Internet Analytics (COM-308): Final Exam

May 31, 2017

Duration: **2h45**.

Total points: **100**.

Number of pages: **15**.

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 indicate it on the corresponding questions, **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**.

Name:
First name:
SCIPER number:
Signature:

Please leave blank.

1	2	3	4	5	6	Total
24	12	16	16	16	16	100

Question 1: Multiple Choice Questions (24 points)

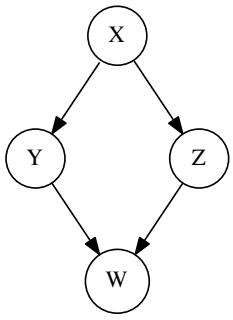
(24 pts) All questions have a single answer. Check the correct one. Grading:

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

1. Consider the undirected line graph with n nodes (n even), defined such that node i is connected to node $i+1$ for $i = 1, \ldots, n-1$. What is the conductance of this graph?
   - ☐ $\frac{2}{n-1}$
   - ☐ $\frac{1}{2(n-1)}$
   - ☐ $\frac{4}{n-1}$
   - ☐ I don't know

2. In which graph will a random walk converge the fastest? (all graphs have n nodes)
   - ☐ Line graph
   - ☐ Lattice graph (i.e., checkerboard graph) with at least one self-loop
   - ☐ Cycle graph (i.e., line graph with an additional edge connecting nodes n and 1)
   - ☐ I don't know

3. Consider the four random variables W, X, Y, Z whose joint distribution can be represented by the Bayesian network shown on the right. Which of the following independence assumptions is fulfilled by this distribution?



   - ☐ $Y \perp Z$
   - ☐ $Y \perp Z | X$
   - ☐ $Y \perp Z | W$
   - ☐ I don't know

4. Consider the following stream of data:

   ```
   a, a, b, c, b, c, a, b, a
   ```

   What is the output of the "heavy-hitters" algorithm, with $\theta = 0.4$?
   - ☐ $\{a\}$
   - ☐ $\{a, b\}$
   - ☐ $\varnothing$ (empty set)
   - ☐ I don't know

5. Let X_A and X_B be two Pareto random variables with exponent $\gamma_A = 2$, $\gamma_B = 3$, respectively. Knowing that $\mathbf{E}[X_A] = \mathbf{E}[X_B] = 6$, which of the following statements is true?
   - ☐ $P(X_A > 6) < P(X_B > 6)$
   - ☐ $P(X_A > 6) > P(X_B > 6)$
   - ☐ $P(X_A > 6) = P(X_B > 6)$
   - ☐ I don't know

6. Consider the following snippet of PySpark code.

```
data = sc.parallelize([(6, 2), (3, 7), (6, 7)])
output = (data.map(lambda x: (x[1], x[0]))
            .filter(lambda x: x[0] > 5)
            .reduceByKey(max)
            .collect())
```

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

☐ `[(6, 9)]`
☐ `[(6, 7)]`
☐ `[(7, 6)]`
☐ I don't know

7. Which of the following functions is **not** a probability density?

☐ $f(x) = \begin{cases} |x|, & \text{if } -\frac{1}{2} \leq x \leq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$
☐ $f(x) = \begin{cases} 2, & \text{if } -\frac{1}{4} \leq x \leq \frac{1}{4} \\ 0, & \text{otherwise} \end{cases}$
☐ $f(x) = \begin{cases} e^{-x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$
☐ I don't know

8. Which is the only possible result of a SVD?

☐ $\begin{bmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}$
☐ $\begin{bmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}$
☐ $\begin{bmatrix} -\frac{3\sqrt{2}}{2} & -\frac{3\sqrt{2}}{2} \\ -\frac{3\sqrt{2}}{2} & \frac{3\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}$
☐ I don't know

