CSCI-UA 9473 - Introduction to Machine Learning
Final I

Augustin Cosse

May 2022

Total: 45 points
Total time: 1h15
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 25pts)

1. *Indicate whether the following statements are true or false (5pts)*

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. *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*

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

Keep in mind that the neural network is a non linear model. Give your answer in the form $w_i \leftarrow w_i + \ldots$ for $1 \leq i \leq n$. [7pts]

3. *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. [3pts]*

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

(b) *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)[3pts]*

(c) *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. [2pts]*

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

5. *Explain the difference between a generative and a discriminative classifier. [2pts]*

Question 2 (Unsupervised 20pts)

1. *Indicate whether the following statements are true or false (5pts)*

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. *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)[5pts]*

3. *Give the pseudo-code for the Kmeans algorithm. How can one handle empty clusters (+pseudo code) [5pts] ?*

4. *Provide the three main types of agglomerative clustering algorithms and explain how the clusters are merged in each type [5pts].*