Announcements

- HW0 due Fri/Sun
- Fri TA session (different room)
- HW1 will be posted Sun night
- read VMLS book ...

EE445 Mod1-Lec2: Linear Algebra II

References:

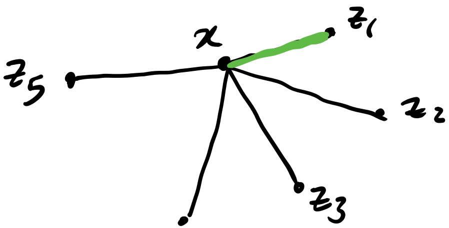
- [VMLS]: Chapters 3, 4, 5
- Topics: Distance and angle, Clustering example (and k-means), Basis & orthonormal vectors

Feature distance and nearest neighbors

$x \in \mathbb{R}^n$ $\quad \|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}$ $\quad \text{dist}(x,y) = \|x-y\|$

- if x, y are feature vectors for two entities, $\|x - y\|$ is the *feature distance*
- for vectors $z_1, \ldots, z_m$, z_j is nearest neighbor of x if

$$\|x - z_j\| \leq \|x - z_i\|, \quad i = 1, \ldots, m$$



- simple ideas that are widely used!

Example: document dissimilarity

- 5 Wikipedia articles: 'Veterans Day', 'Memorial Day', 'Academy Awards', 'Golden Globe Awards', 'Super Bowl'
- word count histograms, dictionary of 4423 words
- pairwise distances shown below:

(more setup details in VMLS sec. 4.4.5)

$$a = \begin{bmatrix} 0 \\ 0 \\ 10 \\ 3 \\ \vdots \end{bmatrix}, b, c, d, e \qquad \|a-b\|, \|a-c\|, \ldots$$

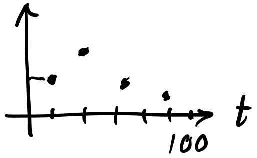
		Veterans' day	Memorial day	Academy A.	Golden G.	Super Bowl
a	Veterans' day	0	0.095	0.130	0.153	0.170
b	Memorial day	0.095	0	0.122	0.147	0.164
	Academy A.	0.130	0.122	0	0.108	0.164
	Golden G.	0.153	0.147	0.108	0	0.181
	Super Bowl	0.170	0.164	0.164	0.181	0

Standard deviation of vector x

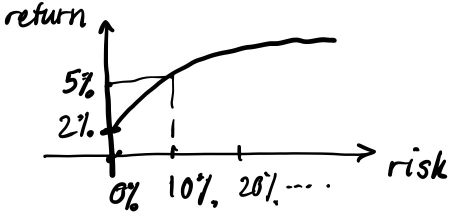
- for n-vector x, average of its entries is: $\mathbf{avg}(x) = \frac{\mathbf{1}^T x}{n}$
- de-meaned (or centered) vector: $\tilde{x} = x - \mathbf{avg}(x)\mathbf{1} = x - \frac{1}{n}\mathbf{1}^T x$
- standard deviation of x is: $\mathbf{std}(x) = \frac{1}{\sqrt{n}} \left\| x - \underbrace{\frac{\mathbf{1}^T x}{n}} \mathbf{1} \right\| = \frac{1}{\sqrt{n}} \sqrt{(x_i - \mathbf{avg}(x))^2}$

- $\mathbf{std}(x)$ measures the typical amount x_i vary from $\mathbf{avg}(x)$
- $\mathbf{std}(x) = 0$ only if $x = \alpha\mathbf{1}$
- notation: μ and σ commonly used for mean, standard deviation

