n\}$
  - $rank(A) = rank(A^T)$
  - $rank(AB) \leq \min\{rank(A), rank(B)\}$
  - $rank(A + B) \leq rank(A) + rank(B)$

MATRIX INVERSES

The *inverse* $\mathbf{A}^{-1}$ of a matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ is defined such that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}, \tag{145}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix. If $\mathbf{A}^{-1}$ exists, $\mathbf{A}$ is said to be *nonsingular*. Otherwise, $\mathbf{A}$ is said to be *singular* (see e.g. [12]).

- This notion generalizes to non-square matrices via left- and right-inverses.
- Not all matrices have inverses.

$$A = \begin{bmatrix} 4 & 3 \\ 3 & 2 \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} -2 & 3 \\ 3 & -4 \end{bmatrix}$$

==> $$AA^{-1} = A^{-1}A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

IDENTITY, TRANSPOSE, AND SYMMETRY

- **Identity Matrix**

$$I_1 = [1],\ I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\ I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},\ \cdots,\ I_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

- **Transpose of a Matrix**

For an $m \times n$ matrix A with $(A)_{ij} = a_{ij}$, its transpose is an $n \times m$ matrix A^T with $(A^T)_{ij} = a_{ji}$.

- **Symmetric Matrix**

  A square matrix equal to its transpose, A = A^T

$$\begin{bmatrix} 1 & 7 & 3 \\ 7 & 4 & -5 \\ 3 & -5 & 6 \end{bmatrix} \stackrel{?}{=} \begin{bmatrix} 1 & 7 & 3 \\ 7 & 4 & -5 \\ 3 & -4 & 6 \end{bmatrix}$$

Are these matrices symmetric?

ORTHOGONAL MATRICES

1. An **orthogonal matrix** is a square matrix with real entries whose columns and rows are orthogonal unit vectors, i.e., $Q^TQ = QQ^T = I$.

2. A matrix **Q is orthogonal** it its transpose is equal to its inverse, i.e., $Q^T = Q^{-1}$.

3. **Examples**

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

identity transformation

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

reflection across x-axis

$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

permutation matrix

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

rotation matrix

EXERCISE 8: ORTHONORMALITY

- Let **M** be a 3x3 matrix, whose columns form an orthonormal basis

$$M = \begin{bmatrix} 2/7 & 6/7 & x \\ 3/7 & 2/7 & y \\ 6/7 & -3/7 & z \end{bmatrix}$$

- **What is meant by "*whose columns form an orthonormal basis*?"**

EXERCISE 8: ORTHONORMALITY

- For example, one constraint is that the L_2 norm for column 3 is 1.

$$M = \begin{bmatrix} 2/7 & 6/7 & x \\ 3/7 & 2/7 & y \\ 6/7 & -3/7 & z \end{bmatrix}$$

- **What are two additional constraints?**

EXERCISE 8: ORTHONORMALITY

- For **M** depicted below …

$$M = \begin{bmatrix} 2/7 & 6/7 & x \\ 3/7 & 2/7 & y \\ 6/7 & -3/7 & z \end{bmatrix}$$

- **Which of the following constraints is correct?**

  (a) z = 3y (b) 2x = 3z (c) y = 2x (d) y = 3z

- **What are the correct values for x, y, and z?**

SOLUTION 8

- For **M** depicted below …

$$M = \begin{bmatrix} 2/7 & 6/7 & x \\ 3/7 & 2/7 & y \\ 6/7 & -3/7 & z \end{bmatrix}$$

- **Which of the following constraints is correct?**

  (a) z = 3y (b) 2x = 3z (c) y = 2x (d) y = 3z

- **What are the correct values for x, y, and z?**

Answer 1:
(b) 2x = 3z

Answer 2:
x = -3/7
y = +6/7
z = -2/7

MATRIX PROPERTIES

$$
\begin{aligned}
(1) \quad (\mathbf{AB})^{-1} &= \mathbf{B}^{-1}\mathbf{A}^{-1} \\
(2) \quad (\mathbf{ABC...})^{-1} &= ...\mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1} \\
(3) \quad (\mathbf{A}^T)^{-1} &= (\mathbf{A}^{-1})^T \\
(4) \quad (\mathbf{A}+\mathbf{B})^T &= \mathbf{A}^T + \mathbf{B}^T \\
(5) \quad (\mathbf{AB})^T &= \mathbf{B}^T\mathbf{A}^T \\
(6) \quad (\mathbf{ABC...})^T &= ...\mathbf{C}^T\mathbf{B}^T\mathbf{A}^T \\
(7) \quad (\mathbf{A}^H)^{-1} &= (\mathbf{A}^{-1})^H \\
(8) \quad (\mathbf{A}+\mathbf{B})^H &= \mathbf{A}^H + \mathbf{B}^H \\
(9) \quad (\mathbf{AB})^H &= \mathbf{B}^H\mathbf{A}^H \\
(10) \quad (\mathbf{ABC...})^H &= ...\mathbf{C}^H\mathbf{B}^H\mathbf{A}^H
\end{aligned}
$$

TRANSPOSE AND TRACE PROPERTIES

- Transpose: $A \in \mathbb{R}^{m \times n}$, then $A^T \in \mathbb{R}^{n \times m}$: $(A^T)_{ij} = A_{ji}$
- Properties:
  - $(A^T)^T = A$
  - $(AB)^T = B^T A^T$
  - $(A + B)^T = A^T + B^T$
- Trace: $A \in \mathbb{R}^{n \times n}$, then: $tr(A) = \sum_{i=1}^{n} A_{ii}$
- Properties:
  - $tr(A) = tr(A^T)$
  - $tr(A + B) = tr(A) + tr(B)$
  - $tr(\lambda A) = \lambda tr(A)$
  - If AB is a square matrix, $tr(AB) = tr(BA)$

DETERMINANT AND PROPERTIES

A determinant is a function of a square matrix that reduces it to a single number.

- $A \in \mathbb{R}^{n \times n}$, $a_1, \ldots, a_n$ the rows of A, $S = \{\sum_{i=1}^{n} \alpha_i a_i | \, 0 \leq \alpha_i \leq 1\}$, then $det(A)$ is the volume of S.
- Properties:
  - $det(I) = 1$
  - $det(\lambda A) = \lambda det(A)$
  - $det(A^T) = det(A)$
  - $det(AB) = det(A)det(B)$
  - $det(A) \neq 0$ if and only if A is invertible.
  - If A invertible, then $det(A^{-}1) = det(A)$

EIGENPROBLEM

§E.2. The Standard Algebraic Eigenproblem

Consider the linear equation system (D.16), namely $\mathbf{A}\mathbf{x} = \mathbf{y}$, in which $\mathbf{A}$ is a square $n \times n$ matrix, while $\mathbf{x}$ and $\mathbf{y}$ are n-vectors. (Entries of $\mathbf{A}$, $\mathbf{x}$ and $\mathbf{y}$ may be real or complex numbers.) Suppose that the right-hand side vector $\mathbf{y}$ is required to be a multiple λ of the solution vector $\mathbf{x}$:

$$\mathbf{A}\,\mathbf{x} = \lambda\,\mathbf{x}, \tag{E.1}$$

or, written in full,

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & \lambda x_1 \\ a_{21}x_2 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & \lambda x_2 \\ \cdots & & \cdots & & \cdots & & \cdots & & \cdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \cdots & + & a_{nn}x_n & = & \lambda x_n. \end{array} \tag{E.2}$$

These equations are trivially verified for $\mathbf{x} = \mathbf{0}$. The interesting solutions, called *nontrivial*, are those for which $\mathbf{x} \neq \mathbf{0}$. Pairing a nontrivial solution with a corresponding λ that satisfies (E.1) or (E.2) gives an *eigensolution pair*, or *eigenpair* for short.

EIGENVALUES

§E.2.1. Characteristic Equation

The eigenystem (E.1) can be rearranged into the homogeneous form

$$(\mathbf{A} - \lambda \mathbf{I})\,\mathbf{x} = \mathbf{0}. \tag{E.3}$$

A nontrivial solution $\mathbf{x} \neq \mathbf{0}$ of (E.3) is possible if and only if the coefficient matrix $\mathbf{A} - \lambda\mathbf{I}$ is singular. Such a condition can be expressed as the vanishing of the determinant

$$|\mathbf{A} - \lambda \mathbf{I}| = \begin{vmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{vmatrix} = 0. \tag{E.4}$$

When this determinant is expanded, we obtain an algebraic polynomial equation in λ of degree n:

$$P(\lambda) = \lambda^n + \alpha_1 \lambda^{n-1} + \cdots + \alpha_n = 0. \tag{E.5}$$

For future use, the diagonal matrix of eigenvalues will be denoted by $\mathbf{\Lambda} = \mathbf{diag}[\lambda_i]$. This will be called the *eigenvalue matrix*.

- Properties:
  - $tr(A) = \sum_{i=1}^{n} \lambda_i$
  - $det(A) = \prod_{i=1}^{n} \lambda_i$
  - $rank(A) = |\{1 \leq i \leq n | \lambda_i \neq 0\}|$

Typically, numerical methods, such as the QR algorithm or power methods are used to calculate eigenvalues.

EIGENVECTORS

§E.2.2. Eigenvectors

Associated to each eigenvalue λ_i there is a nonzero vector $\mathbf{x}_i$ that satisfies (E.1) instantiated for $\lambda \to \lambda_i$:

$$\mathbf{A}\,\mathbf{x}_i = \lambda_i\,\mathbf{x}_i\,. \qquad \mathbf{x}_i \neq 0. \tag{E.6}$$

This $\mathbf{x}_i$ is called a *right eigenvector* or *right characteristic vector*. This is often abbreviated to just *eigenvector*, in which case the "right" qualifier is tacitly understood. If $\mathbf{A}$ is real and so is λ_i, $\mathbf{x}_i$ has real entries. But it λ_i is complex, entries of $\mathbf{x}_i$ will generally be complex.

