CSCI-UA 9473 - Introduction to Machine Learning
Final II

Augustin Cosse

July 2022

Total: 38 points
Total time: 2h00
General instructions: The exam consists of 2 parts, a first part focusing on supervised learning (including 4 subquestions), and a second part focusing on unsupervised learning (including 4 subquestions). Once you are done, make sure to write your name on each page, then take a picture of all your answers and send it by email to acosse@nyu.edu. In case you have any question, you can ask those through the chat. Answer as many questions as you can starting with those you feel more confident with.

Question 1 (Supervised Learning 18pts)

1. [5pts] *Indicate whether the following statements are true or false*

True / False	*The kernel $K(x,y) = e^{x_1 y_1}$, where x_1 and y_1 are the first components in the $\boldsymbol{x}$ and $\boldsymbol{y}$ vectors, is not a valid kernel*
True / False	*Logistic regression cannot be trained with gradient descent*
True / False	*We expect a model with high variance to generalize better than a model with high bias*
True / False	*In linear regression, highly correlated features will lead to unstable estimates for the regression coefficients*
True / False	*A finite non linearly separable dataset can always be made linearly separable in another space*
True / False	*A symmetric matrix is positive semidefinite if all its eigenvalues are non negative*
True / False	*Cross validation can be used to mitigate overfitting*
True / False	*In Support Vector Machines, the only points that contribute to the expression of the classifier are the closest points to the classifier*
True / False	*Support vector machines can be extended to non linearly separable datasets by combining them with an appropriate kernel*

2. [5pts] *We consider the neural network and dataset shown in Fig 3. The intercepts are implicitly assumed.*
   a) [2pts] *Can the neural network correctly classify the dataset? (Motivate your answer)*
   b) [3pts] *Apply the backpropagation algorithm to obtain an expression of the gradient for the mean-squared (RSS) loss of y, with the target value t, with respect to the weights w_{22} and w_{11}, assuming a sigmoid activations for the hidden layer.*

3. [3pts] *Explain why the kernel trick allows us to solve a learning problem (e.g. a regression problem) in a high dimensional feature space without significantly increasing the run time.*

4. [5pts] *We consider a set of training examples $\{\boldsymbol{x}^{(i)}, t^{(i)}\}$ with $\sum_{i=1}^{N} t^{(i)} = 0$ and $\sum_{i=1}^{N} x^{(i)} = 0$. We let $\boldsymbol{X}$ denote the corresponding feature matrix and $\boldsymbol{t} = [t^{(1)}, \dots, t^{(N)}]$ the target vector.*

a) [3pts] *Give the general formulation of the Ridge loss and derive the gradient iterations for that particular loss.*

b) [2pts] *Show that for the centered dataset $\{\boldsymbol{X}, \boldsymbol{t}\}$, the ridge regression estimates $\hat{\beta}_j$ can be obtained by ordinary least squares regression on an augmented dataset obtained by (1) augmenting the centered feature matrix $\boldsymbol{X}$ with p additional rows $\sqrt{\lambda}\boldsymbol{I}$ and (2) augmenting $\boldsymbol{t}$ with p zeros.*

Question 2 (Unsupervised learning 20pts)

1. [5pts] *Indicate whether the following statements are true or false*

True / False	*Both PCA and the EM algorithm can be used to learn a latent representation of the data*
True / False	*Principal components are always orthogonal to each other*
True / False	*K-means is a clustering algorithm that always converge*
True / False	*Considering a predefined number of clusters K, globally minimizing K-means is NP-hard*
True / False	*Hierarchical clustering methods require a predefined number of clusters, much like K-means*
True / False	*Independent Component Analysis can be solved through a maximization of the likelihood and arbitrary priors on the sources*
True / False	*The output of PCA is a new representation of the data that is always of lower dimensionality than the original feature representation*
True / False	*The total number of parameters estimated during the Maximization step of the EM algorithm for a Gaussian mixture model made of 3 Gaussian distribution is 9 lower dimensionality than the original feature representation*
True / False	*Market Basket Analysis tries to find groups of items that frequently appear together in a given set of transactions*

2. [5pts] *We consider a data matrix $\boldsymbol{X}$ and we want to learn the best dimension 2 subspace to represent the data. Explain how you would proceed (all details, including pseudo-code)*