9. Which situation is more subject to overfitting?

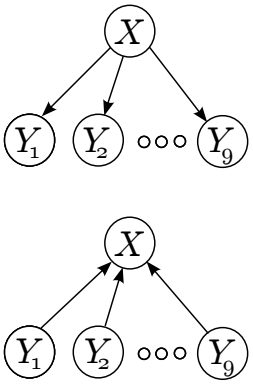
☐ You have few data and a high model complexity
☐ Your regularizer has a large weight
☐ You have a lot of data and a low model complexity
☐ I don't know

10. Which of these algorithms is a **supervised** learning algorithm?

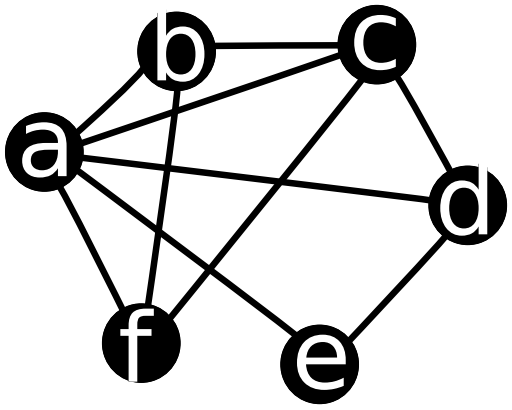
☐ Gaussian mixture model
☐ Naive Bayes
☐ Latent Dirichlet allocation
☐ I don't know

11. Which of the following statements about dimensionality reduction is correct?
    - ☐ PCA projects points into a lower-dimensional space such that pairwise distances are at most $(1-\epsilon)$ times shorter than the original distances (for some constant $0 < \epsilon < 1$).
    - ☐ A random projection always produces a set of points whose pairwise distances are close to the original distances.
    - ☐ To determine the r principal components for PCA, it is enough to know the best rank-r approximation of the data matrix X.
    - ☐ I don't know

12. Compare the number of parameters of the two Bayesian networks on the right (X and Y_i are all Bernoulli random variables).
    - ☐ The network on the top has more parameters.
    - ☐ The network on the bottom has more parameters.
    - ☐ Both networks have the same number of parameters.
    - ☐ I don't know



Question 2: Transitivity in Networks (12 points)



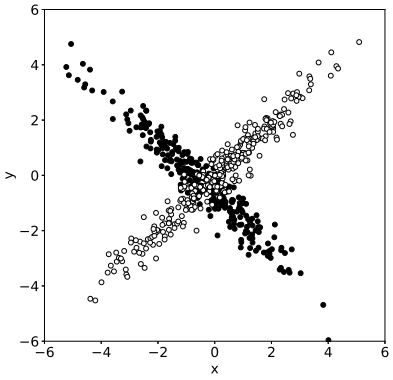
1. (4 pts) What is the clustering coefficient c_d of node d?

2. (8 pts) What is the weighted average clustering coefficient (transitivity) of the entire graph?

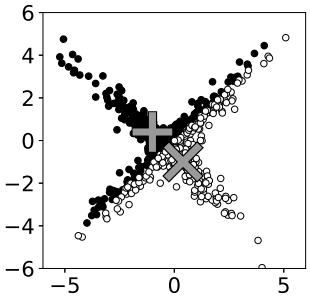
Question 3: Clustering (16 points)

1. (9 pts) The k-means algorithm is run on each of the three datasets below. The first column represents the true black and white clusters, the other ones represent clustering assignments with the centroids represented by the two crosses.

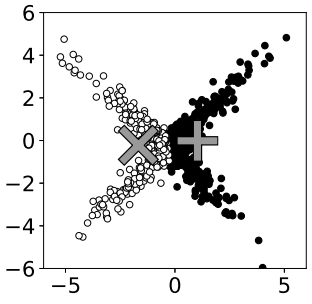
   For each dataset, indicate whether each plot is a plausible, implausible or impossible output of the k-means algorithm with $k = 2$ clusters. Justify with a few words.