The *left eigenvectors* of $\mathbf{A}$ are the right eigenvectors of its transpose, and are denoted by $\mathbf{y}_i$:

$$\mathbf{A}^T\,\mathbf{y}_i = \lambda_i\,\mathbf{y}_i, \qquad \mathbf{y}_i \neq 0. \tag{E.7}$$

EXERCISE 9: EIGENPAIRS

- For the matrix A below, compute the eigenpairs
  - $A = \begin{pmatrix} 5 & 8 & 16 \\ 4 & 1 & 8 \\ -4 & -4 & -11 \end{pmatrix}$

SOLUTION 9 (1 OF 2)

- For the matrix A below, compute the eigenpairs
  - $A = \begin{pmatrix} 5 & 8 & 16 \\ 4 & 1 & 8 \\ -4 & -4 & -11 \end{pmatrix}$
  - $p(\lambda) = \det \begin{bmatrix} 5-\lambda & 8 & 16 \\ 4 & 1-\lambda & 8 \\ -4 & -4 & -11-\lambda \end{bmatrix} = (\lambda - 1)(\lambda + 3)^2$
  - Thus, the eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = -3$ and the corresponding eigenvectors are $x_1 = (-2\ -1\ 1)^T$, $x_2 = (-1\ 1\ 0)^T$, and $x_3 = (-2\ 0\ 1)^T$

SOLUTION 9 (2 OF 2)

Let $A = \begin{bmatrix} 5 & 8 & 16 \\ 4 & 1 & 8 \\ -4 & -4 & -11 \end{bmatrix}$. Then $p(\lambda) = \det \begin{bmatrix} 5-\lambda & 8 & 16 \\ 4 & 1-\lambda & 8 \\ -4 & -4 & -11-\lambda \end{bmatrix} = (\lambda - 1)(\lambda + 3)^2$

after some algebra! Thus, $\lambda_1 = 1$ and $\lambda_2 = -3$ are the eigenvalues of A. Eigenvectors $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$ corresponding to $\lambda_1 = 1$ must satisfy

$$\begin{aligned} 4v_1 + 8v_2 + 16v_3 &= 0 \\ 4v_1 + 8v_3 &= 0 \\ -4v_1 - 4v_2 - 12v_3 &= 0. \end{aligned}$$

Letting $v_3 = t$, we find from the second equation that $v_1 = -2t$, and then $v_2 = -t$. All eigenvectors corresponding to $\lambda_1 = 1$ are multiples of $\begin{bmatrix} -2 \\ -1 \\ 1 \end{bmatrix}$, and so the eigenspace corresponding to $\lambda_1 = 1$ is given by the span of $\begin{bmatrix} -2 \\ -1 \\ 1 \end{bmatrix}$. $\left\{ \begin{bmatrix} -2 \\ -1 \\ 1 \end{bmatrix} \right\}$ is a basis for the eigenspace corresponding to $\lambda_1 = 1$.

Eigenvectors corresponding to $\lambda_2 = -3$ must satisfy

$$\begin{aligned} 8v_1 + 8v_2 + 16v_3 &= 0 \\ 4v_1 + 4v_2 + 8v_3 &= 0 \\ -4v_1 - 4v_2 - 8v_3 &= 0. \end{aligned}$$

The equations here are just multiples of each other! If we let $v_3 = t$ and $v_2 = s$, then $v_1 = -s - 2t$. Eigenvectors corresponding to $\lambda_2 = -3$ have the form

$$\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} s + \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix} t.$$

Thus, the eigenspace corresponding to $\lambda_2 = -3$ is two-dimensional and is spanned by $\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}$. $\left\{ \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix} \right\}$ is a basis for the eigenspace corresponding to $\lambda_2 = -3$.

EIGENVALUE DECOMPOSITION & DIAGONALISATION

- Eigenvalue decomposition of symmetric $M \in \mathbb{R}^{m \times m}$ is

$$M = Q\Sigma Q^T = \sum_{i=1}^{m} \lambda_i \boldsymbol{q}_i \boldsymbol{q}_i^T$$

  - $\Sigma = \text{diag}(\lambda_1, \dots, \lambda_n)$ contains eigenvalues of M
  - Q is orthogonal and contains eigenvectors $\boldsymbol{q}_i$ of M
- If M is not symmetric but *diagonalizable*

$$M = Q\Sigma Q^{-1}$$

  - Σ is diagonal by possibly complex
  - Q not necessarily orthogonal
- When M is symmetric
  - Eigenvalue decomposition is singular value decomposition

SIMILARITY & DIAGONALIZATION: FACTS

1. Matrices A and B are **similar** if there is an invertible matrix P such that $A = P^{-1}BP$

2. If *A and B are similar*, then they **have the same eigenvalues**

3. Even though two similar matrices, A and B, have the same eigenvalues, their eigenvectors are in general different

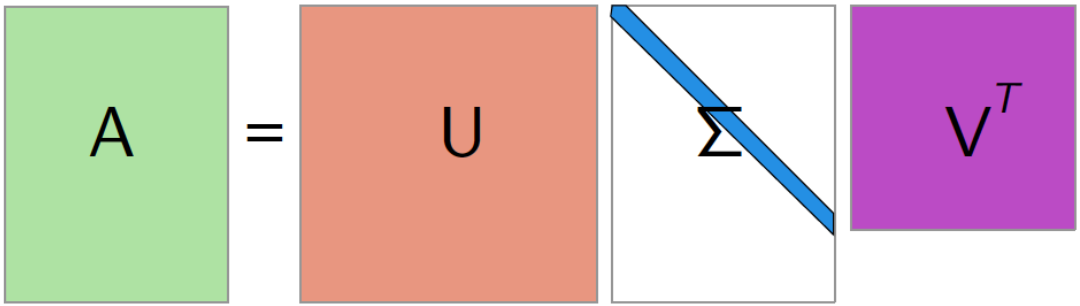
4. Matrix A is **diagonalizable** if A is similar to a diagonal matrix

SINGULAR VALUE DECOMPOSITION

Theorem

For each $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, *there are orthogonal matrices* $\boldsymbol{U}_{m \times m}$, $\boldsymbol{V}_{n \times n}$, *and a diagonal matrix* $\boldsymbol{\Sigma}_{m \times n}$ *with values* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)} \geq 0$ *on the main diagonal such that* $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$.

- $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ is called the **singular value decomposition** (SVD) of $\boldsymbol{A}$
- Values σ_i are the **singular values** of $\boldsymbol{A}$
- Columns of $\boldsymbol{U}$ are the **left singular vectors** of $\boldsymbol{A}$
- Columns of $\boldsymbol{V}$ are the **right singular vectors** of $\boldsymbol{A}$



Ref.: dws.informatik.uni-mannheim.de/en/people/professors/prof-dr-rainer-gemulla/

SINGULAR VALUE DECOMPOSITION

Singular Value Decomposition

Any matrix $A \in \mathbb{R}^{m \times n}$ can be decomposed as follows:

$$A = U \Sigma V^T$$

where $UU^T = VV^T = I$ and Σ is diagonal.

The SVD enables us to compute a least squares solution to a linear system of equations (e.g., using the pseudoinverse).

http://databricks.com/blog/2014/07/21/distributing-the-singular-value-decomposition-with-spark.html

SVD METHOD

- Three mutually compatible viewpoints
- The SVD method …
  1. transforms correlated variables into a set of uncorrelated ones that better expose the various relationships among the original data items

  1. identifies and orders the dimensions along which data points exhibit the most variation

  2. having identified where the most variation is, helps us find the best approximation of the original data points using fewer dimensions

- SVD is a data reduction method

SVD KEY IDEA

- Take a *high dimensional*, *highly variable* set of data points & reduce it to a lower dimensional space that exposes the substructure of the original data more clearly and orders it from most variation to the least

Reduced singular value decomposition is the mathematical technique underlying a type of document retrieval and word similarity method variously called *Latent Semantic Indexing* or *Latent Semantic Analysis.* The insight underlying the use of SVD for these tasks is that it takes the original data, usually consisting of some variant of a word×document matrix, and breaks it down into linearly independent components. These components are in some sense an abstraction away from the noisy correlations found in the original data to sets of values that best approximate the underlying structure of the dataset along each dimension independently. Because the majority of those components are very small, they can be ignored, resulting in an approximation of the data that contains substantially fewer dimensions than the original. SVD has the added benefit that in the process of dimensionality reduction, the representation of items that share substructure become more similar to each other, and items that were dissimilar to begin with may become more dissimilar as well. In practical terms, this means that documents about a particular topic become more similar even if the exact same words don't appear in all of them.

COMPUTING THE SVD

- SVD takes an *nxp* matrix A and decomposes it as $A = U\Lambda V^T$, where
  - A (*nxp*), U (*nxp*), V (*pxp*)
  - $U^TU = I$ (*nxn*), $V^TV = I$ (*pxp*), i.e., U and V are orthogonal

- Calculating the SVD consists of finding the eigenvalues and eigenvectors of AA^T and A^TA

- Note
  - $A = U\Lambda V^T$ and $A^T = V\Lambda U^T$
  - $A^TA = V\Lambda U^TU\Lambda V^T$
  - $A^TA = V\Lambda^2V^T$
  - $A^TAV = V\Lambda^2$

Ref.: web.mit.edu/be.400/www/SVD/Singular_Value_Decomposition.htm

EXERCISE 10: SVD

- Which of the following is the correct SVD for A?

$$A = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix}$$

1. $$A = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 5 & 0 & 0 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{18} & -1/\sqrt{18} & 4/\sqrt{18} \\ 2/3 & -2/3 & -1/3 \end{pmatrix}$$

2. $$A = \begin{pmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{18} & -1/\sqrt{18} & 4/\sqrt{18} \\ 2/3 & -2/3 & -1/3 \end{pmatrix}$$

3. $$A = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{18} & -1/\sqrt{18} & 4/\sqrt{18} \\ 2/3 & -2/3 & -1/3 \end{pmatrix}$$

SOLUTION 9 (1 OF 2)

First we compute the singular values σ_i by finding the eigenvalues of AA^T.

$$AA^T = \begin{pmatrix} 17 & 8 \\ 8 & 17 \end{pmatrix}.$$

The characteristic polynomial is $det(AA^T - \lambda I) = \lambda^2 - 34\lambda + 225 = (\lambda - 25)(\lambda - 9)$, so the singular values are $\sigma_1 = \sqrt{25} = 5$ and $\sigma_2 = \sqrt{9} = 3$.

