CS221 Autumn 2021: Artificial Intelligence: Principles and Techniques

Homework 1: Foundations

Name: Shaurya Goyal
Email ID: shaurya@kgpian.iitkgp.ac.in

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

Welcome to your first CS221 assignment! The goal of this assignment is to sharpen your math, programming, and ethical analysis skills needed for this class. If you meet the prerequisites, you should find these problems relatively innocuous. Some of these problems will occur again as subproblems of later homeworks, so make sure you know how to do them. If you're unsure about them or need a refresher, we recommend going through our prerequisites module or other resources on the Internet, or coming to office hours.

Before you get started, please read the Assignments section on the course website thoroughly.

Problem 1: Optimization and probability

In this class, we will cast a lot of AI problems as optimization problems, that is, finding the best solution in a rigorous mathematical sense. At the same time, we must be adroit at coping with uncertainty in the world, and for that, we appeal to tools from probability.

a. Let $x_1, \ldots, x_n$ be real numbers representing positions on a number line. Let $w_1, \ldots, w_n$ be positive real numbers representing the importance of each of these positions. Consider the quadratic function: $f(\theta) = \sum_{i=1}^{n} w_i(\theta - x_i)^2$. Note that θ here is a scalar. What value of θ minimizes $f(\theta)$? Show that the optimum you find is indeed a minimum. What problematic issues could arise if some of the w_i's are negative?
[**NOTE:** You can think about this problem as trying to find the point θ that's not too far away from the x_i's. Over time, hopefully you'll appreciate how nice quadratic functions are to minimize.]

[**What we expect:** An expression for the value of θ that minimizes $f(\theta)$ and how you got it. A short calculation/argument to show that it is a minimum. 1-2 sentences describing a problem that could arise if some of the w_i's are negative.]

Your Solution: The value of θ that minimizes $f(\theta)$ is $\theta^* = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

Steps:

(a) Take the 1st derivative and equate it to zero, $\frac{\mathrm{d}f(\theta)}{\mathrm{d}\theta} = \sum_{i=1}^{n} 2w_i(\theta - x_i) = 0$

(b) Rearrange and divide by 2 on both sides, $\sum_{i=1}^{n} w_i\theta = \sum_{i=1}^{n} w_i x_i$

(c) Since θ is a scalar, it can be taken out and rearranged to get the equation for θ^*

(d) Check for minima by taking 2nd derivative $\frac{\mathrm{d}^2 f(\theta)}{\mathrm{d}\theta^2} = \sum_{i=1}^{n} 2w_i$

(e) Since w_i is a positive real number for $i = 1 \ldots n$, $\frac{\mathrm{d}^2 f(\theta)}{\mathrm{d}\theta^2}$ is always > 0 which means θ^* is the minima.

If some weights w are negative, $\sum_{i=1}^{n} w_i$ may not be positive. Hence there will be no minima in such a case

b. In this class, there will be a lot of sums and maxes. Let's see what happens if we switch the order. Let $f(\mathbf{x}) = \max_{s \in [-1,1]} \sum_{i=1}^{d} s x_i$ and $g(\mathbf{x}) = \sum_{i=1}^{d} \max_{s_i \in [-1,1]} s_i x_i$, where $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ is a real vector and $[-1, 1]$ means the closed interval from -1 to 1. Which of $f(\mathbf{x}) \leq g(\mathbf{x})$, $f(\mathbf{x}) = g(\mathbf{x})$, or $f(\mathbf{x}) \geq g(\mathbf{x})$ is true for all $\mathbf{x}$? Prove it.
[**HINT:** You may find it helpful to refactor the expressions so that they are maximizing the same quantity over different sized sets.]

[**What we expect:** A short (3-5) line/sentence proof. You should use mathematical notation in your proof, but can also make your argument in words.]

