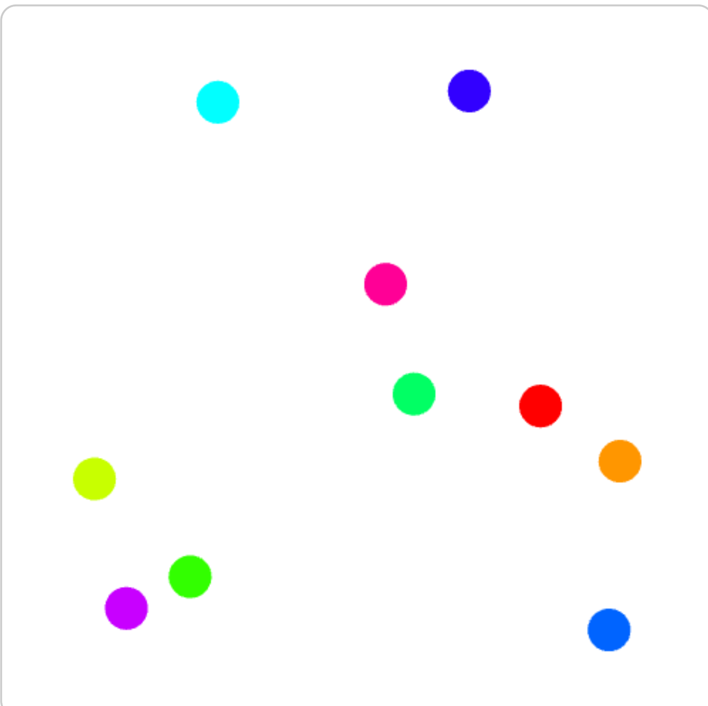
Unsupervised Learning

Supervised learning: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.
Unsupervised learning: $(x_1), (x_2), ..., (x_n)$.
➩ Clustering: separates items into groups.
➩ Novelty (outlier) detection: finds items that are different (two groups).
➩ Dimensionality reduction: represents each item by a lower dimensional feature vector while maintaining key characteristics.
Unsupervised learning applications:
➩ Google news.
➩ Google photo.
➩ Image segmentation.
➩ Text processing.
➩ Data visualization.
➩ Efficient storage.
➩ Noise removal.

Hierarchical Clustering

Hierarchical clustering iteratively merges groups: Link, Wikipedia.
➩ Start with each items as a cluster.
➩ Merge clusters that are closest to each other.
➩ Result in a binary tree with close clusters as children.
▼ TopHat Discussion
ID: Confirm

[1 points] Given the following dataset, use hierarchical clustering to divide the points into 3 groups. Drag one point to another point to merge them into one cluster. Click on a point to move it out of the cluster.



Distance between Points

Distance between points in m dimensional space is usually measured by Euclidean distance (also called L_2 distance): Wikipedia.

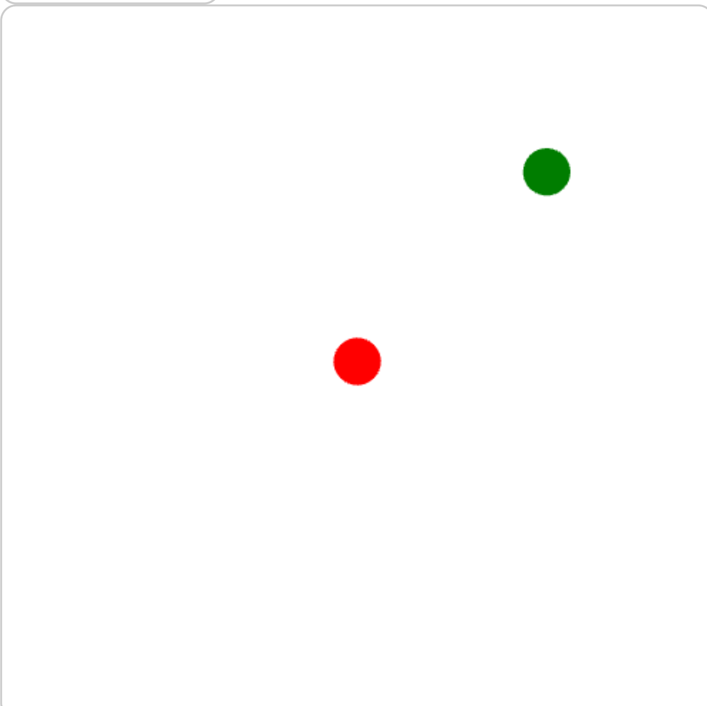
Distances can also be measured by L_1 or L_∞ distances.

