CSCI-UA 9473 - Introduction to Machine Learning
Final I

Augustin Cosse

July 2022

Total: 40 points
Total time: 2h00
General instructions: The exam consists of 2 parts, a first part focusing on supervised learning (including 5 questions), and a second part focusing on unsupervised learning (including 3 questions). 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 20pts)

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

True / False	*A classifier trained on less training data is less likely to overfit*
True / False	*One can perform linear regression using either matrix algebra or using gradient descent*
True / False	*Using cross validation to select the hyperparameters will guarantee that our model does not overfit*
True / False	*The number of parameters in a parametric model is fixed, while the number of parameters in a non-parametric model grows with the amount of training data.*
True / False	*As model complexity increases, bias will decrease while variance will increase*
True / False	*Compared with ordinary least squares regression, ridge regression has smaller bias and larger variance*
True / False	*Compared with ordinary least squares regression, ridge regression has larger bias and smaller variance*
True / False	*Pooling layers in convolutional neural networks reduce the spatial resolution of the image*

2. [*4pts*] *Derive a gradient descent algorithm that minimizes the sum of squared errors for a variant of a perceptron (i.e. one neuron) where the output y of the unit depends on its inputs x_i as follows*

$$y(\boldsymbol{x}) = w_0 + w_1 x_1 + w_1 x_1^3 + w_2 x_2 + w_2 x_2^3 + \ldots + w_n + w_n x_n^3$$

Give your answer in the form $w_i \leftarrow w_i + \ldots$ for $1 \leq i \leq n$.

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 running time.*

4. [*8pts*] *Consider a supervised learning problem in which the training examples are points in a 2-dimensional space. The positive examples are $(1,1)$ and $(-1,-1)$. The negative examples are in $(1,-1)$ and $(-1,1)$.*

(a) [1pt] *Are the positive examples linearly separable from the negative examples in the original space?*

(b) [3pts] *Consider the feature transformation* $\phi(x) = [1, x_1, x_2, x_1x_2]$ *where* x_1 *and* x_2 *are respectively the first and second coordinates of a generic example* x*. The prediction function is* $y(x) = \boldsymbol{w}^T\phi(\boldsymbol{x})$ *in this feature space. Give the coefficients,* $\boldsymbol{w}$ *of a maximum margin decision surface separating the positive examples from the negative examples (You should be able to do this by inspection, without any significant computation)*

(c) [2pts] *Add one training example to the graph so the total five examples can no longer be linearly separated in the feature space* $\phi(x)$ *defined above. Sketch the result in the original space.*

(d) [2pts] *What kernel* $K(x, x')$ *does this feature transformation correspond to?*

Question 2 (Unsupervised 20pts)

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

True / False	*K means returns the global minimum of the clustering problem*
True / False	*Given a data matrix* $\boldsymbol{X} \in \mathbb{R}^{n\times d}$*, where* $d \ll n$*, if we project our data onto a k-dimensional subspace using PCA where k equals the rank of* $\boldsymbol{X}$*, we recreate a perfect representation of our data with no loss*
True / False	*Using 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 Kmeans*
True / False	*Independent Component Analysis is an example of a factor analysis model*
True / False	*To work, Independent Component Analysis requires the sources to follow a Laplace distribution*

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] *Give the pseudo-code for the K-means algorithm. How can one handle empty clusters (+pseudo code)?*

4. [5pts] *Suppose that we have four observations, for which we compute the dissimilarity matrix, given by*

$$D = \begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

For instance, the dissimilarity between the first and second observations is 0.3 and the dissimilarity between the second and fourth observation is 0.8

a) [1,5pts] *On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using complete linkage. Be sure to indicate the observation corresponding to each leaf in the dendrogram.*

b) [1,5pts] *Repeat (a) this time using single linkage clustering*

c) [1pt] *Suppose that we cut the dendrogram in (a) such that two clusters result. Which observations are in each cluster?*

d) [1pt] *Suppose that we cut the dendrogram obtained in (b) such that two clusters result. Which observations are in each cluster?*