Internet Analytics (COM-308): Final Exam
Solutions

August 19, 2020

Duration: **2h45**.

Total points: **100**.

Number of pages: **19**.

Allowed documents: **class notes, lab handouts, homeworks, your own code**.

There should in general be enough room below every question for intermediate calculations and your answer. However, you are allowed to use additional sheets of paper; please **write your name on every sheet**, number them, and staple them to this document before handing in.

The use of **mobile phones, tablets, laptop computers**, and other communication devices is **prohibited**.

Last name:
First name:
SCIPER number:
Signature:

Please leave blank.

1	2	3	4	5	6	7	Total
24	15	10	10	11	10	20	100

Question 1: Multiple Choice Questions (24 points)

(24 pts) All questions have a single answer. Check the correct one. Grading:

- Correct answer: +2 points;
- Wrong answer: −1 point;
- No answer or "I don't know": 0 point.

1. Consider the following snippet of PySpark code.

```
data = sc.parallelize([
    (1, 2), (1, 4), (2, 3), (2, 4), (2, 8),
    (3, 4), (3, 5), (5, 6), (5, 7), (7, 8)
])
output = (data.map(lambda x: (x[0], 1))
              .reduceByKey(lambda x, y: x+y)
              .map(lambda x: (x[1], x[0]))
              .sortByKey().first())
```

What are the contents of the variable `output` at the end of the execution?

☐ `(3, 2)`
■ `(1, 7)`
☐ `(2, 8)`
☐ I don't know

2. Consider the Bayesian network of Figure 1.
How many parameters does this model have?

☐ 27
■ 61
☐ 93
☐ I don't know

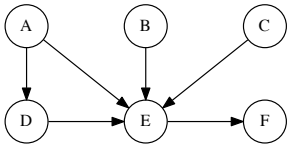
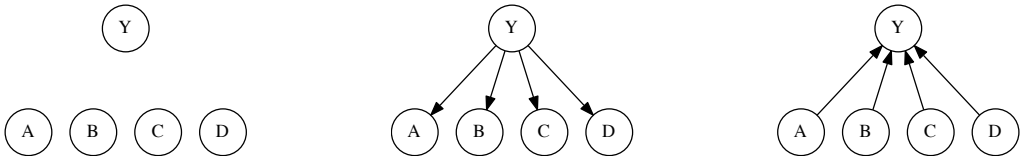


Figure 1: Bayesian network with $A, B, C \in \{0, 1\}$, and $D, E, F \in \{0, 1, 2\}$.

3. Consider the Bayesian network of Figure 1.
Which of the following statements is correct?

☐ $D \perp E|A$
☐ $D \perp C, F|A$
■ $D \perp F|A, E$
☐ I don't know



Network 1 Network 2 Network 3

Figure 2

4. Which one of the three Bayesian networks shown in Figure 2 above correspond to the naive Bayes classifier with features A, B, C, D and class Y.

☐ Network 1
■ Network 2
☐ Network 3
☐ I don't know

5. You have trained a latent factor based recommender system, and find that the RMSE while performing cross-validation is much higher than the RMSE on the training set. What do you suggest as a possible next step?

   ☐ Decrease the strength of regularization
   ☐ Increase the number of latent factors
   ■ Increase the strength of regularization
   ☐ I don't know

6. Consider a (connected) graph G, which consists of two disjoint complete graphs $K_{n/2}$ of equal size, with one additional edge connecting the two. For which n is the modularity highest?

   ☐ $n = 6$
   ☐ $n = 8$
   ■ $n = 10$
   ☐ I don't know

7. We have three independent random variables X, Y, and Z. X is exponential with mean 10. Y is Gaussian with mean 0 and variance 10. Z is Pareto with index $\gamma = 1.5$. What can you conclude about the variance of $X + Y + Z$?

   ☐ It is at most 20
   ☐ It is at least 20 and at most 40
   ■ It is at least 40
   ☐ I don't know

8. Consider a real symmetric $n \times n$ matrix X and its singular value decomposition. Which of the following statements are always true?