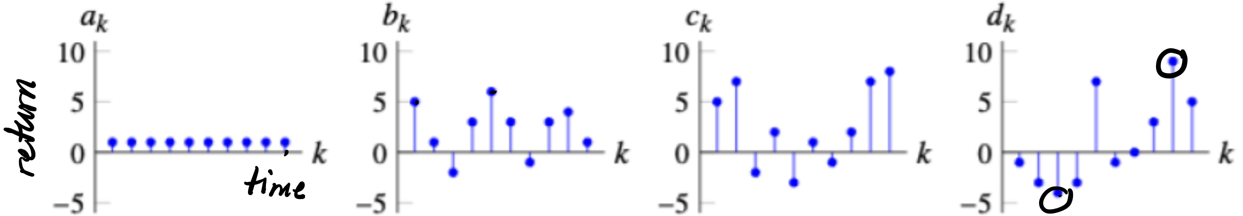
Example: Mean return and risk



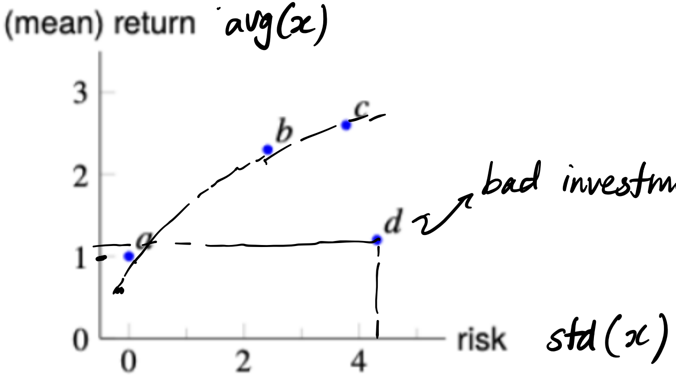
- x is time series of returns (say, in %) on some asset over some period
- $\mathbf{avg}(x)$ is the (mean) return over the period
- $\mathbf{std}(x)$ measures how variable the return is over the period, called the risk
- investments are often compared in terms of return and risk, plotted on a *risk-return* plot



Example: Mean return and risk tradeoff



for each vector a, b, c, d,
compute avg(·) and std(·)
and plot:



Cauchy-Schwartz inequality

- for $a, b \in \mathbf{R}^n$, $|a^T b| \leq \|a\|\|b\|$ or $-\|a\|\|b\| \leq a^T b \leq \|a\|\|b\|$
- written out:

$$|a_1 b_1 + \ldots + a_n b_n| \leq \left(a_1^2 + \ldots + a_n^2\right)^{1/2} \left(b_1^2 + \ldots + b_n^2\right)^{1/2}$$

- can show triangle inequality from this:

$$\begin{aligned} \|a+b\|^2 &= (a+b)^T(a+b) \\ &= a^T a + 2\underbrace{a^T b} + b^T b \\ &\leq \|a\|^2 + 2\|a\|\|b\| + \|b\|^2 \\ &= (\|a\| + \|b\|)^2 \end{aligned}$$

$\|a+b\| \leq \|a\| + \|b\|$ ← triangle ineq.

Derivation of Cauchy-Schwartz

- assume $\alpha = \|a\|$ and $\beta = \|b\|$ are nonzero (ineq. clearly true if either of these is 0)
- one way to derive:

$$
\begin{aligned}
0 \;&\leq\; \|\beta a - \alpha b\|^2 = (\beta a - \alpha b)^T(\beta a - \alpha b) \\
&= \beta^2 a^T a - 2\alpha\beta a^T b + \alpha^2 b^T b \\
&= \beta^2 \|a\|^2 - 2\alpha\beta (a^T b) + \alpha^2 \|b\|^2 \\
&= \|b\|^2\|a\|^2 - 2\|a\|\,\|b\|(a^T b) + \|a\|^2\|b\|^2
\end{aligned}
$$

divide by $\|a\|\,\|b\|$ to get:
$a^T b \leq \|a\|\,\|b\|$.

- apply to $-a$, b to get other half of Cauchy-Schwartz

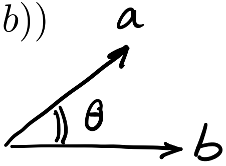
Angle

- *angle* between two nonzero vectors a, b is defined as

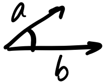
between -1 & 1 (from Cauchy-Schwartz

$$\angle(a,b) = \arccos\left(\overbrace{\frac{a^Tb}{\|a\|\|b\|}}\right)$$

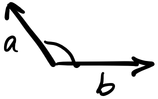
- $\angle(a,b)$ is the number in $[0,\pi]$ that satisfies $\quad a^Tb = \|a\|\|b\|\cos(\angle(a,b))$



- $\theta = \pi/2$: $a^Tb = 0$ orthogonal; $\quad \theta = 0$: a, b aligned
- acute or obtuse angle:



$a^Tb \geq 0$ or



$a^Tb \leq 0$

- spherical distance: if a, b are on a sphere with radius R, distance along the sphere is:

arc-length between a, b $= R\angle(a,b)$

Document dissimilarities by angles

- measure dissimilarity by *angle* between word count histogram vectors
- pairwise angles (in degrees) for the 5 Wikipedia pages:

close to 90°

	Veterans' day	Memorial day	Academy A.	Golden G.	Super Bowl
Veterans' day	0	60.6	85.7	87.0	87.7
Memorial day	60.6	0	85.6	87.5	87.5
Academy A.	85.7	85.6	0	58.7	86.1
Golden G.	87.0	87.5	58.7	0	86.0
Super Bowl	87.7	87.5	86.1	86.0	0

smallest angle (in this set)

Correlation coefficient

- consider vectors a, b and de-meaned vectors $\tilde{a}, \tilde{b}$

$$\tilde{a} = a - \mathbf{avg}(a)\mathbf{1}, \qquad \tilde{b} = b - \mathbf{avg}(b)\mathbf{1}$$