Now we find the right singular vectors (the columns of V) by finding an orthonormal set of eigenvectors of A^TA. It is also possible to proceed by finding the left singular vectors (columns of U) instead. The eigenvalues of A^TA are 25, 9, and 0, and since A^TA is symmetric we know that the eigenvectors will be orthogonal.

For $\lambda = 25$, we have

$$A^TA - 25I = \begin{pmatrix} -12 & 12 & 2 \\ 12 & -12 & -2 \\ 2 & -2 & -17 \end{pmatrix}$$

which row-reduces to $\begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$. A unit-length vector in the kernel of that matrix is $v_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{pmatrix}$.

For $\lambda = 9$ we have $A^TA - 9I = \begin{pmatrix} 4 & 12 & 2 \\ 12 & 4 & -2 \\ 2 & -2 & -1 \end{pmatrix}$ which row-reduces to $\begin{pmatrix} 1 & 0 & -\frac{1}{4} \\ 0 & 1 & \frac{1}{4} \\ 0 & 0 & 0 \end{pmatrix}$.

A unit-length vector in the kernel is $v_2 = \begin{pmatrix} 1/\sqrt{18} \\ -1/\sqrt{18} \\ 4/\sqrt{18} \end{pmatrix}$.

For the last eigenvector, we could compute the kernel of A^TA or find a unit vector perpendicular to v_1 and v_2. To be perpendicular to $v_1 = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$ we need $-a = b$. Then the condition that $v_2^T v_3 = 0$ becomes $2a/\sqrt{18} + 4c/\sqrt{18} = 0$ or $-a = 2c$. So $v_3 = \begin{pmatrix} a \\ -a \\ -a/2 \end{pmatrix}$ and for it to be unit-length we need $a = 2/3$ so $v_3 = \begin{pmatrix} 2/3 \\ -2/3 \\ -1/3 \end{pmatrix}$.

SOLUTION 9 (2 OF 2)

So at this point we know that

$$A = U\Sigma V^T = U \begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{18} & -1/\sqrt{18} & 4/\sqrt{18} \\ 2/3 & -2/3 & -1/3 \end{pmatrix}.$$

Finally, we can compute U by the formula $\sigma u_i = Av_i$, or $u_i = \frac{1}{\sigma} Av_i$. This gives $U = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$. So in its full glory the SVD is:

$$A = U\Sigma V^T = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{18} & -1/\sqrt{18} & 4/\sqrt{18} \\ 2/3 & -2/3 & -1/3 \end{pmatrix}.$$

www.d.umn.edu/~mhampton/m4326svd_example.pdf

LINEAR ALGEBRA IN A NUTSHELL

((*The matrix* A *is* n *by* n))

Nonsingular	Singular
A is invertible	A is not invertible
The columns are independent	The columns are dependent
The rows are independent	The rows are dependent
The determinant is not zero	The determinant is zero
$A\boldsymbol{x}=\mathbf{0}$ has one solution $\boldsymbol{x}=\mathbf{0}$	$A\boldsymbol{x}=\mathbf{0}$ has infinitely many solutions
$A\boldsymbol{x}=\boldsymbol{b}$ has one solution $\boldsymbol{x}=A^{-1}\boldsymbol{b}$	$A\boldsymbol{x}=\boldsymbol{b}$ has no solution or infinitely many
A has n (nonzero) pivots	A has $r<n$ pivots
A has full rank $r=n$	A has rank $r<n$
The reduced row echelon form is $R=I$	R has at least one zero row
The column space is all of $\mathbf{R}^n$	The column space has dimension $r<n$
The row space is all of $\mathbf{R}^n$	The row space has dimension $r<n$
All eigenvalues are nonzero	Zero is an eigenvalue of A
$A^{\mathrm{T}}A$ is symmetric positive definite	$A^{\mathrm{T}}A$ is only semidefinite
A has n (positive) singular values	A has $r<n$ singular values

Due to Prof. Gilbert Strang, MIT

MATRIX FACTORIZATIONS

$\boldsymbol{A = QR}$ = (orthonormal columns in Q) (upper triangular R).

Requirements: A has independent columns. Those are *orthogonalized* in Q by the Gram-Schmidt or Householder process. If A is square then $Q^{-1} = Q^{\mathrm{T}}$.

$\boldsymbol{A = X\Lambda X^{-1}}$ = (eigenvectors in X) (eigenvalues in Λ) (left eigenvectors in X^{-1}).

Requirements: A must have n linearly independent eigenvectors.

$\boldsymbol{S = Q\Lambda Q^{\mathrm{T}}}$ = (orthogonal matrix Q) (real eigenvalue matrix Λ) (Q^{T} is Q^{-1}).

Requirements: S is *real and symmetric*: $S^{\mathrm{T}} = S$. This is the Spectral Theorem.

$$\boldsymbol{A = U\Sigma V^{\mathrm{T}}} = \begin{pmatrix} \text{orthogonal} \\ U \text{ is } m \times m \end{pmatrix} \begin{pmatrix} m \times n \text{ singular value matrix} \\ \sigma_1, \ldots, \sigma_r \text{ on its diagonal} \end{pmatrix} \begin{pmatrix} \text{orthogonal} \\ V \text{ is } n \times n \end{pmatrix}.$$

Requirements: None. This ***Singular Value Decomposition*** (SVD) has the eigenvectors of AA^{T} in U and eigenvectors of $A^{\mathrm{T}}A$ in V; $\sigma_i = \sqrt{\lambda_i(A^{\mathrm{T}}A)} = \sqrt{\lambda_i(AA^{\mathrm{T}})}$.

Due to Prof. Gilbert Strang, MIT

PROBABILITY/STATISTICS

PROBABILITY PARADIGMS

- Frequentist Interpretation
  - probabilities represent long run frequencies of events
  - proportion of the time an event occurs in a long sequence of trials
  - (e.g., if we flip an unbiased coin many times, we expect it to land on heads about half the time)

- Bayesian Interpretation
  - probability is used to quantify our uncertainty about something
  - it is fundamentally related to information rather than repeated trials
  - based on degrees of belief; subjective assessment concerning whether the event will occur (or has occurred)
  - we believe the coin is equally likely to land on heads or tails on the next toss
  - an advantage is that it can be used to model our uncertainty about events that do not have long term frequencies

SETS AND RANDOM EXPERIMENTS

- A **set** is a collection of objects
  - A = set of all possible outcomes upon tossing a fair die
  - A = {1, 2, 3, 4, 5, 6} or A = {i : i = 1, 2, …, 6}

- A **random experiment** consists of a procedure (e.g., flip two coins) and an observation (e.g., number of heads) and has the following characteristics:
  - can be repeated indefinitely under unchanged conditions
  - cannot state particular outcome ahead of time, but can describe the set of all possible outcomes
  - when experiment is repeated, individual outcomes are "random," but after a large number of repetitions, a pattern appears (*statistical regularity*)
  - probability theory describes statistical regularity

SAMPLE SPACE & EVENTS

- **Sample space** (S)
  - set of all possible outcomes of a random experiment
- **Event** (E)
  - set of possible outcomes (a subset of a S)
- **Sample point**
  - an individual element of S, one possible outcome
- Examples
  - E_1 = toss fair coin, observe H or T
  - S_1 = {H, T} outcomes are equally likely P(H) = P(T) = ½
  - E_2 = printer operational status, observe yes or no
  - S_2 = {yes, no} outcomes are not equally likely P(yes) = p, P(no) = 1-p
  - E_3 = test items from an assembly line for defects over 24 hour period
  - S_3 = {0, 1, 2, …} expect some regularity, described in terms of a PMF

PROBABILITY AXIOMS

- Probability axioms assume **probability** is an underlying property of an event, which satisfies certain mathematical rules
- Let **S** be a sample space of a random experiment
- To each event A we associate a real number **P(A)**
- If **P** satisfies the following axioms, then it is the probability of A
  - P1. 0 <= P(A) <= 1
  - P2. P(S) = 1
  - P3. $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$

CONDITIONAL DISTRIBUTION & INDEPENDENCE

- For any two events E and F
  - $P(E \mid F) = \frac{P(EF)}{P(F)}$, provided P(F) > 0 **CONDITIONAL PROBABILITY**
  - $P(E,F) = P(E \mid F)P(F)$ **PRODUCT RULE**
- Any two events E and F are **independent** iff
  - P(EF) = P(E)P(F)
  - P(E|F) = P(E)
  - P(F|E) = P(F)

FUNDAMENTAL RULES

- Given a joint distribution on two events p(E,F) …
- The marginal distribution
  - $p(E) = \sum_f p(E,F) = \sum_f p(E \mid F = f)p(F = f)$ SUM RULE (RULE OF TOTAL PROBABILITY)
  - $p(F) = \sum_e p(F,E) = \sum_e p(F \mid E = e)p(E = e)$ SUM RULE (RULE OF TOTAL PROBABILITY)
- The chain rule
  - $p(X_1,X_2,\ldots,X_n) = p(X_1)\, p(X_2 \mid X_1)\, p(X_3 \mid X_1,X_2)\, p(X_4 \mid X_1,X_2,X_3) \ldots p(X_n \mid X_1, \ldots, X_{n-1})$
- Bayes Theorem (Bayes Rule)
  - $p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{p(X = x)p(Y = y \mid X = x)}{\sum_x p(X = x)p(Y = y \mid X = x)}$

RANDOM VARIABLES

- Let E be an experiment with sample space S
- A function X that maps the points of S to the real line is a random variable

  - Example 1: X = no. of tests until find one operational
    - X(y) = 1
    - X(ny) = 2
    - X(nny) = 3
    - Then, $R_X = \{1, 2, 3, \ldots\}$

  - Example 2: X = no. of heads observed from three coin tosses
    - S = {HHH,HHT,HTH,HTT,THH,THT,TTH,TTT}
    - X(HHH) = 3
    - X(HHT) = X(HTH) = X(THH) = 2
    - X(HTT) = X(THT) = X(TTH) = 1
    - X(TTT) = 0
    - Then, $R_X = \{0, 1, 2, 3\}$

