CS5154/6054 Quiz 17 Key, 10/27/2022

(14.3) $$\vec{w}^T\vec{x} = b$$

DECISION HYPERPLANE

The assignment criterion then is: assign to c if $\vec{w}^T\vec{x} > b$ and to $\bar{c}$ if $\vec{w}^T\vec{x} \le b$. We call a hyperplane that we use as a linear classifier a *decision hyperplane*.

this for Rocchio, observe that a vector $\vec{x}$ is on the decision boundary if it has equal distance to the two class centroids:

(14.4) $$|\vec{\mu}(c_1) - \vec{x}| = |\vec{\mu}(c_2) - \vec{x}|$$

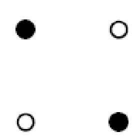
Exercise 14.15

Show that Equation (14.4) defines a hyperplane with $\vec{w} = \vec{\mu}(c_1) - \vec{\mu}(c_2)$ and $b = 0.5 * (|\vec{\mu}(c_1)|^2 - |\vec{\mu}(c_2)|^2)$.

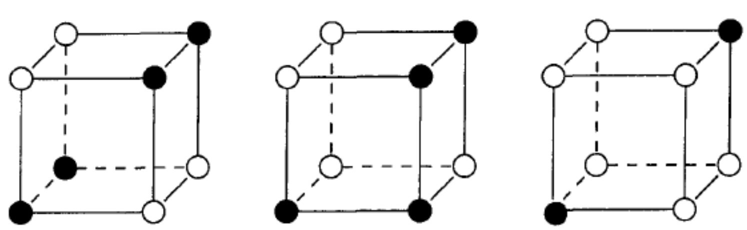
(14.4) is true iff $(\mu_1 - x)^2 = (\mu_2 - x)^2$ or $|\mu_1|^2 - 2\mu_1^T x + |x|^2 = |\mu_2|^2 - 2\mu_2^T x + |x|^2$ or $|\mu_1|^2 - 2\mu_1^T x = |\mu_2|^2 - 2\mu_2^T x$ or $|\mu_1|^2 - |\mu_2|^2 = 2\mu_1^T x - 2\mu_2^T x$ or $2b = 2w^T x$ or $b = w^T x$ (14.3).

Exercise 14.17

Assuming two classes, show that the percentage of non-separable assignments of the vertices of a hypercube decreases with dimensionality M for $M > 1$. For example, for $M = 1$ the proportion of non-separable assignments is 0, for $M = 2$, it is 2/16. One of the two non-separable cases for $M = 2$ is shown in Figure 14.15, the other is its mirror image. Solve the exercise either analytically or by simulation.



► **Figure 14.15** A simple non-separable set of points.



Three non-separable assignments for M = 3. There are 8 corners and $2^8 = 256$ assignments. How many of them are non-separable? Is the percentage of non-separable assignments decreasing from M = 2 to M = 3?

Let us consider the cases when there are 0, 1, 2, 3, and 4 black corners. Cases for 5 to 8 are symmetric. There are 8 chooses k assignments for k blacks and thus 1, 8, 28, 56, and 70 assignments for k = 0 to 4. The cases for k = 0 or 1 are separable. For k = 2, the two black corners must be adjacent to be separable and there are 12 edges or 12 separable assignments for k=2 and 28-12 = 16 non-separable k=2 assignments, including the illustrated case on the right. The illustrations on left and middle are for k=4 with the remaining patterns separable in two ways: all four blacks on one of the six surfaces, or three are adjacent to the other one, which may be any of the eight corners. So for the 70 assignments of k=4, 56 = 70 – 6 – 8 are mirror images of either the left or the middle examples. There are also 16 non-separable for k = 6 and the total non-separables for k = 2, 4, 6 is 16 + 56 + 16 = 88 and this makes 88/256 assignments which is 11/32 > 4/32 = 2/16. Whether the k=3 and k=5 cases are non-separable, this percentage is already larger than the case for M = 2. So the percentage of non-separable assignments does not decrease from M=2 to M=3.