➩ Manhattan distance (L_1): $\left\|x_i - x_j\right\|_1 = \left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + \ldots + \left|x_{im} - x_{jm}\right|$: Wikipedia.

➩ Chebyshev distance (L_∞): $\left\|x_i - x_j\right\|_\infty = \max\left\{\left|x_{i1} - x_{j1}\right|, \left|x_{i2} - x_{j2}\right|, \ldots, \left|x_{im} - x_{jm}\right|\right\}$: Wikipedia

▼ TopHat Discussion

[1 points] Move the green point so that it is within 100 pixels of the red point measured by the Manhattan distance. Highlight the region containing all points within 100 pixels of the red point.



Distance:

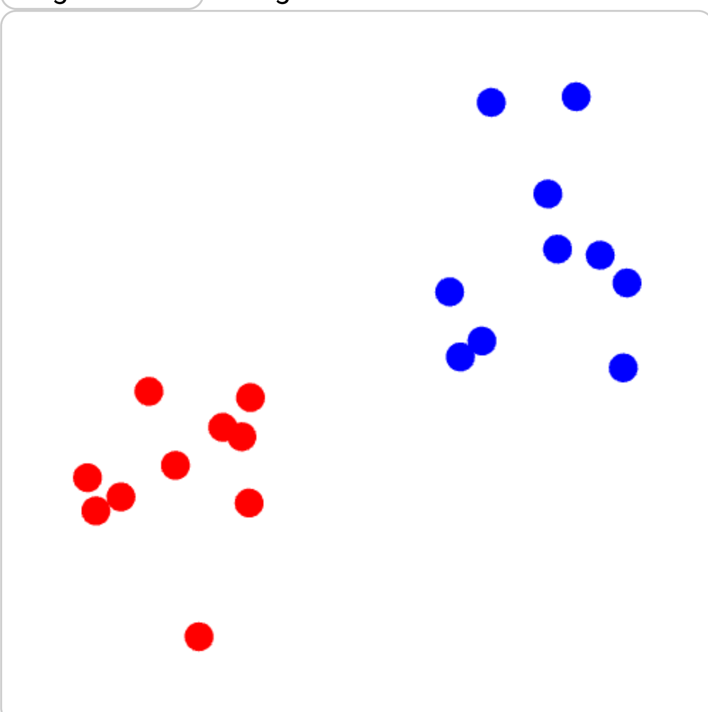
Distance between Clusters

📗 Distance between clusters (group of points) can be measured by single linkage distance, complete linkage distance, or average linkage distance.
➩ Single linkage distance: the shortest distance from any item in one cluster to any item in the other cluster: Wikipedia.
➩ Complete linkage distance: the longest distance from any item in one cluster to any item in the other cluster: Wikipedia.
➩ Average linkage distance: the average distance from any item in one cluster to any item in the other cluster (average of distances, not distance between averages): Wikipedia.

▼ TopHat Discussion

ID: Confirm

📗 [1 points] Highlight the Euclidean distance between the two clusters (red and blue) measured by the Single linkage distance.