DISCRETE RANDOM VARIABLES

- For discrete RVs, the probability distribution is described by the probability mass function (PMF)
  - p(x) = P(X = x) for each x in R_X, where x is a no. and X is the random variable
  - Example (Binomial Distribution)
  - if x = 1, then P(X = 1) = p(1) = P(y) = 0.95
  - if x = 2, then P(X = 2) = p(2) = P(ny) = (0.05)(0.95)
  - if x = 3, then P(X = 3) = p(3) = P(nny) = $(0.05)^2(0.95)$
  - if x = k, then P(X = k) = p(k) = P(nnn…ny) = $(0.05)^{n-1}(0.95)$
- Expected Value (center mass of a distribution)
  - $E[X] = \sum_x xp(x) = \mu$
- Variance $VAR(X) = E[X]^2 - \mu^2$
- Standard Deviation $STD(X) = +(VAR(X))^{1/2}$

JOINT PMF

- Joint PMF of discrete random variables X and Y is
  - p(x,y) = P(X = x, Y = y)
  - satisfy the probability axioms (e.g., P1 and P2)
- Example
  - X = no. of minor defects in a new car
  - Y = no. of major defects in a new car

p(x,y)	x = 0	x = 1	x = 2	x = 3
y = 0	0.1	0.2	0.2	0.1
y = 1	0.05	0.05	0.1	0.1
y = 2	0	0.01	0.04	0.05

  - P(exactly 2 minor defects, 1 major defect) = ?
  - P(≤ 1 minor defect) = ?

JOINT PMF

- Joint PMF of discrete random variables X and Y is
  - p(x,y) = P(X = x, Y = y)
  - satisfy the probability axioms (e.g., P1 and P2)
- Example
  - X = no. of minor defects in a new car
  - Y = no. of major defects in a new car

p(x,y)	x = 0	x = 1	x = 2	x = 3
y = 0	0.1	0.2	0.2	0.1
y = 1	0.05	0.05	0.1	0.1
y = 2	0	0.01	0.04	0.05

  - P(exactly 2 minor defects, 1 major defect) = p(2,1) = 0.1
  - P(≤ 1 minor defect) = p(0,0)+p(0,1)+p(0,2)+p(1,0)+p(1,1)+p(1,2) = 0.41

MARGINAL PMF

- Marginal PMF is
  - $p_X(x) = P(X = x) = \sum_y p(x,y)$
  - $p_Y(y) = P(Y = y) = \sum_x p(x,y)$

p(x,y)	x = 0	x = 1	x = 2	x = 3
y = 0	0.1	0.2	0.2	0.1
y = 1	0.05	0.05	0.1	0.1
y = 2	0	0.01	0.04	0.05

- Example
  - $R_Y = \{0, 1, 2\}$
  - $p_Y(0) = P(Y=0) = p(0,0)+p(1,0)+p(2,0)+p(3,0) = 0.1+0.2+0.2+0.1 = 0.6$
  - $p_Y(1) = P(Y=1) = \sum_x p(x,1) = 0.3$
  - $p_Y(2) = P(Y=2) = \sum_x p(x,2) = p(0,2)+p(1,2)+p(2,2)+p(3,2) = 0.1$

INDEPENDENCE

- Random variables X and Y are independent iff
- $p(x,y) = p_X(x)\, p_Y(y)$ for all x and y
- $P(X = x, Y = y) = P(X = x)\, P(Y = y)$

- Example
  - $p(0,0) = p_X(0)\, p_Y(0)$
  - Is this true?
  - Are X & Y independent?

p(x,y)	x = 0	x = 1	x = 2	x = 3
y = 0	0.1	0.2	0.2	0.1
y = 1	0.05	0.05	0.1	0.1
y = 2	0	0.01	0.04	0.05

COVARIANCE AND CORRELATION

- **Covariance**
  - $COV(XY) = E[XY] - \mu_X \mu_Y$
  - measures how random variables are related
  - positive covariance means the r.v.'s are positively related
  - negative covariance means the r.v.'s are inversely related

- **Correlation**
  - $CORR(XY) = COV(X,Y)/(STD(X)*STD(Y))$
  - $-1 <= CORR(XY) <= 1$
  - measures the strength of a linear relationship between X and Y
  - $CORR(XY) = 0$ => X and Y are uncorrelated
  - $CORR(XY) > 0$, if Y (wt.) generally increases as X (ht.) increases
  - $CORR(XY) < 0$, if Y (heating cost) generally decrease as X (temp.) increases

- Independent random variables are uncorrelated
- Uncorrelated random variables are not necessarily independent

COVARIANCE OF RANDOM VARIABLES

Definition. Let X and Y be random variables (discrete or continuous!) with means μ_X and μ_Y. The **covariance** of X and Y, denoted $\boldsymbol{Cov(X,Y)}$ or $\boldsymbol{\sigma_{XY}}$, is defined as:

$$Cov(X, Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$

That is, if X and Y are discrete random variables with joint support S, then the covariance of X and Y is:

$$Cov(X, Y) = \sum_{(x,y)\in S} \sum (x - \mu_X)(y - \mu_Y) f(x, y)$$

And, if X and Y are continuous random variables with supports S_1 and S_2, respectively, then the covariance of X and Y is:

$$Cov(X, Y) = \int_{S_2} \int_{S_1} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy$$

..., where f(x,y) is either the joint probability mass or density function.

Ref.: https://onlinecourses.science.psu.edu/stat414/node/109

CONDITIONAL DISTRIBUTION

- Recall that for any two events E and F
  - $P(E \mid F) = \frac{P(EF)}{P(F)}$, provided P(F) > 0
- Let X and Y have joint PMF, p(x,y)
- The **conditional PMF of X** given that Y = y is
  - $P_{X|Y}(x \mid y) = \frac{p(x,y)}{P_Y(y)}$
- Similarly, the **conditional PMF of Y** given that X = x is
  - $P_{Y|X}(y \mid x) = \frac{p(x,y)}{P_X(x)}$

PARAMETRIC POINT ESTIMATION

- The general functional form of the pmf or pdf which underlies the data is assumed to be known, but the parent distribution depends upon a (possibly multidimensional) parameter θ, whose value depends to a specified set of values Θ called the parameter space
- The value of θ is unknown
- It is desired to estimate the unknown parameter θ (or perhaps some function of θ) & we want the estimate to be as accurate as possible

LIKELIHOOD FUNCTION

- Suppose that the distribution of X_i depends upon a parameter θ
- Let $p_{x_i}(x_i; \theta)$ be the pmf or pdf of the distribution of the X_i
- If the r.v.'s are mutually independent, the joint pmf or pdf is just
  - $\prod_{i=1}^{n} p(x_i;\theta)$
- Plugging in the observed values $x_1, x_2, \ldots, x_n$ and viewing this as a function of θ, we refer to this as the **likelihood function**
- $L(\theta;\vec{x}) = L(\theta;x_1,\ldots,x_n)$

MAXIMUM LIKELIHOOD ESTIMATE

- Loosely speaking, the method of maximum likelihood seeks to determine the value $\hat{\theta}(\vec{x})$ of the parameter space, which is most compatible with the data
- More formally, if $\vec{x}$ is the observed value of $\vec{X} = (X_1, \ldots, X_n)$ we seek $\hat{\theta}(\vec{x})$ which satisfies
  - $L(\hat{\theta}(\vec{x}); \vec{x}) = \max\{L(\theta; \vec{x}) \mid \theta \in \Theta\}$
- If such a $\hat{\theta}$ exists, we estimate θ by $\hat{\theta}$, i.e., the MLE of θ
- If $\hat{\theta}(\vec{x})$ can be written as a function of the x_i, then the corresponding function of the X_i is the maximum likelihood estimator (or MLE)

EXERCISE 11: PMF

- An urn contains both some black balls and some white balls
- It is known that the ratio of the number of balls of one color to the number of balls of the other color is 3/1
- It is unknown whether black or white balls are more numerous
- All we know for sure is that the probability of drawing a black ball from the full urn is either ¼ or ¾ (i.e., the parameter space)

PROBABILISTIC MODEL

- If we randomly draw with replacement, three balls from the urn (on each draw, giving all of the balls the same chance of being chosen), and if we let

$$X_i = \begin{cases} 1 & \text{if the i}^{\text{th}}\text{ draw results in a black ball} \\ 0 & \text{if the i}^{\text{th}}\text{ draw results in a white ball} \end{cases}$$

- Then X_1, X_2, X_3 are iid Bernoulli(p) r.v.'s with *p* being either ¼ or ¾

SETUP

- Likelihood function for each possible outcome is $\vec{X} = (X_1, X_2, X_3)$

	$\vec{x} : \sum_{i=1}^{3} x_i = 0$	$\vec{x} : \sum_{i=1}^{3} x_i = 1$	$\vec{x} : \sum_{i=1}^{3} x_i = 2$	$\vec{x} : \sum_{i=1}^{3} x_i = 3$
$L(1/4; \vec{x})$				
$L(3/4; \vec{x})$				

- (Exercise) Compute the pmf for $\vec{x}$

SOLUTION 11

- Likelihood function for each possible outcome is $\vec{X} = (X_1, X_2, X_3)$

	$\vec{x} : \sum_{i=1}^{3} x_i = 0$	$\vec{x} : \sum_{i=1}^{3} x_i = 1$	$\vec{x} : \sum_{i=1}^{3} x_i = 2$	$\vec{x} : \sum_{i=1}^{3} x_i = 3$
$L(1/4; \vec{x})$	$\frac{27}{64}$	$\frac{9}{64}$	$\frac{3}{64}$	$\frac{1}{64}$
$L(3/4; \vec{x})$	$\frac{1}{64}$	$\frac{3}{64}$	$\frac{9}{64}$	$\frac{27}{64}$

- Each row gives a likelihood function
- For example, if $x_1 = 0$, $x_2 = 1$, $x_3 = 0$, then $L(1/4; \vec{x}) = \left(\frac{3}{4}\right)\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) = \frac{9}{64}$
- Each row represents a pmf for $\vec{x}$

EXERCISE 12

Let us consider two boxes. Box 1 contains 30 white balls and 10 black balls. Box 2 contains 20 white and 20 black balls.