CS5154/6054 Quiz 18 Key, 11/1/2022

CENTROID

K-means is the most important flat clustering algorithm. Its objective is to minimize the average squared Euclidean distance (Chapter 6, page 131) of documents from their cluster centers where a cluster center is defined as the mean or *centroid* $\vec{\mu}$ of the documents in a cluster ω:

$$\vec{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x}$$

RESIDUAL SUM OF SQUARES

A measure of how well the centroids represent the members of their clusters is the *residual sum of squares* or *RSS*, the squared distance of each vector from its centroid summed over all vectors:

$$\text{RSS}_k = \sum_{\vec{x} \in \omega_k} |\vec{x} - \vec{\mu}(\omega_k)|^2$$

$$\text{RSS} = \sum_{k=1}^{K} \text{RSS}_k \tag{16.7}$$

Exercise 16.20 [⋆⋆⋆]

WITHIN-POINT SCATTER

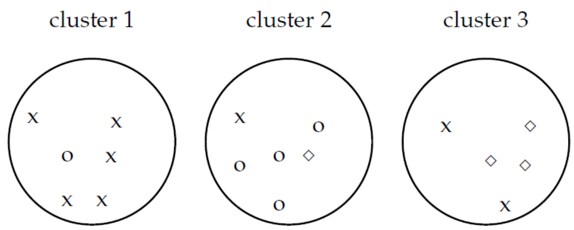
The *within-point scatter* of a clustering is defined as $\sum_k \frac{1}{2} \sum_{\vec{x}_i \in \omega_k} \sum_{\vec{x}_j \in \omega_k} |\vec{x}_i - \vec{x}_j|^2$. Show that minimizing RSS and minimizing within-point scatter are equivalent.

$\sum_k ½ \sum_{xi \text{ in } \omega k} \sum_{xj \text{ in } \omega k} (x_i - x_j)^2 = \sum_k ½ \sum_{xi \text{ in } \omega k} \sum_{xj \text{ in } \omega k} ((x_i - \mu_k) - (x_j - \mu_k))^2 = \sum_k ½ \sum_{xi \text{ in } \omega k} \sum_{xj \text{ in } \omega k} ((x_i - \mu_k)^2 - 2(x_i - \mu_k)(x_j - \mu_k) + (x_j - \mu_k)^2) = \sum_k ½ \sum_{xi \text{ in } \omega k} \sum_{xj \text{ in } \omega k} (x_i - \mu_k)^2 - \sum_k \sum_{xi \text{ in } \omega k} \sum_{xj \text{ in } \omega k} (x_i - \mu_k)(x_j - \mu_k) + \sum_k ½ \sum_{xi \text{ in } \omega k} \sum_{xj \text{ in } \omega k} (x_j - \mu_k)^2 = \sum_k ½ \sum_{xi \text{ in } \omega k} (x_i - \mu_k)^2 \sum_{xj \text{ in } \omega k} 1 - \sum_k \sum_{xi \text{ in } \omega k} (x_i - \mu_k) \sum_{xj \text{ in } \omega k} (x_j - \mu_k) + \sum_k ½ \sum_{xj \text{ in } \omega k} (x_j - \mu_k)^2 \sum_{xi \text{ in } \omega k} 1.$

Notice that $\sum_{xi \text{ in } \omega k} (x_i - \mu_k)^2 = RSS_k$, $\mu_k = \sum_{xi \text{ in } \omega k} x_i / |\omega_k|$, and $\sum_{xi \text{ in } \omega k} (x_i - \mu_k) = \sum_{xi \text{ in } \omega k} x_i - \sum_{xi \text{ in } \omega k} \mu_k = |\omega_k|\mu_k - \mu_k \sum_{xi \text{ in } \omega k} 1 = |\omega_k|\mu_k - \mu_k|\omega_k| = 0$. Thus, the within-point scatter $\sum_k ½ \sum_{xi \text{ in } \omega k} \sum_{xj \text{ in } \omega k} (x_i - x_j)^2 = \sum_k ½ (|\omega_k| RSS_k + |\omega_k| RSS_k) = \sum_k |\omega_k| RSS_k$.

CS5154/6054 Quiz 19 Key, 11/3/2022

Complete the computation of FN and TN for RI.



► **Figure 16.4** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and $\diamond$, 3 (cluster 3). Purity is $(1/17) \times (5+4+3) \approx 0.71$.