   ☐ All singular values of X are strictly positive
   ■ Singular values of X are the absolute values of the eigenvalues of X
   ☐ Singular values of X are the square roots of the eigenvalues of X
   ☐ I don't know

9. We evaluate the similarity of two (non-empty) documents A and B. Which of the following statements is **false**?

   ☐ It is possible that A and B have perfect similarity according to the vector space model, but Jaccard similarity equal to zero over 2-grams
   ☐ The probability $P(s(A) = s(B))$ is an unbiased estimator of the Jaccard similarity of A and B (assuming no hash collision).
   ■ 3-gram similarity can be higher than 2-gram similarity.
   ☐ I don't know

10. Consider a graph G and two random processes X and Y. The values of X are the nodes visited by a random walk on G *after* convergence to the stationary distribution, X_t is the node visited at time t. The values of Y are nodes chosen from G, where the probability of choosing a node n at time t is proportional to the degree of n. You are given G and a sample $S = [S_1, \ldots, S_L]$ which was generated from one of the two processes ($S_i = X_i \ \forall i$ or $S_i = Y_i \ \forall i$). For some large enough L, can you say which process generated it?

- ☐ We cannot say, since the number of times each node of G occurs in S would be nearly the same whether X or Y generated it.
- ■ We can say, even though number of times each node of G occurs in S would be nearly the same whether X or Y generated it.
- ☐ We can say, only because the number of times each node of G occurs in S would be very different depending on whether X or Y generated it.
- ☐ I don't know

11. Consider the graph shown in Figure 3. We denote the PageRank of node i by π_i. Which of the following ordering of the PageRanks is correct when $\theta = 0.999$?

- ☐ $\pi_{10} > \pi_5 > \pi_9 > \pi_8 > \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_6 = \pi_7$
- ☐ $\pi_5 > \pi_{10} > \pi_8 > \pi_9 > \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_6 = \pi_7$
- ■ $\pi_{10} > \pi_5 > \pi_8 > \pi_9 > \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_6 = \pi_7$
- ☐ I don't know

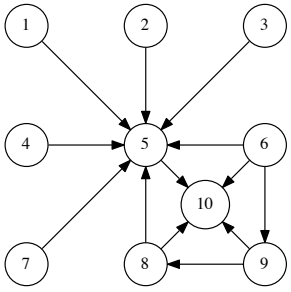


Figure 3: PageRank Graph.

12. Consider again the same graph shown in Figure 3, if $\theta = 0.001$, what would be the node with the highest PageRank?

- ■ 5
- ☐ 6
- ☐ 10
- ☐ I don't know

Question 2: Graph Conductance (15 points)

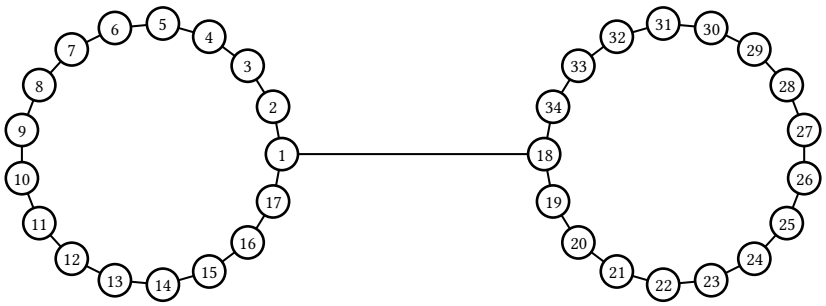


Figure 4: Graph G_1 with $2n = 34$ nodes.

1. (5 pts) Consider the graph G_1 with $2n$ nodes (n odd) composed of two cycle graphs linked by a single edge. An example of such graph with $n = 17$ ($2n = 34$) is shown above in Figure 4. Compute the conductance of the graph as a function of n.

   Solution: The graph G_1 with $2n$ nodes has $m = 2n + 1$ edges. The set leading to the best conductance can be easily identified as the set $S = \{1, \ldots, n\}$, leading to $|\delta S| = 1$ and $\pi_S = \pi_{S'} = 1/2$ by symmetry. Therefore, the conductance can be written as

$$\Phi_{G_1} = \frac{|\delta S|}{2m\pi_S\pi_{S'}} = \frac{1}{2(2n+1)\frac{1}{2}\frac{1}{2}} = \frac{2}{2n+1}.$$

2. (5 pts) Let $p_{ij}(t)$ be the probability that a random walk on G_1 starting at node i reaches node j after t steps. For an arbitrarily large n, among the following choices, which is the smallest t that guarantess that

$$|p_{ij}(t) - \pi_j| \leq 10^{-2}$$

for any nodes i and j? Circle the correct answer and show your calculations.

- $t = n$
- $t = n^2$
- $t = n^3$
- $t = n^4$
- $t = n^5$

Hint: You may find the following approximations useful: for x small, $(1-x)^n \approx e^{-xn}$, and $\log_{10}(e) \approx 0.5$.