Your Solution:

For $f(x)$,

(a) Take s outside the summation since it is a constant, $f(x) = \max s \sum_{i=1}^{d} x_i$

(b) Since x_i is a real number, $\sum x_i$ is also a real number

(c) Therefore it is equivalent to write the function as $f(x) = |\sum x_i|$ as s can be 1 or -1 depending on the sign of $\sum x_i$

Similarly $g(x)$ can be written as $\sum |x_i|$ as each individual term is being maximized.

Using the triangle inequality, we get $|\sum x_i| \leq \sum |x_i|$ or $f(x) \leq g(x)$

c. Suppose you repeatedly roll a fair six-sided die until you roll a 1 or a 2 (and then you stop). Every time you roll a 3, you lose a points, and every time you roll a 6, you win b points. You do not win or lose any points if you roll a 4 or a 5. What is the expected number of points (as a function of a and b) you will have when you stop?
[**HINT:** You will find it helpful to define a recurrence. If you define V as the expected number of points you get from playing the game, what happens if you roll a 3? You lose a points and then get to play again. What about the other cases? Can you write this as a recurrence?]

[**What we expect:** A recurrence to represent the problem and the resulting expression from solving the recurrence (no more than 1-2 lines).]

Your Solution: $V = E[\text{points}] = \frac{2}{6}0 + \frac{1}{6}(V - a) + \frac{1}{6}(V + b) + \frac{2}{6}V$

The terms in order correspond to rolling: 1 or 2 (game end), 3 (lose a), 6 (gain b), 4 or 5 (no change in points, continue rolling)

Solving the recurrence gives the expected number of points as $V = \frac{b - a}{2}$

d. Suppose the probability of a coin turning up heads is p (where $0 < p < 1$), and we flip it 6 times and get $\{\mathrm{T}, \mathrm{H}, \mathrm{H}, \mathrm{H}, \mathrm{T}, \mathrm{H}\}$. We know the probability (likelihood) of obtaining this sequence is $L(p) = (1-p)ppp(1-p)p = p^4(1-p)^2$. What value of p maximizes $L(p)$? Prove/Show that this value of p maximizes $L(p)$. What is an intuitive interpretation of this value of p?
[**HINT:** Consider taking the derivative of $\log L(p)$. You can also directly take the derivative of $L(p)$, but it is cleaner and more natural to differentiate $\log L(p)$. You can verify for yourself that the value of p which maximizes $\log L(p)$ must also maximize $L(p)$ (you are not required to prove this in your solution).]

[**What we expect:** The value of p that maximizes $L(p)$ and the work/calculation used to solve for it. Note that you must prove/show that it is a maximum. A 1-sentence intuitive interpretation of the value of p.]

Your Solution:

$\log L(p) = 4\log(p) + 2\log(1-p)$

Taking 1st derivative $\dfrac{\mathrm{d}\log L(p)}{\mathrm{d}p} = \dfrac{4}{p} - \dfrac{2}{1-p}$

Equating to zero and solving gives $p = \dfrac{2}{3}$

Second derivative test gives $\dfrac{\mathrm{d}^2\log L(p)}{\mathrm{d}p^2} = \dfrac{-4}{p^2} - \dfrac{2}{(1-p)^2}$

Since $0 < p < 1$, we get $\frac{\mathrm{d}^2\log L(p)}{\mathrm{d}p^2} < 0$ and the value of p that maximizes the likelihood of the sequence

Since p corresponds to heads and the sequence had 4 heads on 6 tosses (aka $\frac{4}{6} = \frac{2}{3}$), and since each toss is independent, it makes some sense that p is $\frac{2}{3}$

e. Now for a little bit of practice manipulating conditional probabilities. Suppose that A and B are two events such that $P(A|B) = P(B|A)$. We also know that $P(A \cup B) = \frac{1}{2}$ and $P(A \cap B) > 0$. Prove that $P(A) > \frac{1}{4}$.

[**HINT:** Note that A and B are not necessarily mutually exclusive. Consider how we can relate $P(A \cup B)$ and $P(A \cap B)$.]