Distance: 0

▼ TopHat Quiz

(Past Exam Question) ID: Confirm

📗 [4 points] You are given the distance table. Consider the next iteration of hierarchical agglomerative clustering (another name for the hierarchical clustering method we covered in the lectures) using complete linkage. What will the new values be in the resulting distance table corresponding to the 4 new clusters? If you merge two columns (rows), put the new distances in the column (row) with the smaller index. For

0	28	50	39	77
28	0	95	5	49
50	95	0	70	6
39	5	70	0	57
77	49	6	57	0

$$d = \begin{bmatrix} 0 & 28 & 50 & 39 & 77 \\ 28 & 0 & 95 & 5 & 49 \\ 50 & 95 & 0 & 70 & 6 \\ 39 & 5 & 70 & 0 & 57 \\ 77 & 49 & 6 & 57 & 0 \end{bmatrix}$$

Answer (matrix with multiple lines, each line is a comma separated vector):

Distance between cluster (items, points) 2 and 4 is 5 which is smallest. In next iteration, merge 2 and 4 into single custer.

Assume single linkage:

distance (cluster 1 and (combined 2 and 4)) = min distance (a point in C1, and a point in either C2 or C4)

= min {single linkage dist between C1 and C2, single linkage dist between C1 and C4}

= min{28, 39} = 28

distance (cluster 3 and (combined 2 and 4)) = 70, distance (cluster 5 and (combined 2 and 4)) = 49

	1	3	5	(2,4)
1	0	50	77	28
3		0	6	70
5			0	49
(2,4)				0

Next: Merge 3 & 5 =

	1	(3,5)	(2,4)
1	0	50	28
(3,5)		0	49
(2,4)			0

Number of Clusters

📗 The number of clusters should be chosen based on prior knowledge about the dataset.
📗 The algorithm can also stop merging as soon as all the between-cluster distances are larger than some fixed threshold.
📗 The binary tree generated by hierarachical clustering is often called dendrogram: Wikipedia.

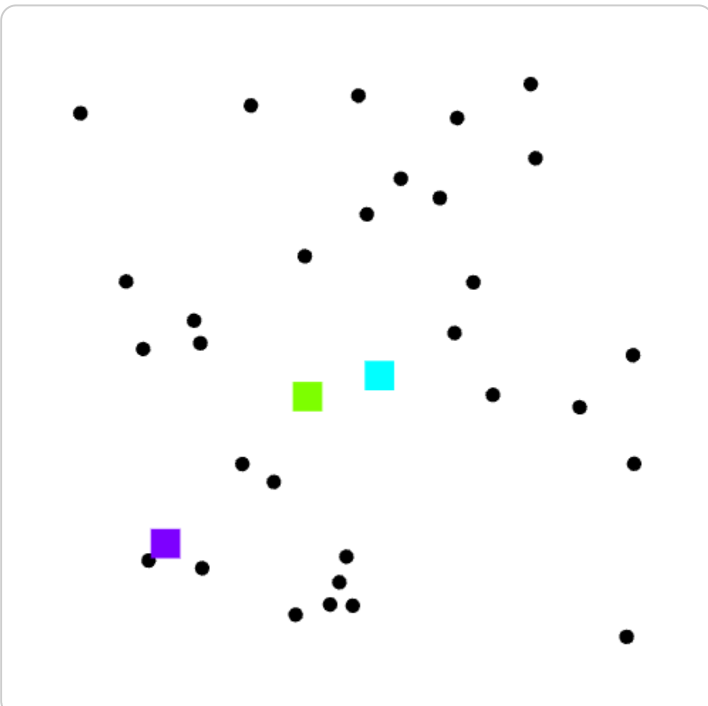
# of clusters

K Means Clustering

📗 K-means clustering (2-means, 3-means, ...) iteratively updates a fixed number of cluster centers: Link, Wikipedia.
➩ Start with K random cluster centers.
➩ Assign each item to its closest center.
➩ Update all cluster centers as the center of its items.
▼ TopHat Discussion

ID: Confirm

📗 [1 points] Given the following dataset, use k-means clustering to divide the points into 3 groups. Move the centers and click on the center to move it to the center of the points closest to the center.