Solution: From Cheeger bound, we have

$$|p_{ij}(t) - \pi_j| \leq \left(1 - \frac{\Phi_{G_1}^2}{8}\right)^t.$$

Using the approximation given in the hint, we get that

$$\left(1 - \frac{\Phi_{G_1}^2}{8}\right)^t \approx e^{-t\frac{\Phi_{G_1}^2}{8}}$$

. Therefore, we want to find t such that

$$\begin{aligned} e^{-t\frac{\Phi_{G_1}^2}{8}} &\leq 10^{-2} \\ \implies -t\frac{\Phi_{G_1}^2}{8}\log_{10} e &\leq -2 \\ \implies t \geq \frac{32}{\Phi_{G_1}^2} &= 8(2n+1)^2 = 32n^2 + 32n + 8 \approx 32n^2. \end{aligned}$$

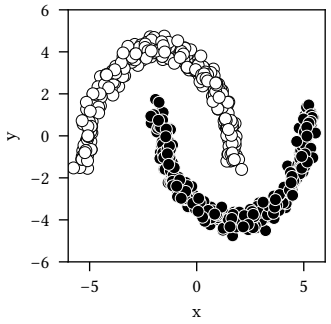
Therefore, we need at least $t = n^3$ steps.

3. (5 pts) Add as many edges as needed to the graph G_1 to reduce the conductance as much as possible. Explain your choice of edge(s).

Solution: To reduce the conductance, we want to increase the probability of a random walk staying in one of the compoenents. To do so, we can make the two cycles complete graph by adding all possible edge between nodes in $\{1, \ldots, n\}$, and similarly for $\{n+1, \ldots, 2n\}$. In terms of computations, this does not change $|\delta S| = 1$ and we still have $\pi_S = \pi_{S'} = 1/2$ by symmetry, but the number of edges m is maximized, leading to a maximum decrease in conductance.

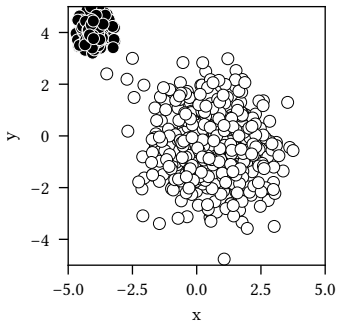
Question 3: Clustering (10 points)

1. (6 pts) Consider applying K-means and the Gaussian Mixture Model (GMM), with $K = 2$ classes, to cluster the following five datasets containing data points from two classes (the class of each point is shown on the figures in black and white colors). For each dataset and each algorithm, what assumption(s), if any, are violated and would lead to many misclassifications?



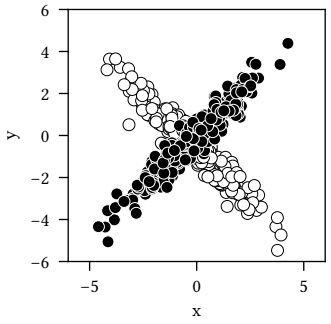
Dataset 1

- **K-Means:** *The two clusters are not linearly separable.*
- **GMM:** *The two clusters are not normally distributed.*



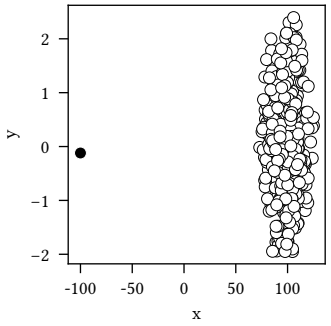
Dataset 2

- **K-Means:** *The two clusters do not have the same variance, which will be an issue and misclassfiy some of the outer points of the whie cluster within the black one.*
- **GMM:** *No assumption is violated, GMM would work well.*



Dataset 3

- **K-Means:** *The two clusters are not linearly separable.*
- **GMM:** *GMM would work well. However, since the cluster centroids overlap, the posterior probability distributions with be close to* $1/2$ *near the centroids.*



Dataset 4

- **K-Means:** *K-Means would likely work well.*
- **GMM:** *GMM would likely identify the cluster centroids correctly.*

2. (4 pts) When applying the GMM algorithm to Dataset 4, what would the likelihood of the model asymptotically converge to? Justify your answer in a few words. *Hint: What would the covariance matrix of the black cluster asymptotically converge to?*

Solution: Since there is only a single point in the black cluster, the M-Step would make the covariance matrix of this cluster tend to zero, leading to a probability density tending to infinity, as thus the likelihood of the model would tend to infinity.