[**What we expect:** A short ($\sim$ 5 line) proof/derivation.]

Your Solution:

Since $P(A|B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{P(B \cap A)}{P(A)} = P(B|A)$, we get $P(A) = P(B)$

Rearranging and substituting values in $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ gives $P(A \cap B) = 2P(A) - P(A \cup B) = 2P(A) - \dfrac{1}{2}$

Because $P(A \cap B) > 0$, we get $2P(A) > \dfrac{1}{2}$ which simplifies to $P(A) > \dfrac{1}{4}$ $\square$

f. Let's practice taking gradients, which is a key operation for being able to optimize continuous functions. For $\mathbf{w} \in \mathbb{R}^d$ (represented as a column vector), and constants $\mathbf{a}_i, \mathbf{b}_j \in \mathbb{R}^d$ (also represented as column vectors), $\lambda \in \mathbb{R}$, and a positive integer n, define the scalar-valued function

$$f(\mathbf{w}) = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} (\mathbf{a}_i^\top \mathbf{w} - \mathbf{b}_j^\top \mathbf{w})^2 \right) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2,$$

where the vector is $\mathbf{w} = (w_1, \ldots, w_d)^\top$ and $\|\mathbf{w}\|_2 = \sqrt{\sum_{k=1}^{d} w_k^2} = \sqrt{\mathbf{w}^T \mathbf{w}}$ is known as the L_2 norm. Compute the gradient $\nabla f(\mathbf{w})$.
[**RECALL:** The gradient is a d-dimensional vector of the partial derivatives with respect to each w_i:

$$\nabla f(\mathbf{w}) = \left(\frac{\partial f(\mathbf{w})}{\partial w_1}, \ldots \frac{\partial f(\mathbf{w})}{\partial w_d} \right)^\top.$$

If you're not comfortable with vector calculus, first warm up by working out this problem using scalars in place of vectors and derivatives in place of gradients. Not everything for scalars goes through for vectors, but the two should at least be consistent with each other (when $d = 1$). Do not write out summations over dimensions, because that gets tedious.]

[**What we expect:** An expression for the gradient and the work used to derive it. ($\sim$ 5 lines). No need to expand out terms unnecessarily; try to write the final answer compactly.]

Your Solution:

Each term in the gradient $\forall k \in 1 \ldots n$ is,

$$\frac{\partial f(\mathbf{w})}{\partial w_k} = 2 \sum_{i=1}^{n} \sum_{j=1}^{n} (\boldsymbol{a_i}^\top \boldsymbol{w} - \boldsymbol{b_j}^\top \boldsymbol{w})(a_{ik} - b_{jk}) + \lambda w_k$$

Imagine stacking the above equation for $\forall k \in 1 \ldots n$ to get,

$$\frac{\partial f(\mathbf{w})}{\partial \mathbf{w}} = 2 \sum_{i=1}^{n} \sum_{j=1}^{n} (\boldsymbol{a_i}^\top \boldsymbol{w} - \boldsymbol{b_j}^\top \boldsymbol{w})(\mathbf{a_i} - \mathbf{b_j}) + \lambda \mathbf{w}$$

I solved this by roughly expanding the main term for n = 2 and writing it element wise rather than vector. Once I understood what 1 differentiated term was, I generalized and condensed it to vector notation again.

Problem 2: Complexity

When designing algorithms, it's useful to be able to do quick back-of-the-envelope calculations to see how much time or space an algorithm needs. Hopefully, you'll start to get more intuition for this by being exposed to different types of problems.

a. Suppose we have an $n \times n$ grid of points, where we'd like to place 4 arbitrary axis-aligned rectangles (i.e., the sides of the rectangle are parallel to the axes). Each corner of each rectangle must be one of the points in the grid, but otherwise there are no constraints on the location or size of the rectangles. For example, it is possible for all four corners of a single rectangle to be the same point (resulting in a rectangle of size 0) or for all 4 rectangles to be on top of each other. How many possible ways are there to place 4 rectangles on the grid? In general, we only care about asymptotic complexity, so give your answer in the form of $O(n^c)$ or $O(c^n)$ for some integer c.