Total distortion: 0

Total Distortion

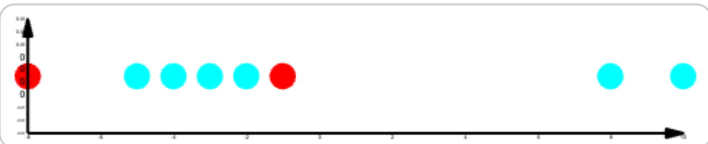
K means clustering tries to minimize the total distances of all items to their cluster centers. The total distance is called total distortion or inertia.

Suppose the cluster centers are $c_1, c_2, \ldots, c_K$, and the cluster center for an item x_i is $c\left(x_i\right)$ (one of $c_1, c_2, \ldots, c_K$), then the total distortion is $\left\|x_1 - c\left(x_1\right)\right\|_2^2 + \left\|x_2 - c\left(x_2\right)\right\|_2^2 + \ldots + \left\|x_n - c\left(x_n\right)\right\|_2^2$.

▼ TopHat Quiz

(Past Exam Question) ID: Confirm

[3 points] Perform k-means clustering on six points: x_1 = $[8]$, x_2 = $[-2]$, x_3 = $[10]$, x_4 = $[-4]$, x_5 = $[-3]$, x_6 = $[-5]$. Initially the cluster centers are at c_1 = $[-1]$, c_2 = $[-8]$. Run k-means for one iteration (assign the points, update center once and reassign the points once). Break ties in distances by putting the point in the cluster with the smaller index (i.e. favor cluster 1). What is the reduction in total distortion? Use Euclidean distance and calculate the total distortion by summing the squares of the individual distances to the center.



Note: the red points are the cluster centers and the other points are the training items.

Answer: dist(x4, c1) = 3 dist(x6, c1) = 4 . Calculate

dist(x4, c2) = 4 dist(x6, c2) = 3

x4 closer to c1

x6 closer to c2

c2: {x6} , c1: {x1, x2, x3, x4, x5}

update c1 ⇒ $\frac{1}{5}$ × (8 + 10 − 2 − 3 − 4) = 1.8

c2 ⇒ −5

c2: {x2, x4, x5, x6}, c1: {x1, x3}

update c1 ⇒ $\frac{1}{4}$(−2, −3, −4, −5) = −3.5

c2 ⇒ $\frac{1}{2}$ (8 + 10) = 9

2-means converges, stop.

Number of Clusters

📗 There are a few ways to choose the number of clusters K.
➩ K can be chosen based on prior knowledge about the items.
➩ K cannot be chosen by minimizing total distortion since the total distortion is always minimized at 0 when $K = n$. K can be chosen by minimizing total distortion plus some regularizer, for example, $\lambda m K \log\left(n\right)$ where λ is a fixed constant.

▼ TopHat Quiz

📗 [1 points] Upload an image and use K-means clustering to group the pixels into K clusters. Find an appropriate value of K: Choose files No file chosen . Click on the image to perform the clustering for 100 iterations.

Number of clusters:

Initial Clusters

There are a few ways to initialize the clusters: Link.
⇨ The initial cluster centers can be randomly chosen in the domain.
⇨ The initial cluster centers can be randomly chosen as K distinct items.
⇨ The first cluster center can be a random item, the second cluster center can be the item that is the farthest from the first item, the third cluster center can be the item that is the farthest from the first two items, ...

High Dimensional Data

📗 Text and image data are usually high dimensional: Link.
➩ The number of features of bag of words representation is the size of the vocabulary.
➩ The number of features of pixel intensity features is the number of pixels of the images.
📗 Dimensionality reduction is a form of unsupervised learning since it does not require labeled training data.

Principal Component Analysis

Principal component analysis rotates the axes ($x_1, x_2, \ldots, x_m$ axes) so that the first K new axes ($u_1, u_2, \ldots, u_K$) capture the directions of the greatest variability of the training data. The new axes are called principal components: Link, Wikipedia.

➩ Find the direction of the greatest variability, u_1.