Question 4: Recommender Systems (10 points)

In this problem, we consider a neighborhood-based collaborative filtering recommender.

Assumptions for all the subquestions below (please read carefully):

- The possible ratings are $-2, -1, 0, +1, +2$ (worst to best).
- Unless specified otherwise, we assume the user-user method is used.
- We ignore all the biases, i.e., assume that $\bar{r} = b_u = b_i = 0$ for all users u and items i. The predicted rating is simply given by

$$\widehat{r}_{ui} = \frac{\sum_{v \in L_u(i)} \text{sim}(u,v) \cdot r_{vi}}{\sum_{v \in L_u(i)} \text{sim}(u,v)}, \tag{1}$$

  where $L_u(i)$ is the set of all users v such that (a) $\text{sim}(u,v) > 0$, and (b) r_{vi} is defined (i.e., there is an entry at (v,i) in R).
- The similarity function is cosine similarity. In addition, $\text{sim}(u,v)$ is non-zero only if the size of the intersection of rated items by u and by v is at least two. In other words, if u and v have only one or zero items in common, the similarity is assumed zero (for example, $\text{sim}(6,7) = 0$).

The set of users is indexed by $1, 2, \ldots, 10$, the set of items by $a, b, \ldots, g$. The rating matrix R is

$$R = \begin{bmatrix} & -1 & +1 & +2 & & & \\ -2 & -1 & +1 & +2 & & & \\ -2 & -1 & +1 & +2 & & & \\ +2 & +1 & -1 & -2 & & & \\ +2 & +1 & -1 & -2 & & & \\ +2 & +1 & -1 & -2 & & & \\ & & & & -1 & +1 & +2 \\ & & & & -1 & +1 & +2 \\ & & & & -1 & +1 & +2 \\ & & & & +1 & -1 & -1 \end{bmatrix}, \tag{2}$$

where an empty field means "undefined".

1. (3 pts) User 1 would most likely hate movie a (i.e., $\widehat{r}_{1a} = -2$). You are working for the social media team of the production company promoting movie a, and you want to improve $\widehat{r}_{1a}$ by adding **one additional user (11) with arbitrary ratings**, so that $\widehat{r}_{1a}$ is maximized. Provide the additional row of R corresponding to user 11, and the corresponding $\widehat{r}_{1a}$.

a	b	c	d	e	f	g
+2	-1	+1	+2			

$$\widehat{r}_{1a} = (-2 - 2 + 2)/3 = -2/3. \tag{3}$$

2. (3 pts) Suppose a new movie h comes out, which nobody has rated so far. By again manipulating the ratings of the fake user 11, how many people can you trick into watching h? More specifically, for how many users can you force a predicted score of 2 for h? Justify your answer.

a	b	c	d	e	f	g	h
+2			-2	-1		+2	+2

The above solution is not unique, of course.

In this solution, 6 users (4, 5, 6, 7, 8, 9) would have a prediction of +2 for h.

A better solution (hard to find, so not needed for full score) is the following:

a	b	c	d	e	f	g	h
+2			-2	+2	-1	+2	+2

The gives positive similarity for users 7–10, so 7 total.

3. (2 pts) Suppose now that the number of users n and the number of items m is very large (e.g., millions). However, the ranking matrix R is sparse: every user has rated at most 3 items, and every item has been rated by at most 100 users. For a given user u, can you find an upper bound on the number of items i for which the user-user neighborhood method can generate a (non-zero) recommendation.

   User u has rated at most 3 items. These 3 items have been rated by at most 300 (297 really) other users v. These users v have each rated at most 2 other items, for a total of 600.

4. (2 pts) Same assumption on R as above (every user has rated at most 3 items, every item rated at most 100 times), but using the item-item method. For a given item i, find an upper bound on the number of users u for which the item-item neighborhood method can generate a (non-zero) recommendation.

   Item i has been rated by at most 100 users. These 100 users have rated at most 200 other items j. These items j have each been rated by at most 99 other users, for a total of 19800.