[**NOTE:** It is unnecessary to consider whether order matters in this problem, since we are asking for asymptotic complexity. You are free to assume either in your solution, as it doesn't change the final answer.]

[**What we expect:** A big-O bound for the number of possible ways to place 4 rectangles and some simple explanation/reasoning for the answer ($\sim$ 2 sentences).]

Your Solution: A grid of $n \times n$ has $n+1$ parallel lines in each direction. Selecting any 2 (with repetition allowed as size 0 is possible) gives

$$(n+1)(n+1) \times (n+1)(n+1) \text{ possible choices for one rectangle}$$

This has complexity $O(n^4)$ as the leading term is n^4

Since there are 4 rectangles, the complexity increases to $O(n^{16})$

b. Suppose we have an $n \times 3n$ grid of points. We start in the upper-left corner (the point at position $(1, 1)$), and we would like to reach the point at the lower-right corner (the point at position $(n, 3n)$) by taking single steps down or to the right. Suppose we are provided with a function $c(i, j)$ that outputs the cost associated with position (i, j), and assume it takes constant time to compute for each position. Note that $c(i, j)$ can be negative. Define the cost of a path as the sum of $c(i, j)$ for all points (i, j) along the path, including both endpoints. Give an algorithm for computing the cost of the minimum-cost path from $(1, 1)$ to $(n, 3n)$ in the most efficient way (with the smallest big-O time complexity). What is the runtime (just give the big-O)?

[**What we expect:** A description of the algorithm for computing the cost of the minimum-cost path as efficiently as possible ($\sim$ 5 sentences). The big-O runtime and a short explanation of how it arises from the algorithm.]

Your Solution: Complexity is $O(n^2)$ as the algorithm calculates all $n \times 3n$ values only once

Algorithm 1 Minimum Cost Dynamic Programming

Require: FinalCost = zero array of dimension(n, 3n), Cost function c(i,j)
  FinalCost[0][0] $\leftarrow c(0, 0)$ ▷ Initialize starting state cost
  for $i = 1 : n - 1$ **do** ▷ Initialize first column of matrix
    FinalCost[i][0] $\leftarrow$ FinalCost[i-1][0] + $c(i, 0)$
  end for
  for $j = 1 : 3n - 1$ **do** ▷ Initialize first row of matrix
    FinalCost[0][j] $\leftarrow$ FinalCost[0][j-1] + $c(0, j)$
  end for
  for $i = 1 : n - 1$ **do** ▷ Calculate remaining values by using dynamic programming
    for $j = 1 : 3n - 1$ **do**
      FinalCost[i][j] $\leftarrow$ minimum(FinalCost[i-1][j] , FinalCost[i][j-1]) + $c(i, j)$
    end for
  end for

FinalCost[n-1][3n-1] is the minimum cost

Problem 3: Ethical Issue Spotting

One of the goals of this course is to teach you how to tackle real-world problems with tools from AI. But real-world problems have real-world consequences. Along with technical skills, an important skill every practitioner of AI needs to develop is an awareness of the ethical issues associated with AI. The purpose of this exercise is to practice spotting potential ethical concerns in applications of AI - even seemingly innocuous ones.

In this question, you will explore the ethics of four different real-world scenarios using the ethics guidelines produced by a machine learning research venue, the NeurIPS conference. The NeurIPS Ethical Guidelines list sixteen non-exhaustive concerns under Potential Negative Social Impacts and General Ethical Conduct (the numbered lists). For each scenario, you will write a potential negative impacts statement. To do so, you will first determine if the algorithm / dataset / technique could have a potential negative social impact or violate general ethical conduct (again, the sixteen numbered items taken from the NeurIPS Ethical Guidelines page). If the scenario does violate ethical conduct or has potential negative social impacts, list one concern it violates and justify why you think that concern applies to the scenario. If you do **not** think the scenario has an ethical concern, explain how you came to that decision. Unlike earlier problems in the homework there are many possible good answers. If you can justify your answer, then you should feel confident that you have answered the question well.

