Homework 1 - 1/23/23

CS 5787 Deep Learning Spring 2023

Your homework solution must be typed. Your submission must cite any references used (including articles, books, code, websites, and personal communications). All solutions must be written in your own words, and you must program the algorithms yourself. If you do work with others, you must list the people you worked with. Submit your solutions as a PDF to Canvas.

Problem 1 - Probability Review

Recall these rules from probability: Bayes' rule is

$P(B|A) = P(A|B)P(B).\ P(A)$

Joint probability's relationship with conditional probability is

$P(A \cap B) = P(A|B)\ P(B) = P(B|A)\ P(A)$.

If two events A and B are independent then their joint probability is $P(A \cap B) = P(A)\ P(B)$.

If two events A and B are not mutually exclusive then,

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

If two events A and B are mutually exclusive then $P(A \cap B) = 0$.

Hint: None of the questions in problem 1 have the same answer.

In this question, assume that the kangaroos have equal probability of having children that are male or female. Kangaroos almost always give birth to a single offspring.

Let $P(O = F)$ be the probability of the older offspring being female and let $P(Y = F)$ be the probability of the younger offspring being female.

Part 1 (3 points)

Suppose a kangaroo has two children. What is the probability that both are female? Show your derivation.

Answer:

Part 2 (3 points)

Suppose a kangaroo has two children and the oldest is female. What is the probability that both are female? Show your derivation.

Answer:

Part 3 (3 points)

Suppose a kangaroo has two children and at least one of them is female. What is the probability that both are female? Show your derivation.

Answer:

Problem 2 - ML Review (13 points)

1. Are there differences between "Machine Learning algorithm" and "classification algorithm" and if so, what are they? (1 point)
2. Explain linear regression and logistic regression, their differences, and when you would apply each. (3 points)
3. Pick a regression loss function, explain how it works, and give an example. What are the advantages/disadvantages of the function you picked? (3 points)
4. Read about Gradient Descent and explain it using a real life example (snowboarding?) (3 points)

5. You're working on solving a problem where you can use AI to make binary decisions given a set of inputs. How would you compute the errors your system might make, and how would you decide when your system is good enough to be deployed? (3 points)

Problem 3 - Setting Up Python

Part 0 - Installing Anaconda and PyTorch

Most assignments can be completed either by running and testing code on your local machine (see “Local installation”) or executing the code on the cloud through Google Colab. If you have time, you might want to try both options (running code locally and on Colab) so that you gain experience and figure out which workflow suits you best.

Local installation:

Python 3.6+, numpy, scipy, and matplotlib are necessary for the homeworks in this class. You can obtain all of these packages in one go by installing Anaconda. For your text editor, we recommend VSCode, although if you have another preference you’re welcome to use any editor you’d like.

After installing your Python 3.6+ environment along with the other toolboxes, you may optionally install PyTorch, but it will not be required until Homework 2.

Google Colab:

Google Colab is a free hosted version of Jupyter Notebook. Unlike Python files which are executed as an entire script, notebooks are executed in blocks of code called cells. To get started, go to the Colab homepage at https://colab.research.google.com/. Create a new notebook. Type in a line of Python code and run it (and start the notebook) by pressing Shift + Enter.

Part 1 - Visualizing CIFAR-10 (4 points)

The CIFAR-10 dataset contains 60,000 RGB images from 10 categories. It can be found here: https://www.cs.toronto.edu/~kriz/cifar.html. It can alternatively be loaded via HuggingFace datasets (make sure to `pip install datasets` before importing).

Using the first CIFAR-10 training batch file, display the first three images from each of the 10 categories as a 3 × 10 image array. The images are stored as rows, and you will need to reshape them into 32 × 32 × 3 images if you load up the raw data yourself. It is okay to use the PyTorch toolbox for loading them or you can roll your own.

Image and Code:

CODE HERE

Part 2 - Playing with NumPy (4 points)

Write a function called gaussian(n, m) that returns an n × m NumPy array, each entry of which is a random number drawn from the standard normal distribution (Gaussian with mean 0 variance 1). Generate a 10x10 grid of 2-D points using this function and then make a scatter plot of the points using Matplotlib.

Scatter Plot and Code:

CODE HERE