Question 5: Word2Vec (11 points)

The negative-sampling based objective for Word2Vec for a single training example, consisting of a positive sample (c, p) and k negative samples (c, n_i) $\forall i = 1, \ldots, k$, can be written as

$$J_{\text{neg}}(v_c, p, U) = \log\left(\sigma\left(u_p^T v_c\right)\right) + \sum_{i=1}^{k} \log\left(\sigma\left(-u_{n_i}^T v_c\right)\right).$$

Here c, p and n_i $\forall i = 1, \ldots, k$ are the indices of the center word, true context word and the k negative context words in the vocabulary $\mathcal{V}$, respectively. v_j and u_j, $j \in \{1, \ldots, |\mathcal{V}|\}$ are center and context word vectors corresponding to the word at index j in the vocabulary. U is a matrix whose columns are u_j $\forall j = 1, \ldots, |\mathcal{V}|$. $\sigma(x)$ is the logistic sigmoid function,

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

1. (4 pts) For the logistic sigmoid function $\sigma(x)$, compute the derivative with respect to x. Write your answer in terms of $\sigma(x)$.

$$\begin{aligned}
\sigma'(x) &= \frac{e^{-x}}{(1 + e^{-x})^2} \\
&= \frac{1}{1 + e^{-x}} \cdot \frac{e^{-x}}{1 + e^{-x}} \\
&= \sigma(x)\left(1 - \sigma(x)\right)
\end{aligned}$$

2. (4 pts) Compute the gradients of J_{neg} with respect to v_c, u_p and u_{n_i}, $\forall i = 1, \ldots, k$.

$$\begin{aligned}
\frac{\partial J}{\partial v_c} &= \frac{1}{\sigma\left(u_p^T v_c\right)} \cdot \sigma\left(u_p^T v_c\right)\left(1 - \sigma\left(u_p^T v_c\right)\right) u_p \\
&\qquad + \sum_{i=1}^{K} \frac{1}{\sigma\left(-u_{n_i}^T v_c\right)} \cdot \sigma\left(-u_{n_i}^T v_c\right)\left(1 - \sigma\left(-u_{n_i}^T v_c\right)\right)\left(-u_{n_i}\right) \\
&= \left(1 - \sigma\left(u_p^T v_c\right)\right) u_p - \sum_{i=1}^{K} \left(1 - \sigma\left(-u_{n_i}^T v_c\right)\right) u_{n_i} \\
&= \left(1 - \sigma\left(u_p^T v_c\right)\right) u_p - \sum_{i=1}^{K} \sigma\left(u_{n_i}^T v_c\right) u_{n_i}
\end{aligned}$$

$$\begin{aligned}\frac{\partial J}{\partial u_p} &= \frac{1}{\sigma\left(u_p^T v_c\right)} \cdot \sigma\left(u_p^T v_c\right)\left(1 - \sigma\left(u_p^T v_c\right)\right) v_c \\ &= \left(1 - \sigma\left(u_p^T v_c\right)\right) v_c\end{aligned}$$

$$\begin{aligned}\frac{\partial J}{\partial u_{n_i}} &= \frac{1}{\sigma\left(-u_{n_i}^T v_c\right)} \cdot \sigma\left(-u_{n_i}^T v_c\right)\left(1 - \sigma\left(-u_{n_i}^T v_c\right)\right)(-v_c) \\ &= \left(1 - \sigma\left(-u_{n_i}^T v_c\right)\right)(-v_c) \\ &= \sigma\left(u_{n_i}^T v_c\right)(-v_c)\end{aligned}$$

3. (3 pts) Looking at the gradients you computed for u_{n_i}, $\forall i = 1, \ldots, k$ in part 2, which among the following strategies for choosing negative context words do you think is likely to lead to a fast convergence of the training procedure? Justify your answer.

   (a) Sample uniformly at random from the vocabulary

   (b) Sample more often the words j whose context vectors u_j have a higher cosine similarity to the center word v_c

   (c) Sample more often the words j whose context vectors u_j have a lower cosine similarity to the center word v_c

   The training procedure using stochastic gradient descent is likely to converge fast when the gradient magnitude is large. Looking at the gradients for u_{n_i}, we can see that the gradient magnitude is maximized when $\sigma(u_{n_i}^T v_c)$ and consequently $u_{n_i}^T v_c$ is larger. This is more likely to happen when the cosine similarity between u_{n_i} and the center word v_c is larger. Thus we should choose the strategy (b), where we sample more often the words j whose context vectors u_j have a higher cosine similarity to the center word v_c

Question 6: Naïve Bayes, pLSI, and LDA (10 points)

1. (10 pts) In this class, we have seen several different generative models for text corpora. Here, we compare four corpora generated with each of the four models below. There are three different topics, and the word distribution of each topic is uniform. The word distributions for each topic are disjoint: all the words starting in "`c`" form the first topic vocabulary, all the words starting in "`p`" the second, and in "`s`" the third.