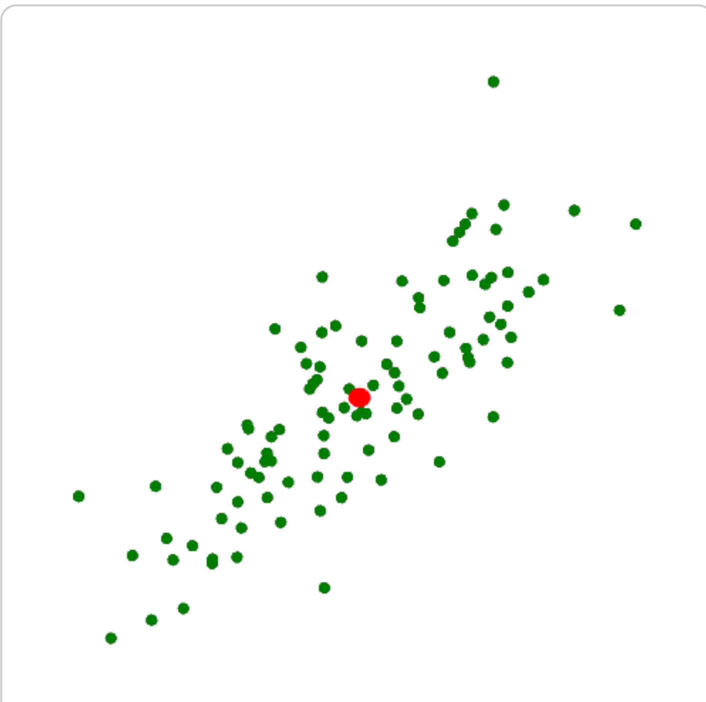
➩ Find the direction of the greatest variability that is orthogonal (perpendicular) to u_1, say u_2.

➩ Repeat until there are K such directions $u_1, u_2, \ldots, u_K$.

▼ TopHat Discussion

ID: Confirm

[1 points] Given the following dataset, find the direction in which the variation is the largest.



Projected variance: 0

Projected points:

Geometry

A vector u_k is a unit vector if it has length 1: $\|u_k\| = u_k^\top u_k = u_{k1}^2 + u_{k2}^2 + \ldots + u_{km}^2 = 1$.

Two vectors u_j, u_k are orthogonal (or uncorrelated) if $u_j^\top u_k = u_{j1} u_{k1} + u_{j2} u_{k2} + \ldots + u_{jm} u_{km} = 0$
.

The projection of x_i onto a unit vector u_k is $u_k^\top x_i u_k = (u_{k1} x_{i1} + u_{k2} x_{i2} + \ldots + u_{km} x_{im}) u_k$ (it is a number $u_k^\top x_i$ multiplied by a vector u_k). Since u_k is a unit vector, the length of the projection is $u_k^\top x_i$.