An alternative to this information-theoretic interpretation of clustering is to view it as a series of decisions, one for each of the $N(N-1)/2$ pairs of documents in the collection. We want to assign two documents to the same cluster if and only if they are similar. A true positive (TP) decision assigns two similar documents to the same cluster, a true negative (TN) decision assigns two dissimilar documents to different clusters. There are two types of errors we can commit. A false positive (FP) decision assigns two dissimilar documents to the same cluster. A false negative (FN) decision assigns two similar documents to different clusters. The *Rand index* (RI) measures the percentage of decisions that are correct. That is, it is simply accuracy (Section 8.3, page 155).

$$\mathrm{RI} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}}$$

As an example, we compute RI for Figure 16.4. We first compute TP + FP. The three clusters contain 6, 6, and 5 points, respectively, so the total number of "positives" or pairs of documents that are in the same cluster is:

$$\mathrm{TP}+\mathrm{FP} = \binom{6}{2} + \binom{6}{2} + \binom{5}{2} = 40$$

Of these, the x pairs in cluster 1, the o pairs in cluster 2, the $\diamond$ pairs in cluster 3, and the x pair in cluster 3 are true positives:

$$\mathrm{TP} = \binom{5}{2} + \binom{4}{2} + \binom{3}{2} + \binom{2}{2} = 20$$

Thus, $\mathrm{FP} = 40 - 20 = 20$.

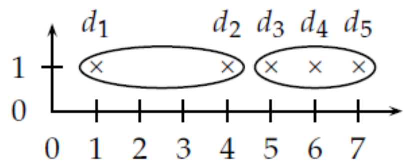
FN and TN are computed similarly, resulting in the following contingency table:

	Same cluster	Different clusters
Same class	TP = 20	FN = 24
Different classes	FP = 20	TN = 72

RI is then $(20+72)/(20+20+24+72) \approx 0.68$.

TP + FN = (8, 2) + (5, 2) + (4, 2) because there are 8 crosses, 5 circles, and 4 diamonds. This is TP + FN = 28 + 10 + 6 = 44. Because TP = 20, this gives FN = 44 – 20 = 24. The total number of pairs is (17, 2) = 136. TN = 136 – TP – FP – FN = 136 – 20 – 20 – 24 = 72.

CS5154/6054 Quiz 20 Key, 11/10/2022



► **Figure 17.7** Outliers in complete-link clustering. The five documents have the x-coordinates $1 + 2\epsilon, 4, 5 + 2\epsilon, 6$ and $7 - \epsilon$. Complete-link clustering creates the two clusters shown as ellipses. The most intuitive two-cluster clustering is $\{\{d_1\}, \{d_2, d_3, d_4, d_5\}\}$, but in complete-link clustering, the outlier d_1 splits $\{d_2, d_3, d_4, d_5\}$ as shown.

distance	d1	d2	d3	d4	d5
d1	0	3 - 2ε	4	5 - 2ε	6 - 3ε
d2	3 - 2ε	0	1 + 2ε	2	3 - ε
d3	4	1 + 2ε	0	1 - 2ε	2 - 3ε
d4	5 - 2ε	2	1 - 2ε	0	1 - ε
d5	6 - 3ε	3 - ε	2 - 3ε	1 - ε	0

Perform single-link and complete-link clustering on the five documents in Fig. 17.7. In each case, there are four mergers of subsets of documents, starting with singleton subsets {d1}, {d2}, {d3}, {d4}, {d5}. Indicate which two subsets will be merged at each merger in each case. After each merger, you may want to update the distance matrix (with one fewer rows/columns).

	Single-link clustering	Complete-link clustering
merger 1	{d3}, {d4} at 1 - 2ε	{d3}, {d4} at 1 - 2ε
merger 2	{d3, d4} and {d5} at 1 - ε	{d3, d4} and {d5} at 2 - 3ε
merger 3	{d2} and {d3, d4, d5} at 1 + 2ε	{d1} and {d2} at 3 - 2ε
merger 4	{d1} and {d2, d3, d4, d5} at 1 + 2ε	{d1, d2} and {d3, d4, d5} at 6 - 3ε

Draw the resulting hierarchies as binary trees or dendrograms for single-link and complete-link clustering.

```
d1 -------------------------------------------- (3-2e) -|
d2----------------------------(1+2e) --|--------------|
d3 – (1-2e) -|-----(1-e)-|------------|
d4 ------------|              |
d5 --------------------------|
```

```
d1 ----------------------------(3-2e)---|
d2 --------------------------------------- |------(6-3e) --|
d3 – (1-2e) -|---(2-3e)-|-------------------------------|
d4 ------------|              |
d5 --------------------------|
```