   (a) **pLSI**: the topic distribution of every document is a free parameter:

   (b) **LDA-high**: the topic distribution of each document is sampled from a Dirichlet distribution with parameter (α, α, α), with $\alpha \gg 1$:

   (c) **LDA-low**: the topic distribution of each document is sampled from a Dirichlet distribution with parameter (α, α, α), with α close to zero:

   (d) **Naïve Bayes (NB)**: every word of every document is generated i.i.d. according to the same word distribution. This word distribution is a mixture over the same per-topic word distributions above, with the same weight for all documents.

   Find the **one-to-one** match between the following four corpora (each one containing three documents of 20 words each) and the four models above (pLSI, LDA-high, LDA-low, NB) that generated it. Each model corresponds to exactly one corpus. Give a short explanation to justify your choices.

   Corpus 1:

   1: `pretzel pasta pretzel pizza pesto pizza peas pancake peas pizza pesto pretzel pasta pretzel pasta peas pesto peas pasta pizza`

   2: `spain sudan surinam spain switzerland switzerland switzerland sweden sudan surinam sudan spain sweden spain surinam switzerland spain spain switzerland spain`

   3: `pancake pancake peas pancake pesto pizza pasta pizza pizza pizza pesto pizza pretzel peas pretzel peas pasta pesto pancake pizza`

   Model: ☐ **pLSI** ☐ **LDA-high** ■ **LDA-low** ☐ **NB**

   Justification: *Topic distributions are very sparse (only 1 topic per document), so it is clearly LDA-low.*

Corpus 2:

1: `peas sweden pasta pretzel camel pizza pasta peas peas caterpillar camel pretzel sweden peas switzerland peas cat pretzel peas cow`

2: `spain pesto switzerland swaziland pizza pretzel camel pizza crab pasta sweden crow crow camel switzerland pancake pasta pretzel switzerland pizza`

3: `pasta cougar crow peas pancake pizza surinam pesto pizza camel pretzel spain pesto peas pancake pancake sudan pizza pesto sweden`

Model: ☐ **pLSI** ☐ **LDA-high** ☐ **LDA-low** ■ **NB**

Justification: *The three documents have roughly the same number of words from each topic, so their topic distributions seem to be the same. Therefore, this corpus was generated by NB.*

Corpus 3:

1: `pancake cow sweden sudan caterpillar pancake cow pizza camel pasta switzerland sudan pasta pancake spain cougar peas pesto cow sudan`

2: `cow crow cougar caterpillar caterpillar cow cow peas crow pretzel pizza caterpillar crab camel camel caterpillar cow cat camel cougar`

3: `sudan sweden peas pancake switzerland surinam cow crab pasta caterpillar spain cow caterpillar cow cougar caterpillar cow cougar crab surinam`

Model: ■ **pLSI** ☐ **LDA-high** ☐ **LDA-low** ☐ **NB**

Justification: *By elimination, this corpus was generated by pLSI.*

Corpus 4:

1: `surinam pancake cow pesto pizza pancake switzerland pancake pretzel camel cat camel swaziland surinam cow sudan surinam pesto cow camel`

2: `cow sweden spain switzerland cat pretzel spain pasta cat switzerland crow crow pretzel pretzel pretzel pasta pesto peas swaziland cat`

3: `cougar pesto cow crow sweden peas cougar pancake pretzel pesto swaziland pesto pizza sweden switzerland spain pancake caterpillar switzerland camel`

Model: ☐ **pLSI** ■ **LDA-high** ☐ **LDA-low** ☐ **NB**

Justification: *Topics are almost perfectly uniform (with 6 or 7 words of each topic per document), so it is clearly LDA-high.*

Question 7: Dimensionality Reduction (20 points)

You have a dataset of images of size 3×3 in which you are trying to find some structure. Given below are the 4 datapoints in the dataset given in the form of matrices. Note that these are 9 dimensional datapoints and the features are numbered from 1 through 9 in row-major order. Thus for example, the usual vector representation of the datapoint x_2 can be obtained as [0 1 0 0 1 0 0 0 1].

$$x_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad x_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$x_3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad x_4 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

You would like to compress the representation of these points through dimensionality reduction using Principal Component Analysis (PCA). All questions can be answered with little or no computation and the justifications should be based on the general properties of covariance matrices or PCA instead of computations wherever possible.