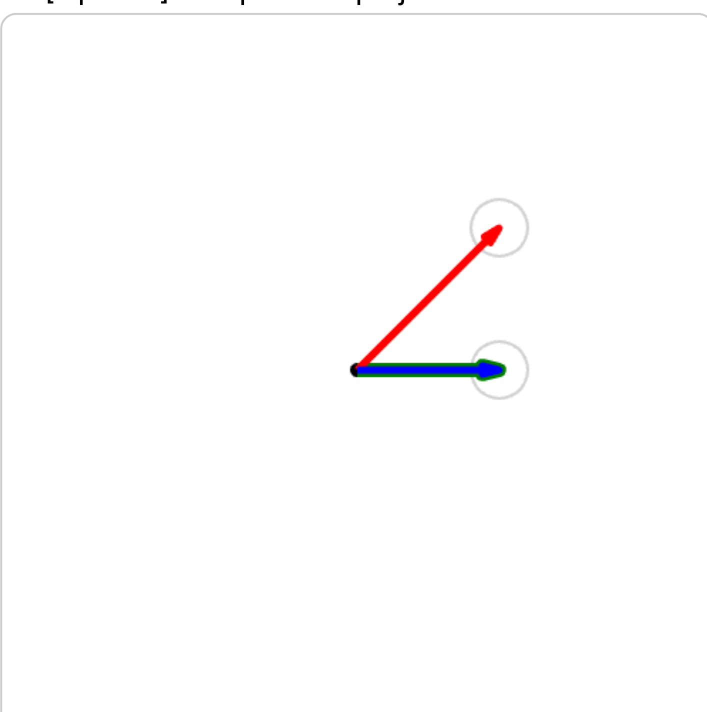
▼ Math Note

The dot product between two vectors $a = (a_1, a_2, \ldots, a_m)$ and $b = (b_1, b_2, \ldots, b_m)$ is usually written as $a \cdot b = a^\top b = \begin{bmatrix} a_1 & a_2 & \ldots & a_m \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_m \end{bmatrix} = a_1 b_1 + a_2 b_2 + \ldots + a_m b_m$. For the purpose of this course, the notation $a^\top b$ will be used instead of $a \cdot b$.

If x_i is projected onto some vector u_k that is not a unit vector, then the formula for projection is $\left(\dfrac{u_k^\top x_i}{u_k^\top u_k}\right) u_k$. Since for unit vector u_k, $u_k^\top u_k = 1$, the two formulas are equivalent.

▼ TopHat Discussion

[1 points] Compute the projection of the red vector onto the blue vector.



Red vector: , blue vector: .
Unit red vector: , unit blue vector: .
Projection: , length of projection: .

Statistics

📗 The (unbiased) estimate of the variance of $x_{1}, x_{2}, \ldots, x_{n}$ in one dimensional space ($m = 1$) is $\dfrac{1}{n - 1} \left(\left(x_{1} - \mu\right)^{2} + \left(x_{2} - \mu\right)^{2} + \ldots + \left(x_{n} - \mu\right)^{2}\right)$, where μ is the estimate of the mean (average) or $\mu = \dfrac{1}{n} \left(x_{1} + x_{2} + \ldots + x_{n}\right)$. The maximum likelihood estimate has $\dfrac{1}{n}$ instead of $\dfrac{1}{n - 1}$.

📗 In higher dimensional space, the estimate of the variance is $\dfrac{1}{n - 1} \left(\left(x_{1} - \mu\right) \left(x_{1} - \mu\right)^\top + \left(x_{2} - \mu\right) \left(x_{2} - \mu\right)^\top + \ldots + \left(x_{n} - \mu\right) \left(x_{n} - \mu\right)^\top\right)$. Note that μ is an m dimensional vector, and each of the $\left(x_{i} - \mu\right) \left(x_{i} - \mu\right)^\top$ is an m by m matrix, so the resulting variance estimate is a matrix called variance-covariance matrix.

📗 If $\mu = 0$, then the projected variance of $x_{1}, x_{2}, \ldots, x_{n}$ in the direction u_{k} can be computed by $u_{k}^\top \Sigma u_{k}$ where $\Sigma = \dfrac{1}{n - 1} X^\top X$, and X is the data matrix where row i is x_{i}.

➩ If $\mu \neq 0$, then X should be centered, that is, the mean of each column should be subtracted from each column.

▼ Math Note

📗 The projected variance formula can be derived by $u_{k}^\top \Sigma u_{k} = \dfrac{1}{n - 1} u_{k}^\top X^\top X u_{k} = \dfrac{1}{n - 1} \left(\left(u_{k}^\top x_{1}\right)^{2} + \left(u_{k}^\top x_{2}\right)^{2} + \ldots + \left(u_{k}^\top x_{n}\right)^{2}\right)$ which is the estimate of the variance of the projection of the data in the u_{k} direction.

