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

- 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:

	Veterans’ day	Memorial day	Academy A.	Golden G.	Super Bowl
Veterans’ day	0	0.095	0.130	0.153	0.170
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

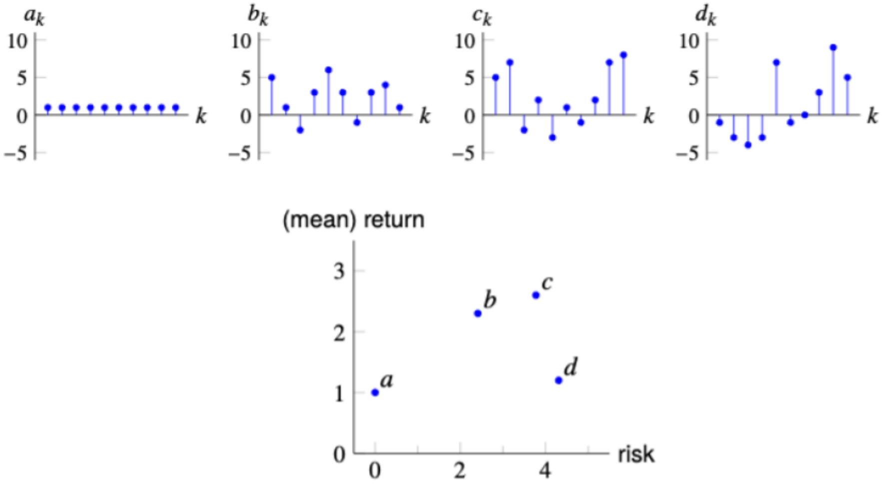
- for n-vector x, average of its entries is:
- de-meaned (or centered) vector:
- standard deviation of x is:

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

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



Cauchy-Schwartz inequality

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

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:

$$\begin{aligned} 0 &\leq \|\beta a - \alpha b\|^2 \\ &= \\ &= \\ &= \end{aligned}$$

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

Angle

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

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

- $\angle(a,b)$ is the number in $[0,\pi]$ that satisfies $\quad a^T b = \|a\|\|b\|\cos(\angle(a,b))$
- $\theta = \pi/2$: $a^T b = 0$ orthogonal; $\quad \theta = 0$: a, b aligned
- acute or obtuse angle:

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

Document dissimilarities by angles

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

	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

Correlation coefficient

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

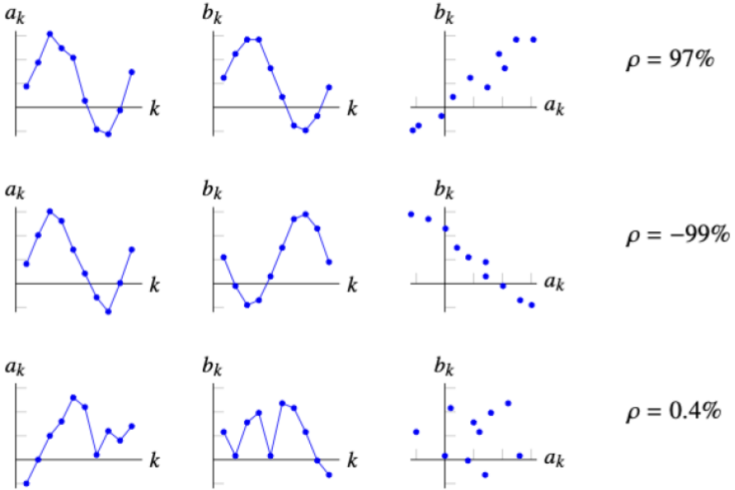
$$\tilde{a} = a - \mathbf{avg}(a)\mathbf{1}, \qquad \tilde{b} = b - \mathbf{avg}(b)\mathbf{1}$$

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

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

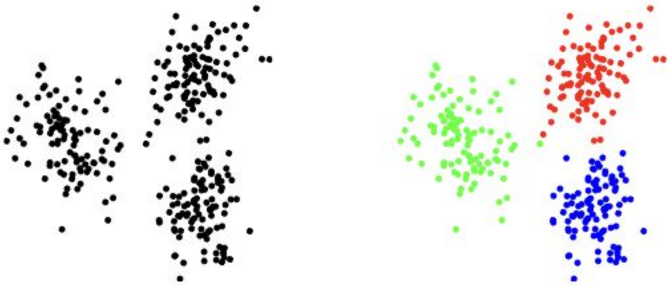
- $\rho = \cos(\angle \tilde{a}, \tilde{b})$
  - $\rho = 0$: a and b are
  - $\rho > 0.8$: a and b are
  - $\rho < -0.8$: a and b are

Examples



Clustering

- given N n-vectors $x_1, \ldots, x_N$, 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;...



Clustering objective

- $G_j \subset \{1, \ldots, N\}$ 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 corresponding representative
- **goal:** choose clatering c_i and representatives z_j to minimize J

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

- 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:

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

- alternating between these two steps gives the famous k-means 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 $e_1, \ldots, e_k$
- any set of $n + 1$ or more n-vectors is linearly dependent

Basis

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

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

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

Orthonormal vectors

- 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:

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

- 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