1. (8 pts) Among the following matrices, which one is the correct covariance matrix for the given data? For the other 5 matrices, give reasons to justify why it cannot be the correct covariance matrix for the data.

(a) $\Sigma_1 = \frac{1}{4} \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$

☐ Correct
■ Incorrect
If incorrect, justify:
Some of the variances (diagonal elements) are negative, which is impossible.

(b) $\Sigma_2 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$

☐ Correct
■ Incorrect
If incorrect, justify:
The factor $1/n$ is missing where n is the number of samples.

(c) $\Sigma_3 = \frac{1}{4}\begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$

☐ Correct
■ Incorrect
If incorrect, justify:
The matrix is not symmetric, which is impossible for a covariance matrix.

(d) $\Sigma_4 = \frac{1}{4}\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$

☐ Correct
■ Incorrect
If incorrect, justify:
Covariances are positive, which does not match the data. For instance the first two elements of the top row change in opposite directions.

(e) $\Sigma_5 = \frac{1}{4}\begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$

■ Correct
☐ Incorrect
If incorrect, justify:

(f) $\Sigma_6 = \frac{1}{4}\begin{bmatrix} 3 & 2 & 1 & 2 \\ 2 & 3 & 2 & 1 \\ 1 & 2 & 3 & 2 \\ 2 & 1 & 2 & 3 \end{bmatrix}$

☐ Correct
■ Incorrect
If incorrect, justify:
The matrix is not 9×9 which is the dimension of the covariance matrix for this data.

2. (2 pts) How many principal components would be needed to capture 100% of the variance? Why?

   We need 2 principal components to capture 100% of the variance since the rank of the covariance matrix is 2.

3. (5 pts) Which pair among the following are the first and second principal components for the data? For the other 4 pairs, give reasons to justify why they cannot be the first and second principal components for the data.

(a) $v_1 = \frac{1}{2}\begin{bmatrix}-1 & 1 & 0\\ 0 & 0 & 0\\ 0 & -1 & 1\end{bmatrix}, v_2 = \frac{1}{2}\begin{bmatrix}1 & -1 & 0\\ 0 & 0 & 0\\ 0 & -1 & 1\end{bmatrix},$

■ Correct
□ Incorrect
If incorrect, justify:

(b) $v_1 = \frac{1}{4}\begin{bmatrix}-1 & 1 & 0\\ 0 & 0 & 0\\ 0 & -1 & 1\end{bmatrix}, v_2 = \frac{1}{4}\begin{bmatrix}1 & -1 & 0\\ 0 & 0 & 0\\ 0 & -1 & 1\end{bmatrix},$

□ Correct
■ Incorrect
If incorrect, justify:
v_1 and v_2 are not unit norm.

(c) $v_1 = \frac{1}{2}\begin{bmatrix}1 & 1 & 0\\ 0 & 0 & 0\\ 0 & -1 & -1\end{bmatrix}, v_2 = \frac{1}{2}\begin{bmatrix}-1 & -1 & 0\\ 0 & 0 & 0\\ 0 & -1 & -1\end{bmatrix},$

□ Correct
■ Incorrect
If incorrect, justify:
v_1 and v_2 are not in the direction of maximum variance of the data. Note that the first two elements of the top row and the last two elements of the bottom row change in opposite directions (negative covariance) while the principal component points in the direction of positive covariance.

(d) $v_1 = \frac{1}{2}\begin{bmatrix}1 & 1 & 0\\ 0 & 0 & 0\\ 0 & -1 & -1\end{bmatrix}, v_2 = \frac{1}{2}\begin{bmatrix}-1 & -1 & 0\\ 0 & 0 & 0\\ 0 & 1 & 1\end{bmatrix},$

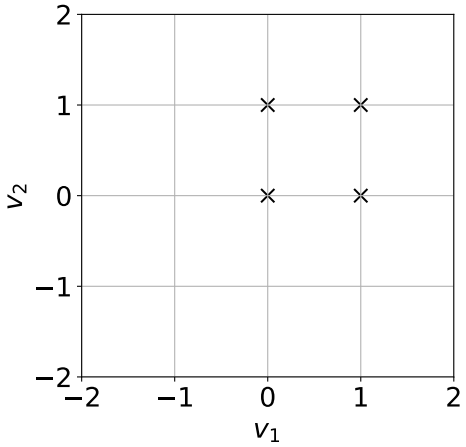
□ Correct
■ Incorrect
If incorrect, justify:
v_1 and v_2 are not orthogonal.