Principal Component Analysis

📗 The goal is to find the direction that maximizes the projected variance: $\max_{u_k} u_k^\top \Sigma u_k$ subject to $u_k^\top u_k = 1$.
➩ This constrained maximization problem has solution (local maxima) u_k that satisfies $\Sigma u_k = \lambda u_k$, and by definition of eigenvalues, u_k is the eigenvector corresponding to the eigenvalue λ for the matrix Σ: Wikipedia.
➩ At a solution, $u_k^\top \Sigma u_k = u_k^\top \lambda u_k = \lambda u_k^\top u_k = \lambda$, which means, the larger the λ, the larger the variability in the direction of u_k.
➩ Therefore, if all eigenvalues of Σ are computed and sorted $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$, then the corresponding eigenvectors are the principal components: u_1 is the first principal component corresponding to the direction of the largest variability; u_2 is the second principal component corresponding to the direction of the second largest variability orthogonal to u_1, ...

▼ TopHat Quiz

(Past Exam Question) ID: Confirm

📗 [3 points] Given the variance matrix $\hat{\Sigma}$ = $\begin{bmatrix} 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 9 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 10 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 \end{bmatrix}$, what is the first principal component? Enter a unit vector.

📗 Answer (comma separated vector): . Calculate

Sigma = $PDP^{-1} = PTP^{T}$

P will contain eigenvectors, D is a diagonal matrix containing eigenvalues on diagonal

PC1 = eigenvector corresponding to largest eigenvalue (max $\{9, 5, 10\} = 10$) = $[\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0]$

PC2 = eigenvector corresponding to e.v. of 9 $[0, 0, 1]$.

Number of Dimensions

📗 There are a few ways to choose the number of principal components K.
➩ K can be selected based on prior knowledge or requirement (for example, $K = 2, 3$ for visualization tasks).
➩ K can be the number of non-zero eigenvalues.
➩ K can be the number of eigenvalues that are larger than some threshold.

Reduced Feature Space

📗 An original item is in the m dimensional feature space: $x_{i} = \left(x_{i 1}, x_{i 2}, \ldots, x_{i m}\right)$.
📗 The new item is in the K dimensional space with basis $u_{1}, u_{2}, \ldots, u_{k}$ has coordinates equal to the projected lengths of the original item: $\left(u_{1}^\top x_{i}, u_{2}^\top x_{i}, \ldots, u_{m}^\top x_{i}\right)$.
📗 Other supervised learning algorithms can be applied on the new features.