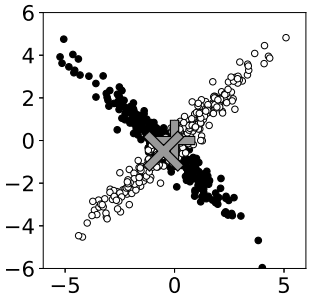
Ground truth
of dataset 1.



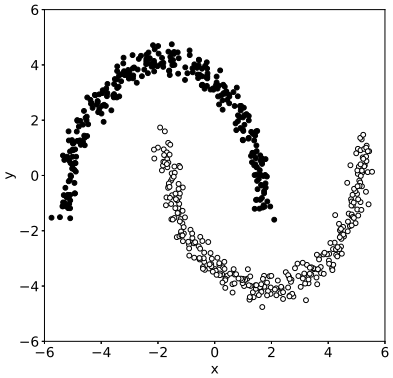
☐ plausible
☐ implausible
☐ impossible



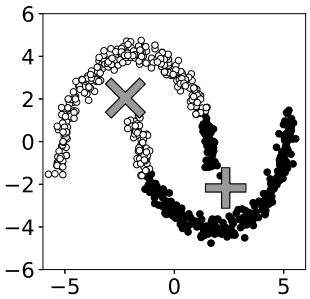
☐ plausible
☐ implausible
☐ impossible



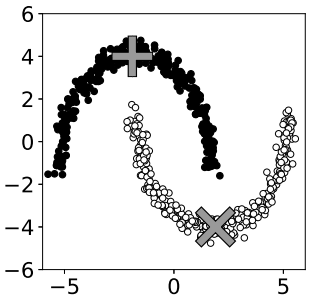
☐ plausible
☐ implausible
☐ impossible



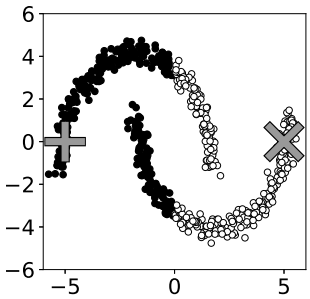
Ground truth
of dataset 2.



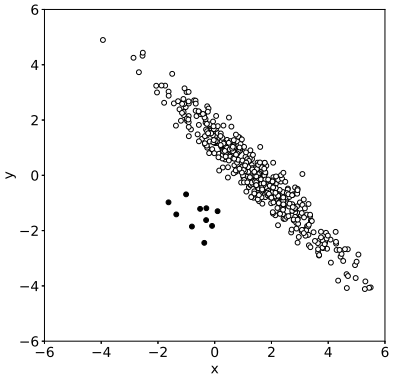
☐ plausible
☐ implausible
☐ impossible



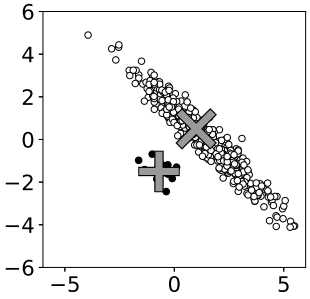
☐ plausible
☐ implausible
☐ impossible



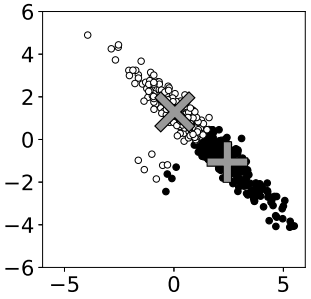
☐ plausible
☐ implausible
☐ impossible



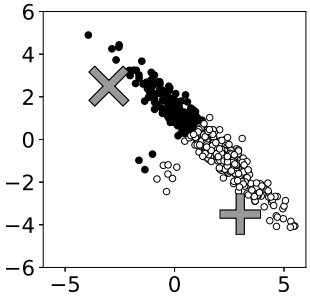
Ground truth
of dataset 3.



