**Machine Learning BLG454E, July, 5 2020 120mins, Final Exam**

**Duration:** 120 minutes.

*Write your answers neatly* ***in the space provided for them****. Type in your name on the header of this Word document. Good work! ☺*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Q1** | **Q2** | **Q3** | **Q4** | TOTAL |
| **10** | **20** | **20** | **50** | **100** |
|  |  |  |  |  |

**Upload the filled-in Word doc file into Ninova before 14:30pm. Late submissions will not be accepted. PDF files will not be accepted.**

**Also, beware of cheating —I don’t need to remind you.**

**Keep your answers CLEAR, CONCISE, and SHORT. This will count in grading ☺**

**Question 1 [10 points]**

For each statement copy paste this check sign (✓) inside the open brackets [ ] only if the statement is correct. [*Hint: there are only two correct answers for each statement*]. [10 points]

**a) Principle component analysis (PCA) is a dimension reduction algorithm, which:**

* [ ] selects existing features by projecting the data onto new directions.
* [✓] extracts new features by projecting the data onto new directions.
* [ ] is suitable for classification as it performs dimensionality reduction in a supervised manner.
* [✓] depends on the covariance matrix of the input data.

**b) Infinite feature selection is:**

* [ ] a wrapper-based feature selection method.
* [✓] filter-based feature selection method.
* [ ] based on examining the sample features independently.
* [✓] based on examining the joint relationships across data features.

**c) Dual Lagrangian optimization for a constrained objective:**

* [✓] is based on minimizing a convex primal function with respect to its primal variables.
* [✓] is based on maximizing a concave dual function with respect to its dual variables.
* [ ] allows to solve the dual problem by optimizing the primal function using gradient descent.
* [ ] allows to solve the dual problem by optimizing the primal function using Newton method.

**d) Linear regression**

* [✓] is a supervised learning technique.
* [ ] it allows to classify samples into different groups.
* [ ] works well on data with input features non-linearly dependent on the output scores.
* [✓] aims to learn a model that captures the relationship between the input features and the target scores.

**e) Bayes classifier:**

* [ ] is a transductive classification technique.
* [✓] produces decision boundaries perpendicular to the lines connecting means of classes if the covariance matrix , where denotes the variance of each class and the identity matrix.
* [✓] produces non-linear boundaries when the discriminant function is quadratic.
* [ ] only depends on the points that lie on the decision boundary separating two classes for estimating the optimal discriminant function.

**Question 2 [20 points]**

Do not change the size of the columns. The gray shaded area is for grading. Don’t use that.

Describe and state **one** difference and **one** similarity between:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **1 Difference** | **1 Similarity** | **grade** |
| Bagging and boosting | Bagging makes random sampling with replacement, boosting makes random sampling but over weighted data | They both use multiple models to get better resulting model |  |
| Bayes classifier and support vector machines | Bayes classifier get help from statistical probability, SVM uses Lagrange function to separate classes | Both are used in classification problems |  |
| Feature selection and feature extraction methods | Feature selection selects some of existing features, feature extraction creating new feature from existing features | Both are dimension reduction methods |  |
| Newton’s method and Gradient descent for optimizing loss functions | Newtons method does not require learning rate, Gradient descent requires | They both used in finding optimal parameters for a model |  |
| PCA (principle component analysis) and CCA (canonical correlation analysis) | CCA uses multiple views of same samples, PCA makes feature extraction from only one view | They both project features |  |

**Question 3 [20 points]**

Do not change the size of the columns. The gray shaded area is for grading. Don’t use that.

|  |  |  |
| --- | --- | --- |
| **Question** | **Answer** | **grade** |
| Briefly explain the key steps of K-Means algorithm for data clustering into K clusters. [8 points] | 1. Selects K random different points 2. Make groups based on sample points’ distance to these selected points 3. Update center points after creating clusters by averaging their values 4. Repeat step 2 – 3 again until we find stable situation which clusters remain same |  |
| Can K-means correctly identify the two clusters in this dataset? Answer by yes or no and justify your choice. [4 points]  ../../../../../../../../../Desktop/Screen%20Shot%202020-06-21%20  dataset 1 | No. Because K means does not count internal relationships of these clusters. By looking visually, they are separable easily but K means selects random points and make clusters based on distance.    For instance, K means clusters these two selected regions since it uses distance to cluster points |  |
| Can hierarchical clustering work on dataset 1? Justify your answer [4 points] | No because hierarchical clustering works with distances between data points as well. |  |
| Which clustering algorithm would handle well the distribution of dataset 1? What is the core idea of such algorithm? [4 points] | Spectral clustering can handle well this dataset since it uses Affinity matrix to examine internal relationships |  |

**Question 3 [50 points] (Read the full annex paper)**

In the attached PDF, you will find a research paper using matrix factorization techniques for biological data analysis.

Siy, P. W., Moffitt, R. A., Parry, R. M., Chen, Y., Liu, Y., Sullards, M. C., ... & Wang, M. D. (2008, October**). Matrix factorization techniques for analysis of imaging mass spectrometry data.** In *2008 8th IEEE International Conference on BioInformatics and BioEngineering* (pp. 1-6). IEEE.

Read the paper in full and answer the following questions in the specified space.

**Important note**: Do not change the size of the columns. The gray shaded area is for grading. Don’t use that. Type in your answer in the middle column. **No copy-pasting sentences from the paper. Use your own words when crafting your answer.**

|  |  |  |
| --- | --- | --- |
| 1) What is the main problem to solve in the Siy et al. paper? [5 points] | In a biological data, proposed models in paper, aims to find concentration and location of molecules by finding most significant molecules and revealing their relationships visually. |  |
| 2) What are the three methods explored to solve this problem? Specify in the field TYPE the category of each method: *supervised, unsupervised, or semi-supervised*. [5 points] | **Method 1 [**Unsupervised**]:** PCA  **Method 2 [**Unsupervised**]:** ICA  **Method 3 [**Unsupervised**]:** NMF |  |
| 3) Briefly explain how ICA algorithm works [5 points] | By maximizing non Gaussianity of component images, ICA tries to separate given inputs. As a result ICA maximizes images’ statistical independence but also keep them uncorrelated so that each image gives maximum information. |  |
| 4) Point out two differences between ICA and PCA? [10 points] | **Difference 1**: ICA makes the data uncorrelated and maximizes their independence, but PCA just makes uncorrelated if the data is non-Gaussian  **Difference 2**: ICA maximizes non Gauussianity of input, PCA maximizes variance |  |
| 5) Point out two differences between NMF and PCA? [10 points] | **Difference 1**: NMF benefits from non negative constraint but PCA benefits from orthogonality constraint  **Difference 2**: PCA works best with Gaussian data to get independent results while NMF does not require the data has Gaussian distribituon |  |
| 6) In NMF, the data matrix is factorized into the product of two matrices and such that  Specify the dimensionality of the different variables stated in the right column based on the paper mathematical notations. [10 points] | The number of samples (images) is:  The original number of features is:  The dimension of matrix:  The dimension of matrix:  Following NMF of , the number of the reduced features is: |  |
| 7) What is (are) the best performing method(s) and why? [5 points] | NMF and ICA performs better than PCA since they produced less noisy results. |  |