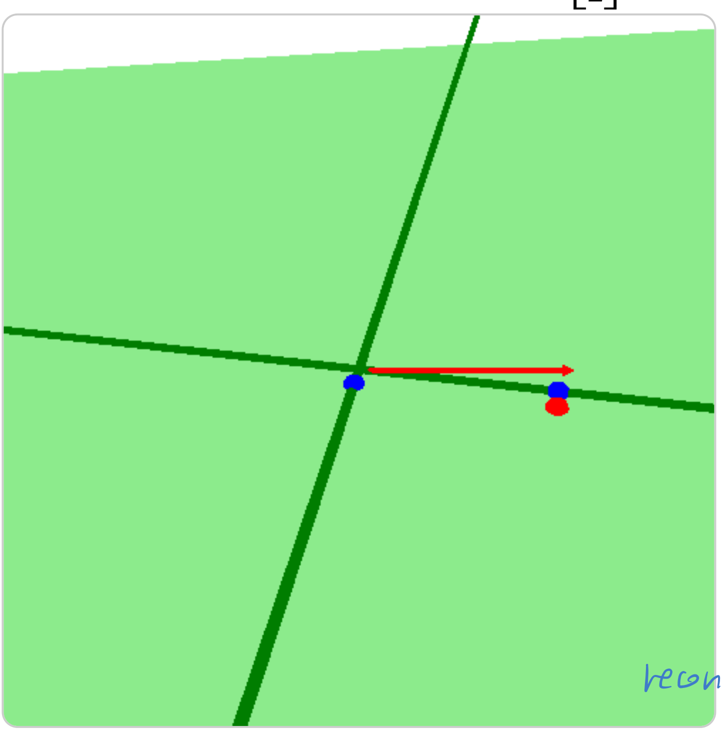
Reconstruction

📗 The original item can be reconstructed using the principal components. If all m principal components are used, then the original item can be perfectly reconstructed: $x_i = u_1^\top x_i u_1 + u_2^\top x_i u_2 + ... + u_m^\top x_i u_m$.
📗 The original item can be approximated by the first K principal components:
$x_i \approx u_1^\top x_i u_1 + u_2^\top x_i u_2 + ... + u_K^\top x_i u_K$.
➩ Eigenfaces are eigenvectors of face images: every face can be written as a linear combination of eigenfaces. The first K eigenfaces and their coefficients can be used to determine and reconstruct specific faces: Link, Wikipedia.

▼ TopHat Quiz

(Past Exam Question) ID: Confirm

📗 [2 points] You performed PCA (Principal Component Analysis) in $\mathbb{R}^3$. If the first principal component is v_1 = $\begin{bmatrix} \frac{3}{\sqrt{19}} \\ \frac{-1}{\sqrt{19}} \\ \frac{3}{\sqrt{19}} \end{bmatrix}$ and the second principal component is v_2 = $\begin{bmatrix} \frac{1}{\sqrt{86}} \\ \frac{9}{\sqrt{86}} \\ \frac{2}{\sqrt{86}} \end{bmatrix}$. What is the new 2D coordinates (new features created by PCA) for the point x = $\begin{bmatrix} 0 \\ 4 \\ 2 \end{bmatrix}$?



$pc1 = [\frac{3}{\sqrt{19}}, -\frac{1}{\sqrt{19}}, \frac{3}{\sqrt{19}}]$

$pc2 = [\frac{1}{\sqrt{86}}, \frac{9}{\sqrt{86}}, \frac{2}{\sqrt{86}}]$

$x = [0, 4, 2]$

$x' = [pc1 \cdot x, pc2 \cdot x]$

$= [\frac{-4+6}{\sqrt{19}}, \frac{36+4}{\sqrt{86}}]$

reconstruct x using $pc1$ & $pc2$ $= \left(\frac{-4+6}{\sqrt{19}}\right) \cdot pc1 + \frac{36+4}{\sqrt{86}} \cdot pc2$

$= [\frac{6}{19} + \frac{40}{86}, \frac{-2}{19} + \frac{360}{86}, \frac{6}{19} + \frac{80}{86}]$

📗 Answer (comma separated vector): . Calculate

Using only PC1: $x' = [\frac{-4+6}{\sqrt{19}} \cdot pc1]$

$= [\frac{6}{19}, \frac{-2}{19}, \frac{6}{19}]$


image_0.jpg
image_1.jpg
image_2.jpg
image_3.jpg
image_4.jpg
image_5.jpg
image_6.jp
image_16.jpg
image_17.jpg
image_18.jpg
image_19.jpg
image_20.jpg
image_21.jpg
image_22.jp
image_32.jpg
image_33.jpg
image_34.jpg
image_35.jpg
image_36.jpg
image_37.jpg
image_38.jp
image_48.jpg
image_49.jpg
image_50.jpg
image_51.jpg
image_52.jpg
image_53.jpg
image_54.jp
image_64.jpg
image_65.jpg
image_66.jpg
image_67.jpg
image_68.jpg
image_69.jpg
image_70.jp
image_80.jpg
image_81.jpg
image_82.jpg
image_83.jpg
image_84.jpg
image_85.jpg
image_86.jp
image_96.jpg
image_97.jpg
image_98.jpg
image_99.jpg
image_100.jpg
image_101.jpg
image_102.jp

📗 Please use Ctrl+F5 or Shift+F5 or Shift+Command+R or Incognito mode or Private Browsing to refresh the cached JavaScript.
📗 Anonymous feedback can be submitted to: Form.

Prev: L2, Next: L4

Last Updated: June 21, 2024 at 2:38 AM