3. [5pts] *In this question you will perform K-means clustering manually, with $K = 2$, on a small example with $n = 6$ observations and $p = 2$ features. The observations are given in Table (1) and represented in Fig 1.*

   (a) [1pt] *Start by computing the centroid for each cluster (red and blue) (and detail your calculations)*

   (b) [1pt] *Perform the assignment step by relying on the Euclidean distance. Report the cluster labels for each observation.*

   (c) [1pt] *Repeat (a) and (b) once and provide the final assignment and the resulting position of the centroids.*

   (d) [1pt] *Assume that we initialize K-means on the dataset shown in Fig 1 with $K = 1$, 2, 3, 4 5 and 6. Assume that we compute the Within-Cluster-Sum of Squares Error (WCSS). Give the general expression of the WCSS. Then sketch the evolution of the WCSS as a function of K.*

4. [5pts] *We consider the datasets shown in Fig 2. One of these figures was obtained by running single linkage agglomerative clustering and stopping at $K = 3$. The other was obtained by running complete linkage agglomerative clustering and stopping at the same value.*

   (a) [2pts] *For each figure, indicate whether it corresponds to the result of the single linkage clustering or the complete linkage clustering.*

   (b) [3pts] *Give the distances that are minimized in single linkage clustering, complete linkage clustering and group average clustering.*

Table 1: K-means dataset

Obs. #	x_1	x_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

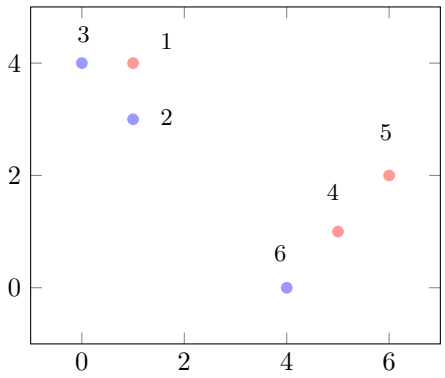

Figure 1: K-means clustering.

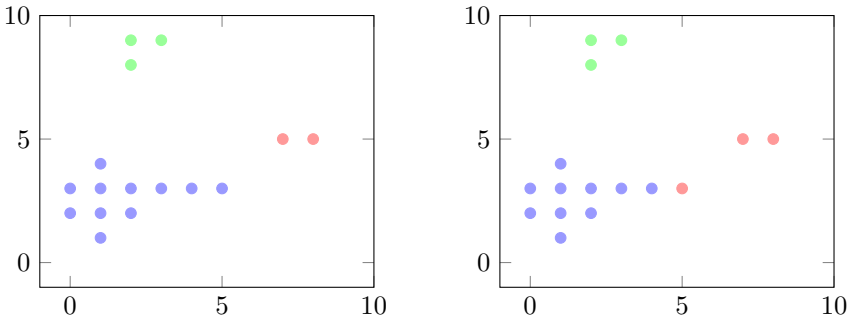

Figure 2: Hierarchical clustering.

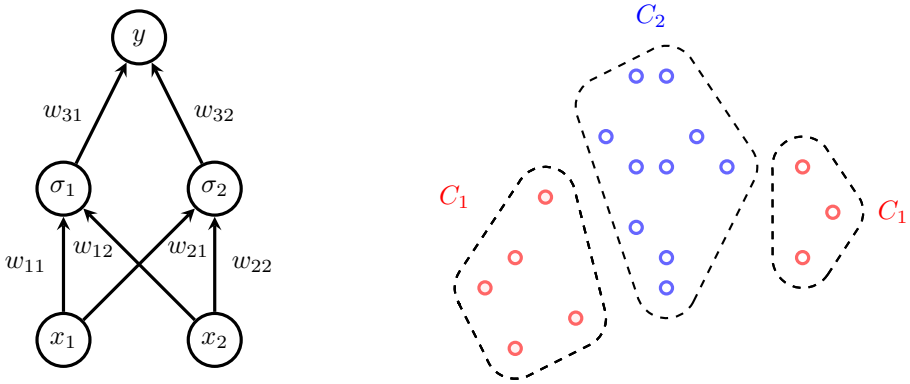

Figure 3: Neural Network and Dataset for Question 1.2 .