☐ plausible
☐ implausible
☐ impossible



☐ plausible
☐ implausible
☐ impossible



☐ plausible
☐ implausible
☐ impossible

2. (7 pts) Consider a Gaussian mixture model in which the priors of all mixture components are given by $\pi_k = 1/K$ for $k = 1, \ldots, K$, and the covariance matrices of all the mixture components are given by $\Sigma_k = \mathbf{I}$, where $\mathbf{I}$ is the $D \times D$ identity matrix. The probability density function of the k-th component is then

$$f(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{I}) = \frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2}\|\mathbf{x} - \boldsymbol{\mu}_k\|_2^2\right).$$

The estimation of the components is done using the following variant of the EM algorithm:

- The E-step is replaced by

$$\gamma_{nk} = \begin{cases} 1, & \text{if } k = \text{argmax}_{k'} \dfrac{\pi_{k'} f(\mathbf{x}_n|\boldsymbol{\mu}_{k'}, \mathbf{I})}{\sum_j \pi_j f(\mathbf{x}_n|\boldsymbol{\mu}_j, \mathbf{I})} \\ 0, & \text{otherwise.} \end{cases}$$

- In the M-step, only the centroids $\{\mu_k\}$ are updated (the component priors $\{\pi_k\}$ and the covariances $\{\Sigma_k\}$ are fixed).

Show that this algorithm is equivalent to the k-means algorithm.

Question 4: Recommender Systems (16 points)

We consider a recommender system with M users and N items. We suppose that $M > N$, and denote by $\mathcal{R}$ the set of user-item pairs that have been rated. We learn a rating model by minimizing the mean-squared error (MSE) on $\mathcal{R}$:

$$\text{MSE} = \sum_{(u,i)\in\mathcal{R}} (r_{ui} - \hat{r}_{ui})^2,$$

where r_{ui} is the true rating and $\hat{r}_{ui}$ is the predicted rating. The predicted ratings are modelled as

$$\hat{r}_{ui} = \bar{r} + b_u + b_i + \boldsymbol{p}_u^\top \boldsymbol{q}_i,$$

where $\boldsymbol{p}_u$ and $\boldsymbol{q}_i$ are vectors in $\mathbf{R}^K$.

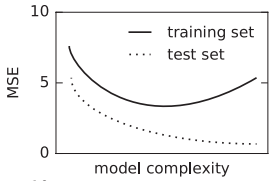
1. (2 pts) For a given user u, what is the partial derivative of the MSE with respect to b_u?

2. (2 pts) For a given item i, what is the gradient of the MSE with respect to $\boldsymbol{q}_i$?

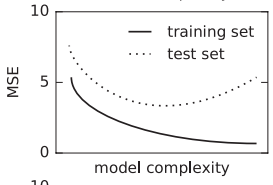
3. (3 pts) Let $K = N$. Find values of $\bar{r}$, $\{b_u\}$, $\{b_i\}$, $\{\boldsymbol{p}_u\}$, $\{\boldsymbol{q}_i\}$ that minimize the MSE. What is the corresponding value of the MSE?

   Hint: the minimum can be achieved with $\bar{r}$, $\{b_u\}$, $\{b_i\}$ *set to* 0.

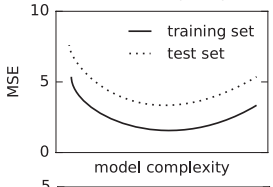
4. (3 pts) We now consider a regularized objective function. Below, we have plotted the MSE, evaluated on the training set $\mathcal{R}$ and on an independent test set, as function of the effective model complexity (low values mean strong regularization, high values mean weak regularization). Indicate whether each plot is *plausible*, *implausible* or *impossible*, and justify with a few words.