Someone draws one ball at random (among the 80 balls). The drawn ball is white. What is the probability that the ball was drawn from the first box ?

SOLUTION 12

Let us consider two boxes. Box 1 contains 30 white balls and 10 black balls. Box 2 contains 20 white and 20 black balls.

Someone draws one ball at random (among the 80 balls). The drawn ball is white. What is the probability that the ball was drawn from the first box ?

Solution Let us denote by B_1 (resp. B_2) the event that the ball was drawn from box 1 (resp. 2). Let us denote by W the event that the drawn ball was white.

We are looking for:

$$\mathbb{P}(B_1 \mid W) = \frac{\mathbb{P}(B_1, W)}{\mathbb{P}(W)}.$$

We compute:

$$\mathbb{P}(B_1, W) = \mathbb{P}(W \mid B_1)\,\mathbb{P}(B_1) = \frac{3}{4} \times \frac{1}{2} = \frac{3}{8},$$

and:

$$\mathbb{P}(W) = \frac{50}{80} = \frac{5}{8}.$$

We conclude:

$$\mathbb{P}(B_1 \mid W) = \frac{\frac{3}{8}}{\frac{5}{8}} = 0.6.$$

EXERCISE 13: CONDITIONAL PROBABILITIES

- A real estate office uses three print shops to make brochures
- 10% of the brochures are made by *Speedy Printers*
- 35% are made by *AAA*
- 55% are made by *PrintRite*
- Speedy makes brochures with errors 5% of the time
- AAA makes errors 2% of the time
- PrintRite makes errors 1% of the time

1. What is the total % of brochures with errors?

2. Examining the brochures, one if found to have an error. Find the probability that it came from Speedy Printers.

SETUP

- Let A = brochure produced by Speedy
- Let B = brochure produced by AAA
- Let C = brochure produced by PrintRite
- Let E = brochure has error
- P(A) = … P(E|A) = …
- P(B) = … P(E|B) = …
- P(C) = … P(E|C) = …

1. What is the total % of brochures with errors?

- P(E) = …

2. Examining the brochures, one is found to have an error. Find the probability that it came from Speedy Printers.

- P(A|E) = …

SOLUTION 13

- Let A = brochure produced by Speedy
- Let B = brochure produced by AAA
- Let C = brochure produced by PrintRite
- Let E = brochure has error
- P(A) = 0.1 P(E|A) = 0.05
- P(B) = 0.35 P(E|B) = 0.02
- P(C) = 0.55 P(E|C) = 0.01

- What is the total % of brochures with errors?
  - P(E) = P(E|A)P(A) + P(E|B)P(B) + P(E|C)P(C) = 0.0175 = 1.75%

- Examining the brochures, one is found to have an error. Find the probability that it came from Speedy Printers.
  - P(A|E) = P(AE)/P(E) = P(E|A)P(A)/P(E) = .05*.1/.0175 = .2857

EXERCISE 14: STATISTICAL MEASURES

- Is the mean robust against outliers? True or False
- Is the median robust against outliers? True or False
- Is the standard deviation robust against outliers? True or False

SOLUTION 14

- Is the mean robust against outliers? False
- Is the median robust against outliers? True
- Is the standard deviation robust against outliers? False

LONG TAIL DISTRIBUTION

- Feature in some probability distributions
  - Zipf, Power, Pareto
- In long-tailed distributions high-frequency population is followed by a low-frequency population, which gradually tails off asymptotically
  - events at the far end of the tail have a very low probability of occurrence
- Relevance
  - collaborative filters (recommender systems)
  - effect observed in some recommender systems that illustrate biases in data

This power law graph illustrates popularity ranking. The right (yellow) is the long tail; the left (green) are the few that dominate. In this example, the areas of both regions are equal.

PROBABILITY & STATISTICS COOKBOOK

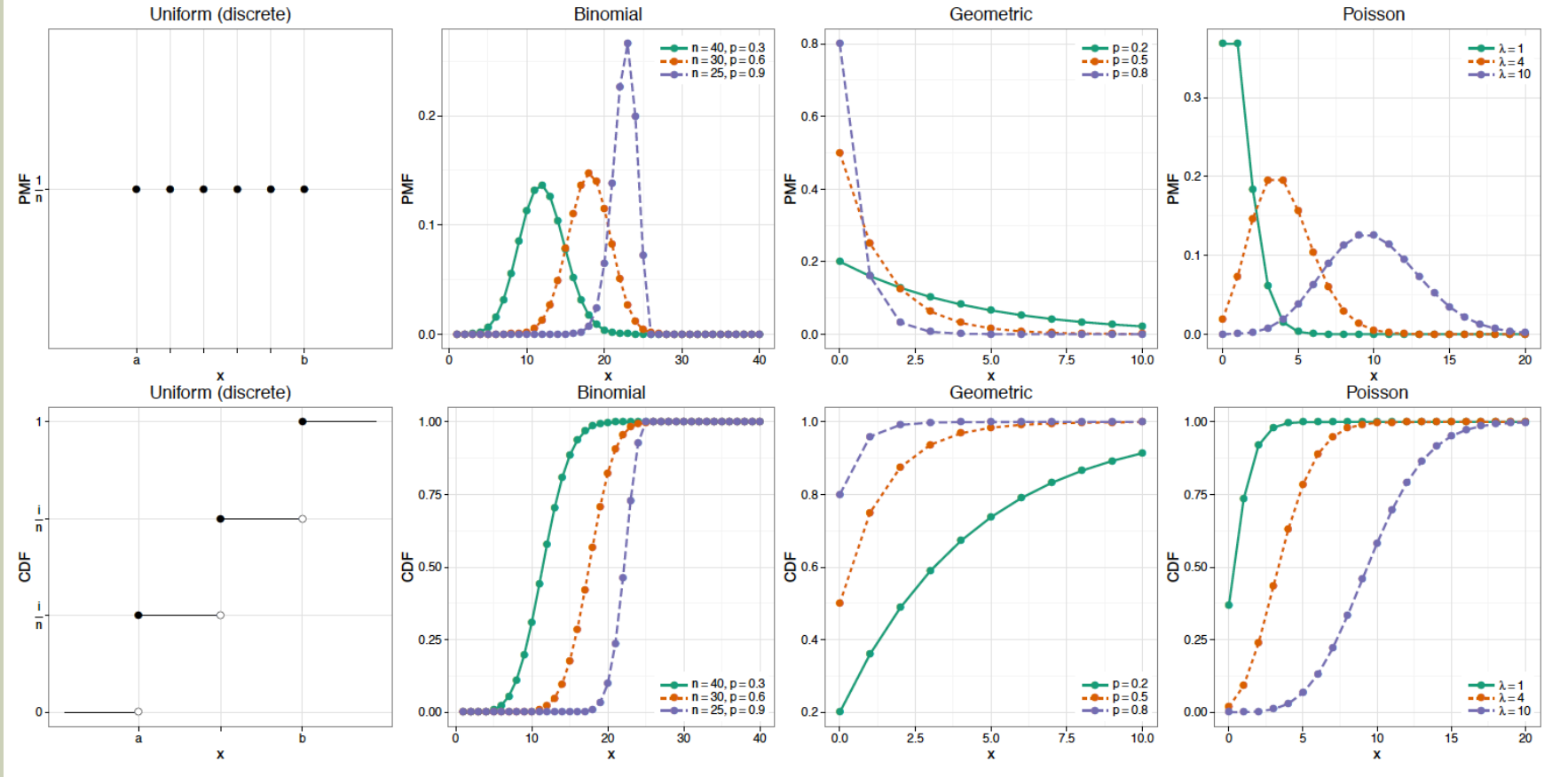
1 Discrete Distributions

	Notation[1]	$F_X(x)$	$f_X(x)$	$\mathbb{E}[X]$	$\mathbb{V}[X]$	$M_X(s)$
Uniform	$\text{Unif}\{a,\dots,b\}$	$\begin{cases} 0 & x < a \\ \frac{\lfloor x \rfloor - a + 1}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$	$\frac{I(a \le x \le b)}{b-a+1}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$	$\frac{e^{as}-e^{-(b+1)s}}{s(b-a)}$
Bernoulli	$\text{Bern}(p)$	$(1-p)^{1-x}$	$p^x(1-p)^{1-x}$	p	$p(1-p)$	$1-p+pe^s$
Binomial	$\text{Bin}(n,p)$	$I_{1-p}(n-x,x+1)$	$\binom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$	$(1-p+pe^s)^n$
Multinomial	$\text{Mult}(n,p)$		$\frac{n!}{x_1!\dots x_k!}p_1^{x_1}\cdots p_k^{x_k} \quad \sum_{i=1}^k x_i = n$	$\begin{pmatrix} np_1 \\ \vdots \\ np_k \end{pmatrix}$	$\begin{pmatrix} np_1(1-p_1) & -np_1p_2 \\ -np_2p_1 & \ddots \end{pmatrix}$	$\left(\sum_{i=0}^k p_i e^{s_i}\right)^n$
Hypergeometric	$\text{Hyp}(N,m,n)$	$\approx \Phi\left(\frac{x-np}{\sqrt{np(1-p)}}\right)$	$\frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$	$\frac{nm}{N}$	$\frac{nm(N-n)(N-m)}{N^2(N-1)}$	
Negative Binomial	$\text{NBin}(r,p)$	$I_p(r,x+1)$	$\binom{x+r-1}{r-1}p^r(1-p)^x$	$r\frac{1-p}{p}$	$r\frac{1-p}{p^2}$	$\left(\frac{p}{1-(1-p)e^s}\right)^r$
Geometric	$\text{Geo}(p)$	$1-(1-p)^x \quad x \in \mathbb{N}^+$	$p(1-p)^{x-1} \quad x \in \mathbb{N}^+$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^s}{1-(1-p)e^s}$
Poisson	$\text{Po}(\lambda)$	$e^{-\lambda}\sum_{i=0}^x \frac{\lambda^i}{i!}$	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	λ	$e^{\lambda(e^s-1)}$