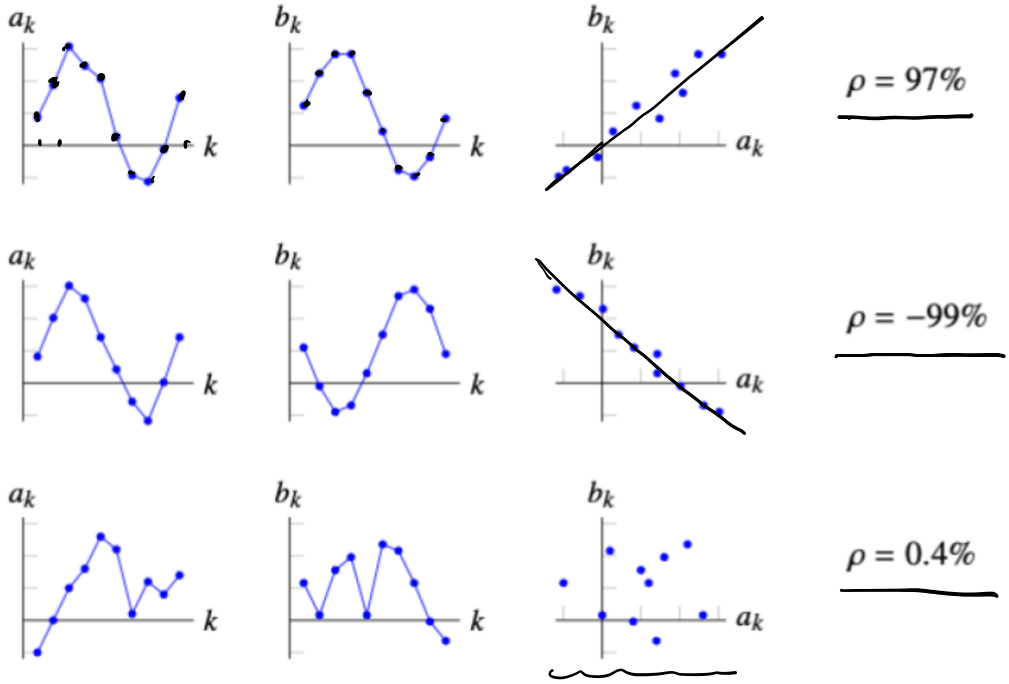
- correlation coefficient between a and b (with $\tilde{a}, \tilde{b} \neq 0$):

$$\rho = \frac{\tilde{a}^T \tilde{b}}{\|\tilde{a}\| \|\tilde{b}\|}$$

$-1 \leq \rho \leq 1$ (from Cauchy-Schwartz)

- $\rho = \cos(\angle \tilde{a}, \tilde{b})$
  - $\rho = 0$: a and b are uncorrelated
  - $\rho > 0.8$: a and b are highly correlated
  - $\rho < -0.8$: a and b are highly anti-correlated

Examples

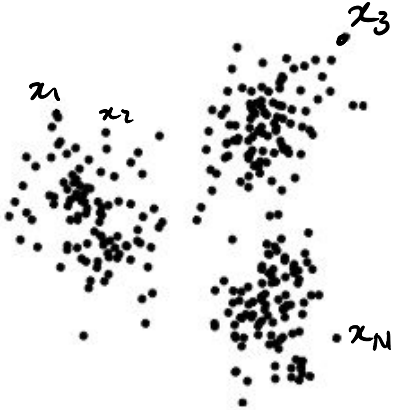


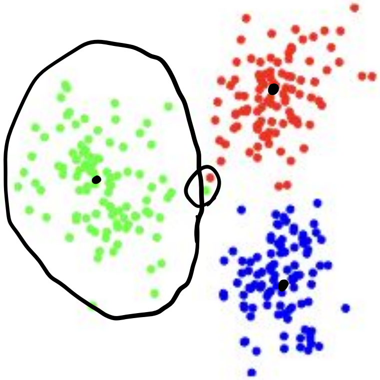
Clustering

$x_1 \in \mathbb{R}^n, \dots, x_N \in \mathbb{R}^n$

- given N n-vectors $x_1, \dots, x_N$, the goal is to cluster (partition) into k groups
- want vectors in the same group to be close
- examples: topic discovery/document classification; patient clustering;...

$k=3$:





Clustering objective

ex: $\{1,2,3,4,5\}$

$G_1 = \{1,4\}$ $G_2 = \{2,3\}$

$c_1 = 1$
$c_2 = 2$
$c_3 = 2$
$c_4 = 1$

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

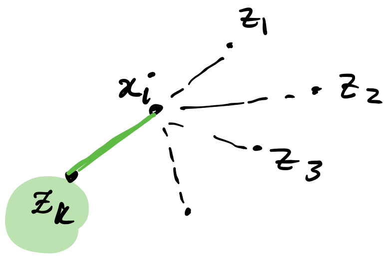
$$J = \frac{1}{N} \sum_{i=1}^{N} \|x_i - z_{c_i}\|^2$$

mean square distance from vectors to their group's representative

- **goal:** choose ~~clatering~~ clustering c_i and representatives z_j to minimize J

(we want to choose c_i and z_j jointly. Later we'll see that's a computationally intractable problem. But we can solve for one when the other is fixed...)