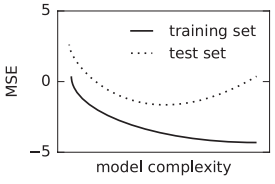
☐ plausible
☐ implausible
☐ impossible



☐ plausible
☐ implausible
☐ impossible



☐ plausible
☐ implausible
☐ impossible



☐ plausible
☐ implausible
☐ impossible

5. (6 pts) Consider the following Spark RDD.

```
# (userId, movieId, rating) triplets.
triplets = sc.parallelize([(1, 1, 3.5), (4, 9, 5.0), (3, 2, 1.5), ...])
```

(a) Write PySpark code to transform this RDD into one that maps users to number of ratings, using only RDD transformations (such as `map`, `filter`, `reduce`, `reduceByKey`, ...). For example, the output `[(1, 35), (3, 266)]` would mean "user 1 has 35 ratings, and user 3 has 266 ratings".

(b) Write PySpark and Python code to compute the mean of all ratings, starting from the `triplets` RDD.

Question 5: Document Classification (16 points)

We want to classify cooking recipes as either French or Swiss, based on the words in the recipe. We make the slightly simplifying assumption that only the words in the table below appear in a recipe. The table below gives the conditional probability of each word given the class (French or Swiss).

	Aromat	Garlic	Gruyere	Oysters	Pastis	Kirsch	Sour-cream	Truffles
$P(.\|F)$	0	0.5	0	0.1	0.1	0	0.2	0.1
$P(.\|S)$	0.2	0.3	0.2	0	0	0.1	0.1	0.1

We assume that most recipes are French, i.e., $P(F) = 0.9$.

1. (6 pts) What are the posterior probabilities and the likely class of the following delicacies: "Garlic + truffles", "Gruyere + Aromat", "Oysters + Gruyere"?

2. (6 pts) If we had an infinitely long recipe of the form "Garlic sour-cream garlic sour-cream...", what would its posterior probability be?

3. (4 pts) How large would you have to make the prior $P(S)$ in order to get a different posterior in the previous question?

Question 6: Topic Models (16 points)

Each question can be answered independently.

1. (8 pts) From the following corpus,

$$\begin{aligned} \{D_1 &= \text{“football football football football football sport sport sport”}, \\ D_2 &= \text{“money money money money money sport sport sport”}, \\ D_3 &= \text{“politics politics politics politics politics money money money”}\}, \end{aligned}$$

we extract $K = 3$ topics using *latent Dirichlet allocation* (*LDA*) with different priors α and β. The prior on **the topics per document** takes the following values:

$$\begin{aligned} \alpha_1 &= (100, 100, 100), \\ \alpha_2 &= (0.01, 0.01, 0.01), \\ \alpha_3 &= (10, 1, 1). \end{aligned}$$

The prior on the **words per topic** takes the following values:

$$\begin{aligned} \beta_1 &= (100, 100, 100, 100), \\ \beta_2 &= (0.01, 0.01, 0.01, 0.01). \end{aligned}$$

Note that the β_i's are defined over the words and are hence the same for each topic. We describe each topic T_i as a combination of words, and each document D_i as a combination of topics.

For each of the following cases, determine what are the values of α and β. Justify your choices in a few words.

(a)

$$\begin{aligned} T_1 &= 0.59\text{“sport”} + 0.27\text{“money”} + 0.11\text{“football”} + 0.03\text{“politics”} \\ T_2 &= 0.51\text{“politics”} + 0.33\text{“money”} + 0.15\text{“football”} + 0.01\text{“sport”} \\ T_3 &= 0.44\text{“football”} + 0.37\text{“sport”} + 0.15\text{“money”} + 0.04\text{“politics”} \\ D_1 &= 0.33T_1 + 0.33T_2 + 0.33T_3 \\ D_2 &= 0.33T_1 + 0.33T_2 + 0.33T_3 \\ D_3 &= 0.33T_1 + 0.33T_2 + 0.33T_3 \end{aligned}$$

$$\begin{aligned} \alpha &= \square\ \alpha_1\ \square\ \alpha_2\ \square\ \alpha_3 \\ \beta &= \square\ \beta_1\ \square\ \beta_2 \end{aligned}$$