[1] We use the notation $\gamma(s,x)$ and $\Gamma(x)$ to refer to the Gamma functions (see §22.1), and use $\text{B}(x,y)$ and I_x to refer to the Beta functions (see §22.2).

github.com/mavam/stat-cookbook/releases/download/0.2.3/stat-cookbook.pdf

PLOTS OF THE PMF AND CDF



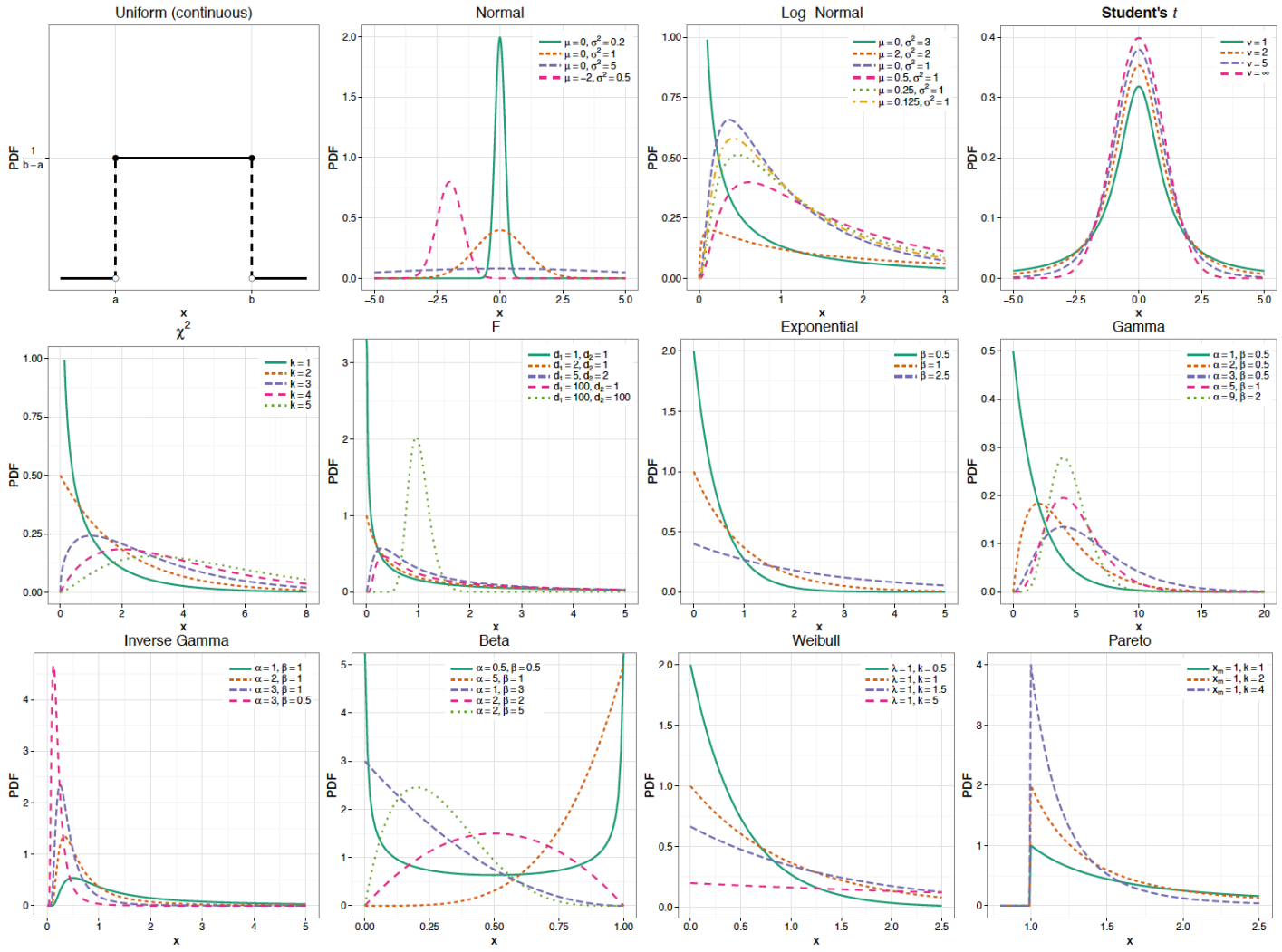
PROBABILITY & STATISTICS COOKBOOK

1.2 Continuous Distributions

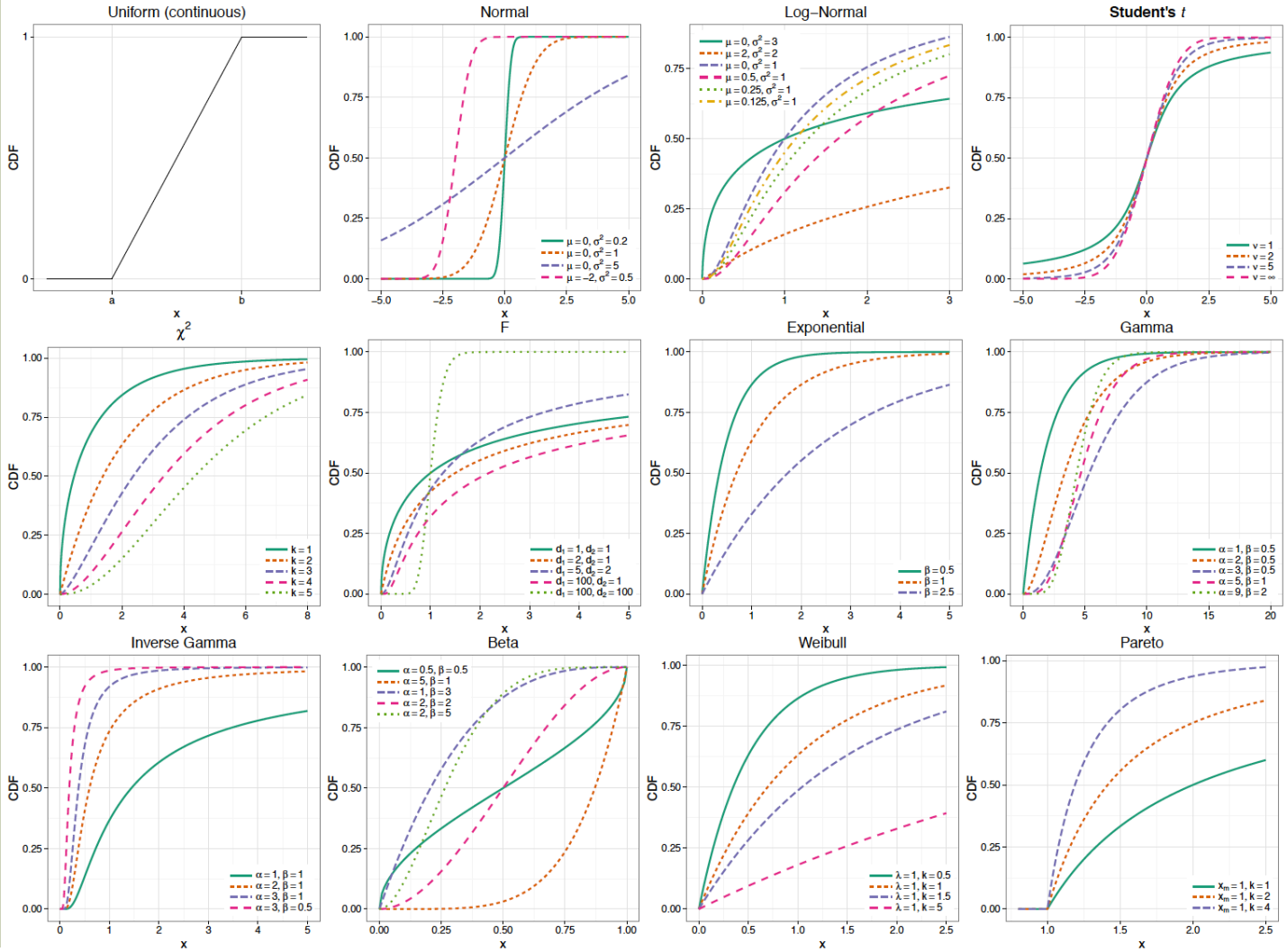
	Notation	$F_X(x)$	$f_X(x)$	$\mathbb{E}[X]$	$\mathbb{V}[X]$	$M_X(s)$
Uniform	$\text{Unif}(a,b)$	$\begin{cases} 0 & x<a \\ \frac{x-a}{b-a} & a<x<b \\ 1 & x>b \end{cases}$	$\frac{I(a<x<b)}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{sb}-e^{sa}}{s(b-a)}$
Normal	$\mathcal{N}(\mu,\sigma^2)$	$\Phi(x)=\int_{-\infty}^{x}\phi(t)\,dt$	$\phi(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$	μ	σ^2	$\exp\left\{\mu s+\frac{\sigma^2 s^2}{2}\right\}$
Log-Normal	$\ln\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{2}+\frac{1}{2}\,\text{erf}\left[\frac{\ln x-\mu}{\sqrt{2\sigma^2}}\right]$	$\frac{1}{x\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\}$	$e^{\mu+\sigma^2/2}$	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$	
Multivariate Normal	$\text{MVN}(\mu,\Sigma)$		$(2\pi)^{-k/2}\lvert\Sigma\rvert^{-1/2}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$	μ	Σ	$\exp\left\{\mu^T s+\frac{1}{2}s^T\Sigma s\right\}$
Student's t	$\text{Student}(\nu)$	$I_x\left(\frac{\nu}{2},\frac{\nu}{2}\right)$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}$	$0 \quad \nu>1$	$\begin{cases} \frac{\nu}{\nu-2} & \nu>2 \\ \infty & 1<\nu\le 2 \end{cases}$	
Chi-square	χ^2_k	$\frac{1}{\Gamma(k/2)}\gamma\left(\frac{k}{2},\frac{x}{2}\right)$	$\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}$	k	$2k$	$(1-2s)^{-k/2}\ s<1/2$
F	$\text{F}(d_1,d_2)$	$I_{\frac{d_1 x}{d_1 x+d_2}}\left(\frac{d_1}{2},\frac{d_1}{2}\right)$	$\frac{\sqrt{\frac{(d_1 x)^{d_1}d_2^{d_2}}{(d_1 x+d_2)^{d_1+d_2}}}}{x\text{B}\left(\frac{d_1}{2},\frac{d_1}{2}\right)}$	$\frac{d_2}{d_2-2}$	$\frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$	
Exponential*	$\text{Exp}(\beta)$	$1-e^{-x/\beta}$	$\frac{1}{\beta}e^{-x/\beta}$	β	β^2	$\frac{1}{1-\frac{s}{\beta}}\ (s<\beta)$
Gamma*	$\text{Gamma}(\alpha,\beta)$	$\frac{\gamma(\alpha,\beta x)}{\Gamma(\alpha)}$	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\left(\frac{1}{1-\frac{s}{\beta}}\right)^\alpha\ (s<\beta)$
Inverse Gamma	$\text{InvGamma}(\alpha,\beta)$	$\frac{\Gamma\left(\alpha,\frac{\beta}{x}\right)}{\Gamma(\alpha)}$	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}e^{-\beta/x}$	$\frac{\beta}{\alpha-1}\ \alpha>1$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}\ \alpha>2$	$\frac{2(-\beta s)^{\alpha/2}}{\Gamma(\alpha)}K_\alpha\left(\sqrt{-4\beta s}\right)$
Dirichlet	$\text{Dir}(\alpha)$		$\frac{\Gamma\left(\sum_{i=1}^k\alpha_i\right)}{\prod_{i=1}^k\Gamma(\alpha_i)}\prod_{i=1}^k x_i^{\alpha_i-1}$	$\frac{\alpha_i}{\sum_{i=1}^k\alpha_i}$	$\frac{\mathbb{E}[X_i](1-\mathbb{E}[X_i])}{\sum_{i=1}^k\alpha_i+1}$	
Beta	$\text{Beta}(\alpha,\beta)$	$I_x(\alpha,\beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$1+\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{s^k}{k!}$
Weibull	$\text{Weibull}(\lambda,k)$	$1-e^{-(x/\lambda)^k}$	$\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k}$	$\lambda\Gamma\left(1+\frac{1}{k}\right)$	$\lambda^2\Gamma\left(1+\frac{2}{k}\right)-\mu^2$	$\sum_{n=0}^{\infty}\frac{s^n\lambda^n}{n!}\Gamma\left(1+\frac{n}{k}\right)$
Pareto	$\text{Pareto}(x_m,\alpha)$	$1-\left(\frac{x_m}{x}\right)^\alpha\ x\ge x_m$	$\alpha\frac{x_m^\alpha}{x^{\alpha+1}}\ x\ge x_m$	$\frac{\alpha x_m}{\alpha-1}\ \alpha>1$	$\frac{x_m^2\alpha}{(\alpha-1)^2(\alpha-2)}\ \alpha>2$	$\alpha(-x_m s)^\alpha\Gamma(-\alpha,-x_m s)\ s<0$