CS5154/6054 Quiz 21 Key, 11/15/2022

$$C = U\Sigma V^T, \quad CC^T = U\Sigma V^T\, V\Sigma U^T = U\Sigma^2 U^T. \quad \begin{matrix} 2.16 & 0.00 \\ 0.00 & 1.59 \end{matrix}$$

	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
voyage	1	0	0	1	1	0
trip	0	0	0	1	0	1

	1	2
ship	−0.44	−0.30
boat	−0.13	−0.33
ocean	−0.48	−0.51
voyage	−0.70	0.35
trip	−0.26	0.65

ship	-0.95	-0.48
boat	-0.28	-0.52
ocean	-1.04	-0.81
voyage	-1.51	0.56
trip	-0.56	1.03

The above is the term document matrix C, the U_2 matrix which is the reduced U matrix from the SVD USV^T, and $U_2\Sigma_2$. Consider each row in the right table as a 2D vector representation of a word. Compute the dot products of the vector representations of the words. (Only fill out the unshaded cells.)

dot product	ship	boat	ocean	voyage	trip
ship		0.52	1.38	1.16	0.038
boat			0.71	0.13	-0.38
ocean				1.12	-0.25
voyage					1.42
trip					

Perform a single-link clustering of the five words, using the dot product as the similarity between them.

	subtree 1	subtree 2	the new subtree	combination similarity
Merger 1	{voyage}	{trip}	{v, t}	1.42
Merger 2	{ship}	{ocean}	{s, o}	1.38
Merger 3	{s, o}	{v, t}	{s, o, v, t}	1.16
Merger 4	{b}	{s, o, v, t}	{s, b, o, v, t}	0.71

Draw a dendrogram for this clustering.

```
voyage -----|
trip  ---------|----------|
                          |----------|
ship -------------|------|           |
ocean -----------|                   |
                                     |
boat --------------------------------|
```

CS5154/6054 Quiz 22 Key, 11/17/2022

Exercise 18.11

Assume you have a set of documents each of which is in either English or in Spanish. The collection is given in Figure 18.4.

1. Construct the appropriate term-document matrix C to use for a collection consisting of these documents. For simplicity, use raw term frequencies rather than normalized tf-idf weights. Make sure to clearly label the dimensions of your matrix.

2. Write down the matrices U_2, Σ_2' and V_2 and from these derive the rank 2 approximation C_2.

DocID	Document text
1	hello
2	open house
3	mi casa
4	hola Profesor
5	hola y bienvenido
6	hello and welcome

► **Figure 18.4** Documents for Exercise 18.11.

C	1	2	3	4	5	6
Profesor	0	0	0	1	0	0
and	0	0	0	0	0	1
bienvenido	0	0	0	0	1	0
casa	0	0	1	0	0	0
hello	1	0	0	0	0	1
hola	0	0	0	1	1	0
house	0	1	0	0	0	0
mi	0	0	1	0	0	0
open	0	1	0	0	0	0
welcome	0	0	0	0	0	1
y	0	0	0	0	1	0

Below are U and V. The six singular values are 1.90, 1.85, 1.41, 1.41, 1.18, and 0.77.

	dim1	dim2	dim3	dim4	dim5	dim6
Profesor	0.28	0	0	0	-0.72	0
and	0	-0.5	0	0	0	0.5
bienvenido	0.45	0	0	0	0.45	0
casa	0	0	-0.7	0	0	0
hello	0	-0.7	0	0	0	-0.7
hola	0.72	0	0	0	-0.28	0
house	0	0	0	-0.7	0	0
mi	0	0	-0.7	0	0	0
open	0	0	0	-0.7	0	0
welcome	0	-0.5	0	0	0	0.5
y	0.45	0	0	0	0.45	0

	dim1	dim2	dim3	dim4	dim5	dim6
d1	0	-0.38	0	0	0	-0.92
d2	0	0	0	-1	0	0
d3	0	0	-1	0	0	0
d4	0.53	0	0	0	-0.85	0
d5	0.85	0	0	0	0.53	0
d6	0	-0.92	0	0	0	0.38

The top two dimensions (dim1 with singular value 1.90 and dim2 with singular value 1.85) can be considered as latent topics. Collect terms and documents with non-zero U and V at these dimensions and speculate the "topics" they present.