(b)

$$
\begin{aligned}
T_1 &= 0.25\text{“money”} + 0.25\text{“sport”} + 0.25\text{“politics”} + 0.25\text{“football”} \\
T_2 &= 0.25\text{“money”} + 0.25\text{“sport”} + 0.25\text{“politics”} + 0.25\text{“football”} \\
T_3 &= 0.25\text{“money”} + 0.25\text{“sport”} + 0.25\text{“politics”} + 0.25\text{“football”} \\
D_1 &= 0.86T_1 + 0.07T_2 + 0.07T_3 \\
D_2 &= 0.86T_1 + 0.07T_2 + 0.07T_3 \\
D_3 &= 0.86T_1 + 0.07T_2 + 0.07T_3
\end{aligned}
$$

$$
\begin{aligned}
\alpha &= \square\ \alpha_1 \square\ \alpha_2 \square\ \alpha_3 \\
\beta &= \square\ \beta_1 \square\ \beta_2
\end{aligned}
$$

(c)

$$
\begin{aligned}
T_1 &= 0.62\text{“football”} + 0.36\text{“sport”} + 0.01\text{“money”} + 0.01\text{“politics”} \\
T_2 &= 0.62\text{“money”} + 0.36\text{“sport”} + 0.01\text{“football”} + 0.01\text{“politics”} \\
T_3 &= 0.62\text{“politics”} + 0.36\text{“money”} + 0.01\text{“sport”} + 0.01\text{“football”} \\
D_1 &= 0.98T_1 + 0.01T_2 + 0.01T_3 \\
D_2 &= 0.01T_1 + 0.98T_2 + 0.01T_3 \\
D_3 &= 0.01T_1 + 0.01T_2 + 0.98T_3
\end{aligned}
$$

$$
\begin{aligned}
\alpha &= \square\ \alpha_1 \square\ \alpha_2 \square\ \alpha_3 \\
\beta &= \square\ \beta_1 \square\ \beta_2
\end{aligned}
$$

(d)

$$
\begin{aligned}
T_1 &= 0.25\text{“money”} + 0.25\text{“sport”} + 0.25\text{“politics”} + 0.25\text{“football”} \\
T_2 &= 0.25\text{“money”} + 0.25\text{“sport”} + 0.25\text{“politics”} + 0.25\text{“football”} \\
T_3 &= 0.25\text{“money”} + 0.25\text{“sport”} + 0.25\text{“politics”} + 0.25\text{“football”} \\
D_1 &= 0.01T_1 + 0.98T_2 + 0.01T_3 \\
D_2 &= 0.98T_1 + 0.01T_2 + 0.01T_3 \\
D_3 &= 0.01T_1 + 0.01T_2 + 0.98T_3
\end{aligned}
$$

$$
\begin{aligned}
\alpha &= \square\ \alpha_1 \square\ \alpha_2 \square\ \alpha_3 \\
\beta &= \square\ \beta_1 \square\ \beta_2
\end{aligned}
$$

2. (3 pts) Let X be an $M \times N$ matrix of TF-IDF vectors, where M is the number of terms and N is the number of documents. We extract K topics from X using *LDA*. Explain how you would compute the similarity between any two documents.

3. (5 pts) Now, we decompose $X \approx USV^\intercal$ into a low-rank approximation by singular value decomposition, where U is the $M \times K$ matrix of left-singular vectors, S is the $K \times K$ diagonal matrix of singular values and V is the $N \times K$ matrix of right-singular values. Let $\mathbf{q}$ be a one-hot single-term query vector of size $M \times 1$ (i.e., $q_i = 1$ if the index of the query term is i and 0 otherwise).

   Show how to efficiently obtain the index d^* of the document in X that is the most relevant to $\mathbf{q}$ in terms of cosine similarity.

   Hint: You implemented this in the last lab. Start by mapping $\mathbf{q}$ *onto the latent concept-space.*