* We use the *rate* parameterization where $\beta=\frac{1}{\lambda}$. Some textbooks use β as *scale* parameter instead [6].

PLOTS OF THE PDF

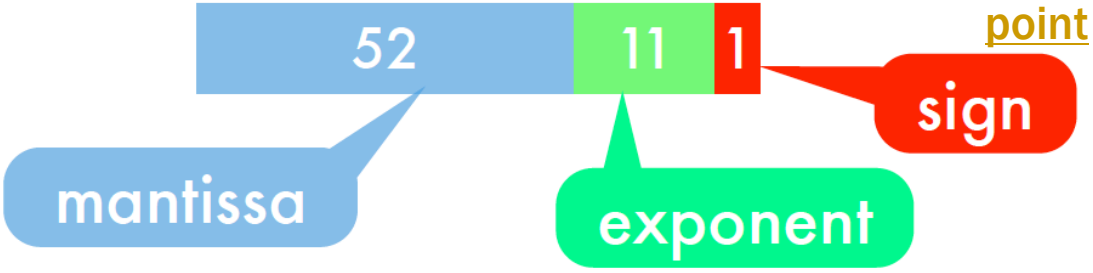


PLOTS OF THE CDF



Logarithms are good

- Floating point numbers



blog.smola.org/post/987977550/log-probabilities-semirings-and-floating-point

$$\pi = \log p$$

- Probabilities can be very small. In particular products of many probabilities. Underflow!
- Store data in mantissa, not exponent

$$\prod_i p_i \to \sum_i \pi_i \qquad \sum_i p_i \to \max \pi + \log \sum_i \exp\left[\pi_i - \max \pi\right]$$

- Known bug e.g. in Mahout Dirichlet clustering

MACHINE LEARNING

ML ALGORITHM: BASICS

- ML algorithms consist both of
  - a loss (or cost) function and
  - an optimization technique
- Loss is the penalty incurred when the estimate of the target provided by the ML model does not equal the target exactly
  - loss functions quantify this penalty as a single value
  - examples include a *logistic loss* function and a *squared loss* function
- Optimization techniques seek to minimize the loss
  - (e.g., stochastic gradient descent)

Given a prediction (p) and a label (y), a loss function $\ell(p, y)$ measures the discrepancy between the algorithm's prediction and the desired output. VW currently supports the following loss functions, with squared loss being the default:

Loss	Function	Minimizer	Example usage
Squared	$\frac{1}{2}(p-y)^2$	Expectation (mean)	Regression *Expected return on stock*
Quantile	$\tau(p-y)\mathbb{I}(y \le p) + (1-\tau)(y-p)\mathbb{I}(y \ge p)$	Median	Regression *What is a typical price for a house?*
Logistic	$\log(1+\exp(-yp))$	Probability	Classification *Probability of click on ad*
Hinge	$\max(0, 1-yp)$	0-1 approximation	Classification *Is the digit a 7?*
Classic	Squared loss without importance weight aware updates	Expectation (mean)	Regression *squared loss often performs better than classic.*

TRAINING PARAMETERS

- Learning algorithms require hyperparameters (i.e., training parameters) that allow you to control the quality of the resulting model

- Common hyperparameters associated with learning algorithms
  - *learning rate* - impacts the speed of the algorithm converging to the optimal weights
  - *regularization* - helps prevent models from overfitting training data examples by penalizing extreme weight values
  - number of passes – determining the optimum number of sequential passes over the training data

CLASSIFICATION: NAÏVE BAYES

Theory

A naive Bayes classifier models a joint distribution over a label Y and a set of observed random variables, or *features*, $(F_1, F_2, \ldots, F_n)$, using the assumption that the full joint distribution can be factored as follows (features are conditionally independent given the label):

$$P(F_1, \ldots, F_n, Y) = P(Y) \prod_i P(F_i|Y)$$

To classify a datum, we can find the most probable label given the feature values for each pixel, using Bayes theorem:

$$\begin{aligned} P(y|f_1, \ldots, f_m) &= \frac{P(f_1, \ldots, f_m|y)P(y)}{P(f_1, \ldots, f_m)} \\ &= \frac{P(y)\prod_{i=1}^{m} P(f_i|y)}{P(f_1, \ldots, f_m)} \end{aligned}$$

$$\begin{aligned} \arg\max_y P(y|f_1, \ldots, f_m) &= \arg\max_y \frac{P(y)\prod_{i=1}^{m} P(f_i|y)}{P(f_1, \ldots, f_m)} \\ &= \arg\max_y P(y)\prod_{i=1}^{m} P(f_i|y) \end{aligned}$$

posterior

prior likelihood

Because multiplying many probabilities together often results in underflow, we will instead compute ***log probabilities*** which have the same argmax:

$$\begin{aligned} \arg\max_y \log P(y|f_1, \ldots, f_m) &= \arg\max_y \log P(y, f_1, \ldots, f_m) \\ &= \arg\max_y \left\{ \log P(y) + \sum_{i=1}^{m} \log P(f_i|y) \right\} \end{aligned}$$

Using this rule, we can calculate the most probable estimate of Y.

OPTIMIZATION

CONVEX OPTIMIZATION

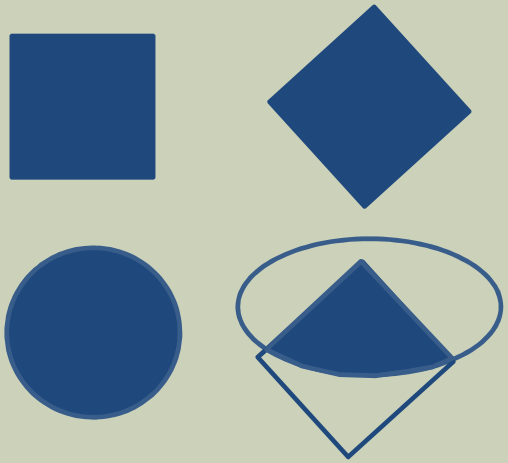
- Find minimum of a function subject to solution constraints
- Business/economics/ game theory
  - Resource allocation
  - Optimal planning and strategies
- Statistics and Machine Learning
  - All forms of regression and classification
  - Unsupervised learning

CONVEX SETS

- A set C is convex if $\forall x, y \in C$ and $\forall \alpha \in [0,1]$

$$\alpha x + (1 - \alpha) y \in C$$

  - Line segment between points in C also lies in C
- **Examples**
  - Intersection of halfspaces
  - L_p balls
  - Intersection of convex sets



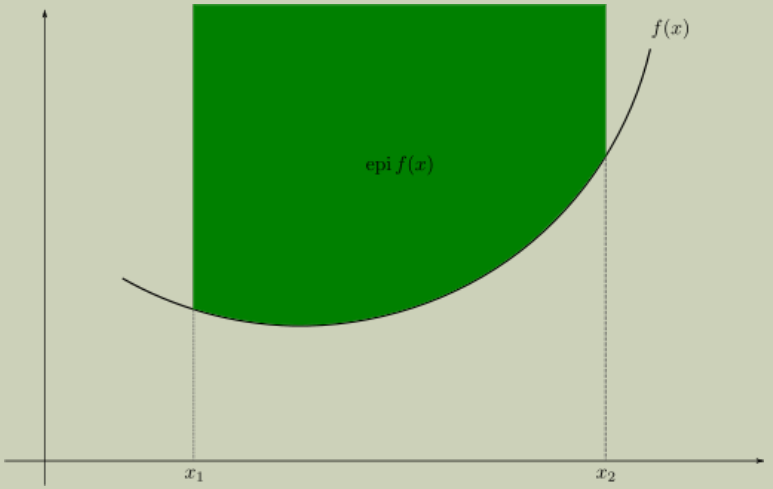
CONVEX FUNCTIONS

A real-valued function f is convex if $\mathrm{dom} f$ *is* convex and $\forall x, y \in \mathrm{dom} f$ and $\forall \alpha \in [0,1]$

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

That is, *f* is convex if the line segment between any two points on the graph of the function lies above the graph.