Partitioning vectors given representatives



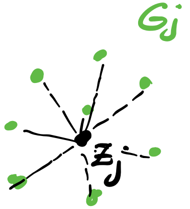
- suppose representatives $z_1, \ldots, z_k$ are given
- how do we assign vectors to groups, i.e., choose $c_1, \ldots, c_N$?

- c_i appears only in term $\|x_i - z_{c_i}\|^2$ (in objective J)
- to minimize, choose c_i so that $\|x_i - z_{c_i}\|^2 = \min_j \|x_i - z_j\|^2$
- i.e., assign each vector to its *nearest representative*

Choosing representatives given partition

these may not be among the x_i's

- given partition $G_1, \ldots, G_k$, how to choose representatives $z_1, \ldots, z_k$ to minimize J?
- J splits into sum of k sums:



$$J = J_1 + \ldots + J_k, \qquad J_j = 1/N \sum_{i \in G_j} \|x_i - z_j\|^2$$

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

$|G_j|$ = # of elements in set G_j

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

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

a heuristic algorithm

[see TA session on 4/8: clustering via k-means, applications]

Linear independence

- set of n-vectors $\{a_1, ..., a_k\}$ is *linearly dependent* if

$$\beta_1 a_1 + \ldots + \beta_k a_k = 0$$

  holds for some $\beta_1, \ldots, \beta_k$ that are not all zero
- equivalent to: at least one a_i is a linear combination of the others
- $\{a_1, a_2\}$ is linearly dependent only if one a_i is a multiple of the other
- set of n-vectors $\{a_1, ..., a_k\}$ is *linearly independent* if

$$\beta_1 a_1 + \ldots + \beta_k a_k = 0$$

  holds only when $\beta_1 = \ldots = \beta_k = 0$
- example: coordinate vectors $e_1, \ldots, e_k$ $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$
- "independence-dimension ineq.": any set of $n+1$ or more n-vectors is linearly dependent

Linear combination of linearly independent vectors: unique coeff's

- suppose x is a linear combination of linearly independent vectors $a_1, \ldots, a_k$

$$x = \beta_1 a_1 + \cdots + \beta_k a_k$$

then coeff's $\beta_1, \ldots, \beta_k$ are unique, i.e., if we also have

$$x = \gamma_1 a_1 + \cdots + \gamma_k a_k$$

then $\beta_i = \gamma_i$.

- proof:

$$(\beta_1 - \gamma_1) a_1 + \cdots + (\beta_k - \gamma_k) a_k = 0$$

$\Rightarrow$ by linear independence, $\beta_1 - \gamma_1 = \cdots = \beta_k - \gamma_k = 0$

- this means we can deduce the coeff's from x (will see on slide 21)

Basis

$a_1, \dots, a_n \in \mathbb{R}^n$

- a set of n linearly independent n- vectors $a_1 \dots, a_n$ is called a *basis*
- any n-vector b can be expressed as a linear combination of them:

$$b = \beta_1 a_1 + \dots + \beta_n a_n$$

  for some $\beta_1, \dots, \beta_n$
- and these coefficients are *unique*
- formula above is called expansion of b in the $a_1, \dots, a_n$ basis
- example: $e_1, \dots, e_n$ is a basis, $b = b_1 e_1 + \dots + b_n e_n$

Orthonormal vectors

means $a_i^T a_j = 0$

- set of n-vectors $a_1, \ldots, a_k$ are (mutually) *orthogonal* if $a_i \perp a_j$ for $i \neq j$
- they are normalized if $\|a_i\| = 1$, $i = 1, \ldots, k$
- express using inner products:

$$a_i^T a_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

$a_i^T a_i = \|a_i\|^2 = 1$

- when $k = n$, $a_1, \ldots, a_n$ are an *orthonormal basis*
- ex:

$$\begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}, \quad \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix},$$

Orthonormal expansion

- if $a_1, \ldots, a_n$ is an orthnormal basis, we have for any n-vector x,

$$x = (a_1^T x)a_1 + \ldots + (a_n^T x)a_n$$

- called *orthonormal expansion* of x (in the orthonormal basis)
- to verify, take inner product of both sides with a_i

later, we'll see an iterative algorithm to check if $a_1, \ldots, a_k$ are independent, called "Gram-Schmidt orthogonalization" algorithm