"topic"	terms and documents involved	meaning of topic
dim1	Profesor, bienvenido, hola, y, d4, d5	Spanish (greeting)
dim2	and, hello, welcome, d1, d6	English (greeting)

K-means is the most important flat clustering algorithm. Its objective is to minimize the average squared Euclidean distance (Chapter 6, page 131) of documents from their cluster centers where a cluster center is defined as the mean or *centroid* $\vec{\mu}$ of the documents in a cluster ω:

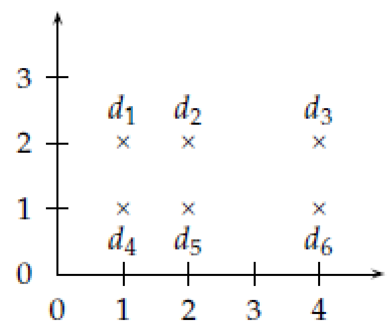
$$\vec{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x}$$

A measure of how well the centroids represent the members of their clusters is the *residual sum of squares* or *RSS*, the squared distance of each vector from its centroid summed over all vectors:

$$\text{RSS}_k = \sum_{\vec{x} \in \omega_k} |\vec{x} - \vec{\mu}(\omega_k)|^2$$

$$\text{RSS} = \sum_{k=1}^{K} \text{RSS}_k$$

RSS is the objective function in K-means and our goal is to minimize it. Since



► **Figure 16.7** The outcome of clustering in K-means depends on the initial seeds. For seeds d_2 and d_5, K-means converges to $\{\{d_1, d_2, d_3\}, \{d_4, d_5, d_6\}\}$, a suboptimal clustering. For seeds d_2 and d_3, it converges to $\{\{d_1, d_2, d_4, d_5\}, \{d_3, d_6\}\}$, the global optimum for $K = 2$.

Exercise 16.10

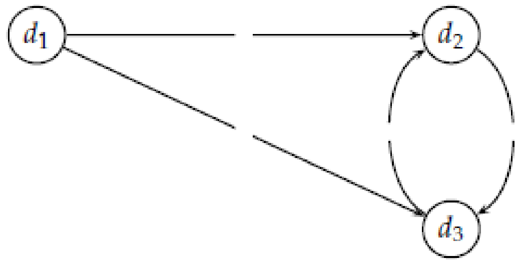
Compute RSS for the two clusterings in Figure 16.7.

First compute the centroids for each clustering.

Clustering 1: $\mu_1 = (7/3, 2)$, $\text{RSS}_1 = (4/3)^2 + (1/3)^2 + (5/3)^2 = 14/3$. Same is RSS_2. RSS is $28/3 = 9.333$.

Clustering 2: $\mu_1 = (1.5, 1.5)$, $\text{RSS}_1 = 4\,(0.5^2 + 0.5^2) = 2$. $\text{RSS}_2 = 2(0.5^2) = 0.5$. RSS is 2.5.

CS5154/6054 Quiz 23 Key, 11/22/2022



$$
\begin{aligned}
\vec{h} &\leftarrow A\vec{a} \\
\vec{a} &\leftarrow A^T\vec{h},
\end{aligned}
$$

▶ **Figure 21.7** Web graph for Exercise 21.22.

1. Write down the adjacency matrix **A** for the web graph in Figure 21.7. The entry A_{ij} is 1 if there is a hyperlink from d_i to d_j, and 0 otherwise. Initialize the vector **a** as $(1, 1, 1)^T$, repeatedly compute **h** = **Aa** and **a** = $\mathbf{A^T h}$, and normalize the resulting **h** and **a** so that the maximum element is 1, until there are no changes. Show your matrix multiplications, normalizations, and termination of the loop.

   A is [[0 1 1][0 0 1][0 1 0]] and A^T is [[0 0 0][1 0 1][1 1 0]]. h = Aa = $(2\ 1\ 1)^T$, normalized to h = $(1\ 0.5\ 0.5)^T$. a = A^Th = $(0\ 0.5\ 0.5)^T$, normalized to a = $(0, 1, 1)^T$. Second iteration: h = Aa = $(2\ 1\ 1)^T$, normalized to h = $(1\ 0.5\ 0.5)^T$, and is the same. Loop terminates.

