CSCI-UA 9473 - Introduction to Machine Learning
Midterm II

Augustin Cosse

June 2022

Total: 35 points
Total time: 2h00
General instructions: The exam consists of 3 questions (each question consisting itself of several 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 (16pts)

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

True / False	*When maximizing a likelihood function through gradient descent, the k^{th} iterate is given by adding the gradient of the function (scaled by the learning rate) to the $(k-1)^{th}$ iterate.*
True / False	*The linear model selected by LASSO will, in general, contain more vanishing coefficients than the model learned by RIDGE*
True / False	*When learning a linear regression model, adding features will increase the bias*
True / False	*Assuming that $\boldsymbol{x}$ is used to denote the feature vectors and that t is used to denote the targets, discriminative classifiers learn a model for $p(\boldsymbol{x}\|t)$ while generative classifiers learn a model for $p(t\|\boldsymbol{x})$*
True / False	*Linear Discriminant Analysis relies on the assumption that the examples within a class are distributed according to a Multivariate Gaussian distribution*
True / False	*The naive Bayes classifier assumes conditional independence of the features with respect to the class*
True / False	*The naive Bayes classifier is an instance of a generative classifier*
True / False	*Applying the one-vs-one approach to a multiclass classification problem with K classes in a D dimensional space, requires learning $K(K-1)/D$ classifiers. can be done by minimizing the binary cross entropy loss*

2. [5*pts*] *We consider a $d = 2$ dimensional dataset with 2 pairs $\{\boldsymbol{x}_i, t_i\}_{i=1}^2$, i.e. $\boldsymbol{x}_i = (x_{i1}, x_{i2}) \in \mathbb{R}^2$. We assume that $x_{i1} = x_{i2}$ for $i = 1, 2$ as well as $t_1 + t_2 = 0$ and $\sum_{i=1}^2 x_{i1} = \sum_{i=1}^2 x_{i2} = 0$ so that the bias $\beta_0 = 0$. Answer the following questions*

   a) *Write the ridge regression optimization problem in this setting [1pt]*

   b) *Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$ [1pt]*

   c) *Write down the LASSO optimisation problem in this setting [1pt]*

   d) *Argue that in this setting, the coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique - In other words, there are many possible solutions to the optimisation problem in 2c. Describe those solutions. [2pts]*

3. [3pts] *We consider the dataset shown in Fig. 1. Draw on top of this dataset the least squares classifier and the logistic regression classifier. Briefly motivate your answer.*

4. [3pts] *We collect data for a group of students in a machine learning class with variables* x_1 = "number of hours studied", x_2 = "undergrad GPA" *and* t = "receives an A". *We fit a logistic regression model to the data and produce estimated coefficients* $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

   (a) *Estimate the probability that a student who studies for 40h and has an undergrad GPA of 3.5 gets an 'A' in the class.*

   (b) *How many hours would the student in part (a) need to study to have a* 50% *chance of getting an 'A' in the class?*

Question 2 (13pts)

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

True / False	*The training of a neural network does not depend on the initialization*
True / False	*If all the activations in a neural network are set to the identity, the model collapses to a linear model in the inputs*
True / False	*Neural networks with step activations can efficiently be trained through backpropagation*
True / False	*The sigmoid function is often used as the output activation when neural networks are used for regression problems*
True / False	*When training a neural network, a training epoch refers to one sweep through the entire dataset*
True / False	*Reducing the number of units in a neural network will increase the bias*
True / False	*Neural networks are parametric models*

2. [5pts] *Consider a neural network with two hidden layers:* $d = 2$ *dimensional inputs,* 2 *units in the first hidden layer,* 2 *units in the second hidden layer and a single output.*

   a) Draw a picture of the network

   b) Write out an expression for $y(x)$ *assuming ReLU activation functions. Be as explicit as possible.*

   c) How many parameters are there?

3. [3pts] *Consider the dataset given in table 1. Can this boolean function be represented by a single neuron with logistic activation function? If yes, give the value of the weights. If not motivate your answer with a short sentence.*

Question 3 (6pts) *We consider a set of training pairs* $\left\{t^{(i)}, \boldsymbol{x}^{(i)}\right\}$ *that satisfy the relation* $t^{(i)} = \beta_0 + \sum_{j=1}^{d} x_j^{(i)} \beta_j + \varepsilon^{(i)}$ *where* $\varepsilon^{(i)}$ *are independent and identically distributed from a* $N(0, \sigma^2)$ *distribution.*

1. [2pts] *Write the likelihood for the data.*

2. [2pts] *We first assume the following prior for* β: $\beta_1, \ldots, \beta_d$ *are independent and identically distributed according to a Laplace distribution with mean* 0 *and common scale parameter* b. *I.e.* $p(\beta) = \frac{1}{2b} \exp(-\|\beta\|_1 / b)$. *Write out the posterior* $p(\beta|t)$ *for* β *in this setting.*

3. [2pts] *Now assume the following prior for* β: $\beta_1, \ldots, \beta_d$ *are independent and identically distributed according to a normal distribution with mean zero and variance* c. *Write out the posterior* $p(\beta|t)$ *in this setting.*

x_1	x_2	$y(x_1, x_2)$
1	1	0
0	0	0
1	0	1
0	1	0

Table 1: Dataset used for Question 2

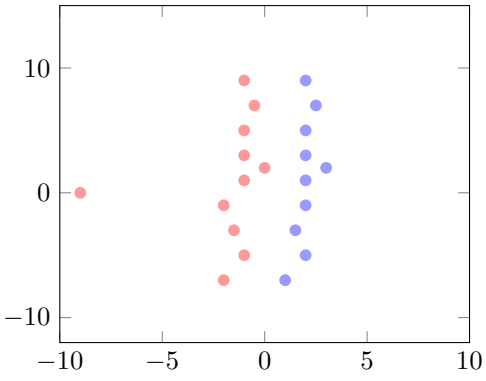

Figure 1: Training set for Question 1.