Introduction to Machine Learning
Fall 2019
MidTerm Exam
Date: October 2019
Time: 1h15

Name: _____

Class: CSCI-UA 9473

Lecturer: *Augustin Cosse*

This Exam is on 55 points. The contribution of each question to the final grade is indicated in bold. Remember that talking is not allowed. Calculators and phones are not permitted. Write your answers in the blank spaces or use the other side of the page and indicate the corresponding question number. You can use additional paper but you should then clearly indicate the connection to the corresponding question.

1. (5 points) Explain the difference between
   - Supervised and unsupervised learning
   - Parametric and non parametric models

2. (10 points) Regression

   2.5pts We consider a set of N pairs $\{\boldsymbol{x}_i, t_i\}_{i=1}^N$ where $\boldsymbol{x}_i \in \mathbb{R}^D$ and $t_i \in \mathbb{R}$. We would like to learn a linear regression model $y(\boldsymbol{x}) \approx \boldsymbol{\beta}^T\boldsymbol{x} + \beta_0$ such that $y(\boldsymbol{x}_i)$ can be used to predict the targets t_i as well as possible. Explain how you would proceed to learn the $\boldsymbol{\beta}, \beta_0$. Give the mathematical expression of the loss that you would minimize and explain each of the terms.

   2.5pts When there is a risk of overfitting, for example when a large number of features are used, it is possible to control the complexity of the model and hence the variance, by adding a regularization term. Give three examples of such penalties.

   2pts You are asked to find the parameters $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\beta_0 \in \mathbb{R}$ that minimize the RSS criterion with ℓ_2 regularization (you can assume that this regularization term is weighted by a parameter λ). Write down the pseudo code of the gradient algorithm that you would use to find those parameters. You can assume that the dataset is given by $\{\boldsymbol{x}_i, t_i\}_{i=1}^N$.

   3pts We want to incorporate, in the pseudo-code, a few lines that will select an optimal value for λ. Explain how you would proceed and give the modified pseudo code.

3. (6 points) Classification

   1pts We have seen in class that learning a binary classifier can be done by minimizing the same RSS criterion as the one that was used in regression. Explain this idea

   5pts We now want to extend this idea to learn a multiclass classifier. Give three possible approaches to get a multiclass classifier from the minimization of the RSS criterion. Explain each of those approaches, including how the final class of a new prototype can be determined once the parameters of the classifiers have been learned.

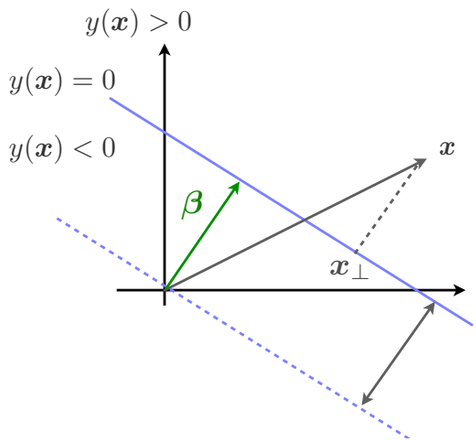

Figure 1: Material for Question 5.

4. (7 points) Perceptron Learning rule

   2pts Give the diagram and mathematical formulation $y(\boldsymbol{x}) = \ldots$ of the perceptron classifier. Explain each of the terms and indicate them on the diagram.

   3pts The parameters of the perceptron can be learned through the so-called perceptron learning rule which is repeated until all the prototypes have been correctly classified. Explain this rule and give the pseudo-code for the Perceptron algorithm

   2pts The Perceptron algorithm comes with an associated Theorem. Explain this Theorem (i.e when does the algorithm work and how efficient is it?).

5. (9 points) Support Vector Machines

   4pts Consider Fig. 1. Explain this Figure and use it to give the expression of the distance of a point $\boldsymbol{x}$ to the plane $y(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x} + \beta_0$.

   3pts Using the expression that you derived for the distance of a point to a plane, explain (in particular give the mathematical expression of the classifier) how you would proceed to learn a maximal margin classifier.

   2pts Maximal margin classifiers (also known as Support vector machines or sparse vector machines) are particularly efficient for two main reasons. Explain those reasons with a drawing.

6. (10 points) We consider a one hidden layer neural network with 10 hidden units. We associate to each neuron an activation function $\sigma(x)$ given by the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$

   1pts Represent the sigmoid activation. What values can be returned by this activation?

   4pts Represent the neural network with a diagram and give the associated mathematical expression (You don't need to represent all the connections as long as your explanation of each of the terms and their connection to the diagram is clear).

   1pts We still consider the sigmoid activation given above. Show that the derivative of this function satisfies the relation $\sigma'(x) = \sigma(x)(1 - \sigma(x))$.

   4pts We want to minimize the log-loss. As a reminder, this objective can be defined for a set of n prototypes $\{\boldsymbol{x}_n, t_n\}_{n=1}^N$ as

$$\ell = -\frac{1}{N}\sum_{n=1}^{N} t_n \log(y(\boldsymbol{x}_n)) + (1 - t_n)\log(1 - y(\boldsymbol{x}_n)) \tag{1}$$

   We place ourselves in a stochastic gradient descent framework where the gradient updates are done with respect to a single pair $\{\boldsymbol{x}_n, t_n\}$ at a time. Explain how the output weights can be learned in such a framework[1]

   Bonus, 3pts So far, we have considered a single output. Explain how the model could be extended to a multi-class classification problem.

[1] Start by computing $\frac{d\ell}{dy_n}$. Then use the chain rule as well as $\frac{\partial y_n}{\partial a_{\text{out}}}$ to show that $\frac{d\ell}{da_{\text{out}}} = y_n - t_n$

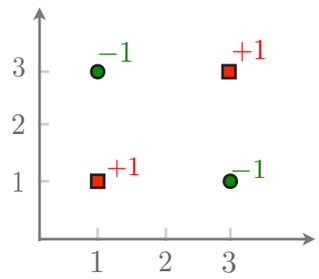

Figure 2: Question 8.

7. (5 points) Describe the difference between the Bayesian and the Frequentist frameworks in parameter estimation. Illustrate the difference between those two frameworks using Bayes' Theorem. Give an estimator from each framework.

8. (3 points) Consider the dataset shown in Fig. 2.

   - Is this dataset linearly separable?
   - What Kernel would you use to learn a classifier from that dataset? What would be the (general) expression of the resulting classifier?