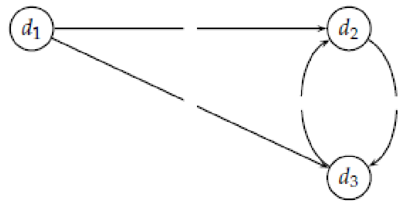
2. What is the matrix $\mathbf{AA^T}$? Show that the final **h** vector is an eigenvector of this matrix and find the corresponding eigenvalue.

   AA^T is [[2 1 1][1 1 0][1 0 1]] and AA^Th is [[2 1 1][1 1 0][1 0 1]]$(1\ 0.5\ 0.5)^T = (3\ 1.5\ 1.5)^T$ = 3h. The corresponding eigenvalue $\lambda_h = 3$.

3. What is the matrix $\mathbf{A^T A}$? Show that the final **a** vector is an eigenvector of this matrix and find the corresponding eigenvalue.

   $A^T A$ is [[0 0 0][0 2 1][0 1 2]] and $A^T A$a is [[0 0 0][0 2 1][0 1 2]]$(0\ 1\ 1)^T = (0\ 3\ 3)^T$ = 3a. The corresponding eigenvalue $\lambda_a = 3$.

CS5154/6054 Quiz 24 Key, 11/29/2022



► **Figure 21.7** Web graph for Exercise 21.22.

web graph is defined as follows: if there is a hyperlink from page i to page j, then $A_{ij} = 1$, otherwise $A_{ij} = 0$. We can readily derive the transition probability matrix P for our Markov chain from the $N \times N$ matrix A:

1. If a row of A has no 1's, then replace each element by 1/N. For all other rows proceed as follows.
2. Divide each 1 in A by the number of 1's in its row. Thus, if there is a row with three 1's, then each of them is replaced by 1/3.
3. Multiply the resulting matrix by $1 - \alpha$.
4. Add α/N to every entry of the resulting matrix, to obtain P.

First let us find P for α=0.5 and your P should only contain values 1/6, 2/3, and 5/12, like (21.3). Show the original 3x3 matrix A, and the transformation after Steps 2, 3, and 4.

A = [[0 1 1][0 0 1][0 1 0]]. After step 2 it is [[0 ½ ½][0 0 1][0 1 0]]. After step 3 it is [[0 ¼ ¼][0 0 ½][0 ½ 0]]. After step 4 it is [[1/6 5/12 5/12][1/6 1/6 2/3][1/6 2/3 1/6]].

Starting with initial probability vector x_0 = [1 0 0], compute $x_1 = x_0P$ and then $x_2 = x_1P$ and you may have already found the steady-state probability π such that πP = π.

$x_1 = x_0P$ = [1/6 5/12 5/12], $x_2 = x_1P$ = [(1/6)(1/6)+(5/12)(1/6)+(5/12)(1/6) (1/6)(5/12)+(5/12)(1/6)+(5/12)(2/3) (1/6)(5/12)+(5/12)(2/3)+(5/12)(1/6)] = [1/6 5/12 5/12] = π

Let us now treat α as a variable. Starting with A and find P for a general teleport probability α.

A = [[0 1 1][0 0 1][0 1 0]]. After step 2 it is [[0 ½ ½][0 0 1][0 1 0]]. After step 3 it is [[0 (1-α)/2 (1-α)/2][0 0 1-α][0 1-α 0]]. After step 4 it is [[α/3 (3-α)/6 (3-α)/6][α/3 α/3 (3-2α)/3][α/3 (3-2α)/3 α/3]].

Again, starting with initial probability vector x_0 = [1 0 0], compute $x_1 = x_0P$ and then $x_2 = x_1P$. You may also have already found the steady state probability π such that πP = π.

$x_1 = x_0P$ = [α/3 (3-α)/6 (3-α)/6], $x_2 = x_1P$ = [(α/3)(α/3)+((3-α)/6)(α/3)+((3-α)/6)(α/3) (α/3)((3-α)/6)+((3-α)/6)(α/3)+((3-α)/6)((3-2α)/3) (α/3)((3-α)/6)+((3-α)/6)((3-2α)/3)+((3-α)/6)(α/3)] = [α/3 (3-α)/6 (3-α)/6] = π.