(e) $v_1 = \frac{1}{2}\begin{bmatrix}-1 & 0 & 1\\ 0 & 0 & 0\\ -1 & 0 & 1\end{bmatrix}, v_2 = \frac{1}{2}\begin{bmatrix}1 & 0 & -1\\ 0 & 0 & 0\\ -1 & 0 & 1\end{bmatrix},$

□ Correct
■ Incorrect
If incorrect, justify:
v_1 and v_2 are not in the direction of maximum variance. Note that there is no variance along the dimensions in the bottom-left and top-right corner.

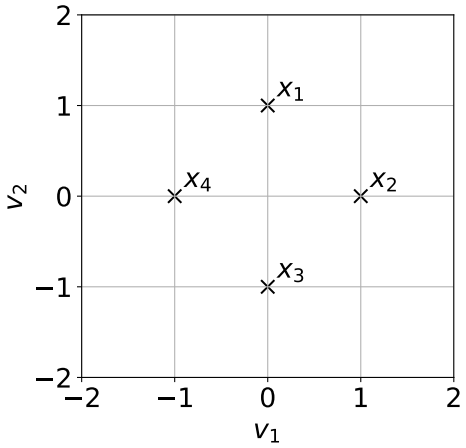
4. (5 pts) You now project the datapoints on the axes defined by the first 2 principal components. Which one of the following is the correct plot of the points? Label the points x_1, x_2, x_3, x_4 on the plot. For the other 4 plots, give reasons why it cannot be the correct plot of the points.

(a)



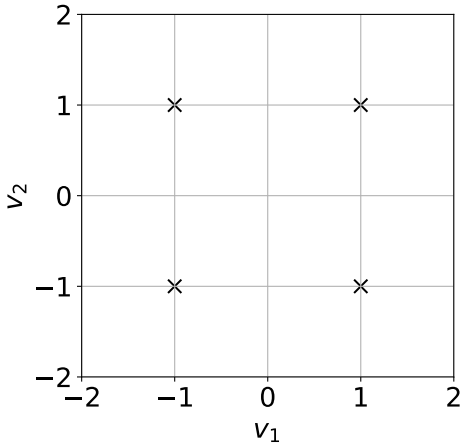
☐ Correct
■ Incorrect
If incorrect, justify:
The mean of the projected points is not zero.

(b)



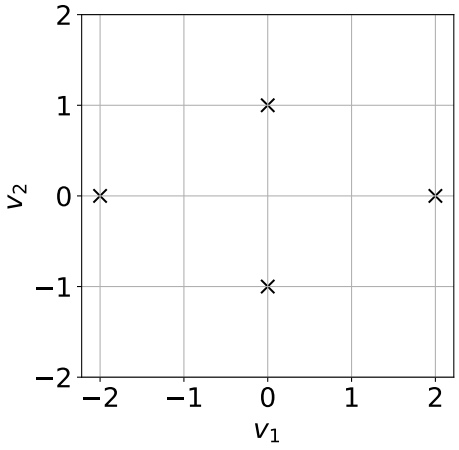
■ Correct
☐ Incorrect
If incorrect, justify:

(c)



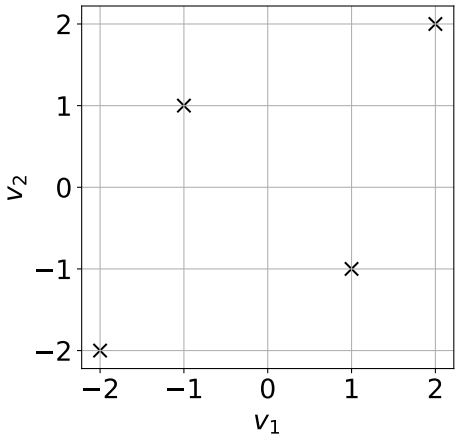
☐ Correct
■ Incorrect
If incorrect, justify:
The maximum variance of this projection is along the diagonal of the square, and this is not aligned with either of the principal components, which are along the axes.

(d)



☐ Correct
■ Incorrect
If incorrect, justify:
The two principal components should have equal variance by symmetry in this problem.

(e)



☐ Correct
■ Incorrect
If incorrect, justify:
There cannot be any correlation among the principal components, whereas here the components along v_1 and v_2 are positively correlated.