Examples: $f(x) = x^2$ or $f(x) = e^x$

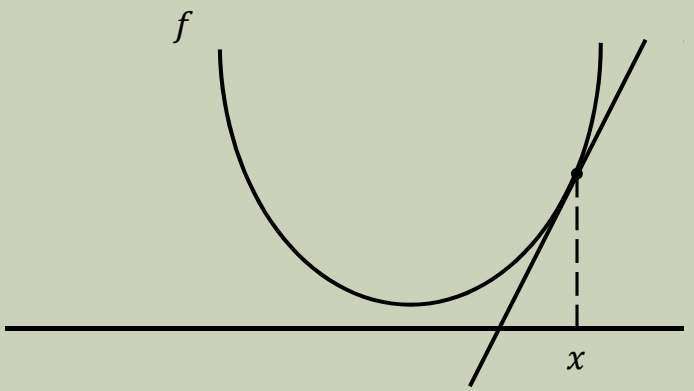


GRADIENT

- **Gradient** is a generalization of the derivative of a function in several dimensions
- If $f(x_1,\ldots,x_n)$ is a *differentiable*, scalar-valued function in R^n, its gradient is a vector whose components are the *n* partial derivatives of f
- Gradient represents the slope of the tangent of the graph of the function

Gradient ∇f at x gives linear approximation

$$\nabla f = \left[\frac{\delta f}{\delta x_1} \quad \ldots \quad \frac{\delta f}{\delta x_d}\right]^T$$



For example, in R^3, the linear approximation of a function f(x) at a point (x_0,y_0,z_0) is

$$L(x) = f(x_0,y_0,z_0) + f_x(x_0,y_0,z_0)(x-x_0) + f_y(x_0,y_0,z_0)(y-y_0) + f_z(x_0,y_0,z_0)(z-z_0)$$

GRADIENT DESCENT

- To minimize f move down gradient
  - But not too far!
  - Optimum when $\nabla f = 0$
- Given f, learning rate α, starting point x_0

  $x = x_0$

  Do until $\nabla f = 0$

  $$x = x - \alpha \nabla f$$

Technique used to find a local minimum in *f*.

STOCHASTIC GRADIENT DESCENT (1/2)

- Many learning problems have extra structure

$$f(\theta) = \sum_{i=1}^{n} L(\theta; \boldsymbol{x}_i)$$

- Computing gradient requires iterating over all points, can be too costly
- Instead, compute gradient at single training example

STOCHASTIC GRADIENT DESCENT (2/2)

- Given $f(\theta) = \sum_{i=1}^{n} L(\theta; \boldsymbol{x}_i)$, learning rate α, starting point θ_0

  $\theta = \theta_0$

  Do until $f(\theta)$ nearly optimal

  For $i = 1$ to n <u>in random order</u>

  $$\theta = \theta - \alpha \nabla L(\theta; \boldsymbol{x}_i)$$

- Finds nearly optimal θ

LEARNING ALGORITHMS

- Stochastic Gradient Descent (SGD)
- Alternating Least Squares (ALS)
- SGD and ALS are two approaches used in collaborative filtering for recommender systems
- The goal is to estimate factor vectors, q_i and p_u corresponding to **items** and **users**, respectively, given existing **ratings** (r_{ui})
- This is formulated as an optimization problem as depicted below

$$\min_{q^*,p^*} \sum_{(u,i)\in\kappa} (r_{ui} - q_i^T p_u)^2 + \lambda(\| q_i \|^2 + \| p_u \|^2)$$

OPTIMIZATION IN RECOMMENDER SYSTEMS

- Collaborative filtering based recommender systems employ learning algorithms, such as SGD & ALS to solve the following min. problem:

$$\min_{q^*,p^*} \sum_{(u,i)\in K} (r_{ui} - q_i^T p_u)^2 + \lambda(\| q_i \|^2 + \| p_u \|^2)$$

- Given training data, i.e., existing ratings (r_{ui}), we seek to estimate factor vectors, q_i and p_u corresponding to items (i) and users (u), respectively
- **Stochastic Gradient Descent (SGD)**
  - loops through all ratings in a training set, computes the associated prediction error, & adjusts parameters using the gradient at a single example
- **Alternating Least Squares (ALS)**
  - rotates between fixing one of the unknowns q_i or p_u; when one is fixed the other can be computed by solving the least-squares problem

The **general least-squares problem** is to find an $\mathbf{x}$ that makes $\|\mathbf{b} - A\mathbf{x}\|$ as small as possible. If A is $m \times n$ and $\mathbf{b}$ is in $\mathbb{R}^m$, a **least-squares solution** of $A\mathbf{x} = \mathbf{b}$ is an $\hat{\mathbf{x}}$ in $\mathbb{R}^n$ such that $\|\mathbf{b} - A\hat{\mathbf{x}}\| \leq \|\mathbf{b} - A\mathbf{x}\|$ for all $\mathbf{x}$ in $\mathbb{R}^n$.

MATHEMATICAL LIBRARIES

Freely Available Software for Linear Algebra (September 2016)

Here is a list of freely available software for the solution of linear algebra problems. The interest is in software for high-performance computers that's available in "open source" form on the web for solving problems in numerical linear algebra, specifically dense, sparse direct and iterative systems, and sparse iterative eigenvalue problems. Please let us know about updates and corrections.

A survey of Iterative Linear System Solver Packages can be found at:
http://www.netlib.org/utk/papers/iterative-survey/

Thanks,
Jack and
Ahmad

SUPPORT ROUTINES	License	Support	Type		Language			Mode			Dense	Sparse
			Real	Complex	F77/ F95	C	C++	Shared	Accel.	Dist		
Armadillo	Mozilla	yes	X	X			X	X			X	
Armas	LGPL	yes	X			X		X			X	
ATLAS	BSD like	yes	X	X	X	X		X			X	
BLAS	PD	yes	X	X	X	X		X			X	
BLIS	New BSD	yes	X	X	X	X		X			X	
Blitz++	LGPLv3+	yes	X	X			X	X			X	
clMath	Apache	yes	X	X		X	X	X	O		X	
KBLAS	BSD	yes	X	X		X	X	X	C		X	
librsb	LGPLv3	yes	X	X	X	X	X	X				X
LINALG *	?	?										
MR3-SMP	New BSD	yes	X	X	X	X		X			X	
MTL	Own	yes	X				X	X			X	
NEWMAT	Own	yes	X				X	X			X	
NIST Sparse BLAS	PD	yes	X	X		X	X	X				X
OpenBLAS	BSD	yes	X	X	X	X		X			X	

http://www.netlib.org/utk/people/JackDongarra/la-sw.html

Reading Assignment

MATRIX FACTORIZATION TECHNIQUES FOR RECOMMENDER SYSTEMS

Yehuda Koren, *Yahoo Research*
Robert Bell and Chris Volinsky, *AT&T Labs—Research*

As the Netflix Prize competition has demonstrated, matrix factorization models are superior to classic nearest-neighbor techniques for producing product recommendations, allowing the incorporation of additional information such as implicit feedback, temporal effects, and confidence levels.

Modern consumers are inundated with choices. Electronic retailers and content providers offer a huge selection of products, with unprecedented opportunities to meet a variety of special needs and tastes. Matching consumers with the most appropriate products is key to enhancing user satisfaction and loyalty. Therefore, more retailers have become interested in recommender systems, which analyze patterns of user interest in products to provide personalized recommendations that suit a user's taste. Because good personalized recommendations can add another dimension to the user experience, e-commerce leaders like Amazon.com and Netflix have made recommender systems a salient part of their websites.

Such systems are particularly useful for entertainment products such as movies, music, and TV shows. Many customers will view the same movie, and each customer is likely to view numerous different movies. Customers have proven willing to indicate their level of satisfaction with particular movies, so a huge volume of data is available about which movies appeal to which customers. Companies can analyze this data to recommend movies to particular customers.

RECOMMENDER SYSTEM STRATEGIES

Broadly speaking, recommender systems are based on one of two strategies. The *content filtering* approach creates a profile for each user or product to characterize its nature. For example, a movie profile could include attributes regarding its genre, the participating actors, its box office popularity, and so forth. User profiles might include demographic information or answers provided on a suitable questionnaire. The profiles allow programs to associate users with matching products. Of course, content-based strategies require gathering external information that might not be available or easy to collect.

A known successful realization of content filtering is the Music Genome Project, which is used for the Internet radio service Pandora.com. A trained music analyst scores

Published by the IEEE Computer Society

REFERENCE SOURCES

REFERENCE SOURCES

- Some of the slide content was drawn from these sources:
  - dws.informatik.uni-mannheim.de/en/teaching/courses-for-master-candidates/ie-673-data-mining-and-matrices/
  - databricks.com/blog/2014/07/21/distributing-the-singular-value-decomposition-with-spark.html
  - web.mit.edu/be.400/www/SVD/Singular_Value_Decomposition.htm
  - www.d.umn.edu/~mhampton/m4326svd_example.pdf
  - matthias.vallentin.net/probability-and-statistics-cookbook/cookbook-en.pdf
  - blog.smola.org/post/987977550/log-probabilities-semirings-and-floating-point
  - www.netlib.org/utk/people/JackDongarra/la-sw.html
  - gams.nist.gov/cgi-bin/serve.cgi/Class/D2
  - math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf
  - math.mit.edu/~gs/linearalgebra/