

CSCI-UA 9473 - Introduction to Machine Learning

Midterm III

Augustin Cosse

June 2022

Total: 31 points

Total time: 2h

General instructions: The exam consists of 2 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. [5pts] Indicate whether the following statements are true or false

True / False When the feature space is larger, overfitting is more likely

True / False Gradient descent applied to the least squares loss for the linear regression problem can get stuck at local minimas

True / False Logistic regression is a generative classifier

True / False Logistic regression is an example of a regression model

Suppose we have a dataset with five features explained below

x_1	GPA
x_2	IQ
x_3	Level (1 for college, 0 for High School)
x_4	Interaction between GPA and IQ
x_5	Interaction between GPA and Level

The target is “starting salary after graduation(in thousands of dollars)”. Suppose that we use a least squares approach to learn the model and got $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = .07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = .01$ and $\hat{\beta}_5 = -10$. Indicate whether the following are true or false

True / False For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates

True / False For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates

True / False For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough

True / False For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough

2. [5pts] Explain how the binary classifier can be extended to a multiclass classification problem. Give three possible extensions and provide the associated pseudo code

3. [6pts] Consider real valued variables x and t . The t variable is generated, conditional on x , from the following process

$$\begin{aligned}\varepsilon &\sim N(0, \sigma^2) \\ t &= \beta x + \varepsilon\end{aligned}$$

where every ε is an independent variable which is drawn from a Gaussian distribution with mean 0 and standard deviation σ . This is a one feature linear regression model, where β is the only weight parameter. The conditional probability distribution of t is given by

$$p(t|x, \beta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(t - \beta x)^2\right)$$

- (a) [2pts] Assume we have a training dataset of n pairs $(x^{(i)}, t^{(i)})$ for $i = 1, \dots, n$ and σ is known. Which of the following equations correctly represent the maximum likelihood problem for estimating β ? (Say yes or no to each possibility, keeping in mind that several of them might be right)

$$\operatorname{argmax}_{\beta} \sum_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(t^{(i)} - \beta x^{(i)})^2\right)$$

$$\operatorname{argmax}_{\beta} \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(t^{(i)} - \beta x^{(i)})^2\right)$$

$$\operatorname{argmax}_{\beta} \sum_{i=1}^n \exp\left(-\frac{1}{2\sigma^2}(t^{(i)} - \beta x^{(i)})^2\right)$$

$$\operatorname{argmax}_{\beta} \prod_{i=1}^n \exp\left(-\frac{1}{2\sigma^2}(t^{(i)} - \beta x^{(i)})^2\right)$$

$$\operatorname{argmax}_{\beta} \frac{1}{2} \sum_{i=1}^n (t^{(i)} - \beta x^{(i)})^2$$

$$\operatorname{argmin}_{\beta} \frac{1}{2} \sum_{i=1}^n (t^{(i)} - \beta x^{(i)})^2$$

- (b) [2pts] Derive the maximum likelihood estimator of the parameter β in terms of the training examples $t^{(i)}$ and $x^{(i)}$. (suggestion: start with the simplest form of the problem you found above and use the fact that the maximum/minimum can be found by setting the derivatives to zero)
- (c) [2pts] We now consider a prior on β . Assume that $\beta \sim N(0, \lambda^2)$ so that

$$p_{\lambda}(\beta) = \frac{1}{\sqrt{2\pi}\lambda} \exp\left(-\frac{1}{2\lambda^2}\beta^2\right)$$

We let β_{MLE} and β_{MAP} denote the Maximum Likelihood and Maximum A Posteriori estimators. Complete the table below

	$p_{\lambda}(\beta)$: wider/narrower/same ?	$ \beta_{MLE} - \beta_{MAP} $ increase/decrease?
As $\lambda \rightarrow \infty$		
As $\lambda \rightarrow 0$		

Question 2 (15pts)

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

- True / False A neural network with a single hidden layer (and one output unit) can learn the XOR dataset
- True / False A neural network with a single unit and a sigmoid activation is equivalent to a logistic regression classifier
- True / False Gradient descent applied to the least squares loss for the training of neural network can get stuck at local minimas
- True / False Increasing the number of units in the hidden layer of a one hidden layer neural network will increase the bias
- True / False Increasing the number of units in the hidden layer of a one hidden layer neural network will increase the variance

2. [4pts] Give the expression of the perceptron classifier and explain how this classifier can be trained through the perceptron learning rule

3. [6pts] We consider a two hidden layers neural network $y(\mathbf{x}; W)$, $\mathbf{x} \in \mathbb{R}^2$ with a final sigmoid activation (output unit). The first hidden layer consists of 3 units and the second hidden layer consists of 2 units. The weights from the first and second layers (including the intercepts) are respectively stored in the matrices $W_1 \in \mathbb{R}^{3 \times 3}$ and $W_2 \in \mathbb{R}^{2 \times 4}$. The weights associated to the output unit are stored in the vector $w_{\text{out}} \in \mathbb{R}^3$. All the hidden units have ReLU activations

(a) [2pts] Sketch the ReLU and sigmoid functions

(b) [2pts] Sketch the network

(c) [2pts] Give the detailed expression of $y(\mathbf{x}; W)$ as a function of \mathbf{x} , W_1 , W_2 and w_{out} .