Each of the scenarios is drawn from a real AI research paper. The ethics of AI research closely mirror the potential real-world consequences of deploying AI, and the lessons you'll draw from this exercise will certainly be applicable to deploying AI at scale. As a note, you are **not** required to read the original papers, but we have linked to them in case they might be useful. Furthermore, you are welcome to respond to anything in the linked article that's not mentioned in the written scenario, but the scenarios as described here should provide enough detail to find at least one concern.

[**What we expect:** A 2-5 sentence paragraph for each of the scenarios where you either A. identify at least one ethical concern from the NeurIPS Ethical Guidelines and justify why you think it applies, or B. state that you don't think a concern exists and justify why that's the case. Chosen scenarios may have anywhere from zero to multiple concerns that match, but you are only required to pick one concern (if it exists) and justify your decision accordingly. Furthermore, copy out and underline the ethical checklist item to which you are referring as part of your answer (i.e.: Severely damage the environment). We have also included a citation in the example solution below, but you are not required to add citations to your response.]

Example Scenario

You work for a U.S. hospital that has recently implemented a new intervention program that enrolls at-risk patients in programs to help address their chronic medical issues proactively before the patients end up in the hospital. The intervention program automatically identifies at-risk patients by predicting patients' risk scores, which are measured in terms of healthcare costs. However, you notice that for a given risk score tier, the Black patients are considerably sicker when enrolled than white patients, even though their assigned illness risk score is identical. You manually re-assign patients' risk scores based on their current symptoms and notice that the percentage of Black patients who would be enrolled has increased from 17%

to over 45% [1].

Example Solution

This algorithm has likely encoded, contains, or potentially exacerbates bias against people of a certain rac since the algorithm predicts healthcare costs. Because access to medical care in the U.S. is unequal, Black patients tend to have lower healthcare costs than their white counterparts [2]. Thus the algorithm will incorrectly predict that they are at lower risk.

a. An investment firm develops a simple machine learning model to predict whether an individual is likely to default on a loan from a variety of factors, including location, age, credit score, and public record. After looking through their results, you find that the model predicts mainly based on location and that the model mainly accepts loans from urban centers and denies loans from rural applicants [3]. Furthermore, looking at the gender and ethnicity of the applicants, you find that the model has a significantly higher false positive rate for Black and male applicants than for other groups. In a false positive prediction, a model misclassifies someone who does not default as likely to default.

Your Solution:

Have a detrimental effect on people's livelihood or economic security and exacerbates bias against people of a certain race

Jobs in the rural area are generally farm based jobs which are highly dependent on weather, kind of crops, methods used etc. So there is a higher possibility of default due to factors that are out of a person's control. Whereas in urban area, income is from work that is much less dependent on such factors.

I do not know if the population of US is skewed according to race and rural/urban so I cannot comment on that aspect

b. Stylometry is a way of predicting the author of contested or anonymous text by analyzing the writing patterns in the anonymous text and other texts written by the potential authors. Recently, highly accurate machine learning algorithms have been developed for this task. While these models are typically used to analyze historical documents and literature, they could be used for deanonymizing a wide range of texts, including code [4].

Your Solution: Develop or extend harmful forms of surveillance

People who are writing sensitive articles can be found out and arrested in tense social climates. Or just a general invasion of privacy of someone who wants to share their writing but prefer to be anonymous

Deanonymizing code can lead to bias in reviews as the rank/position/fame/past papers of a person can bias a reviewer about the quality of work

c. A research group scraped millions of faces of celebrities off of Google images to develop facial recognition technology [5]. The celebrities did not give permission for their images to be used in the dataset and many of the images are copyrighted. For copyrighted photos, the dataset provides URL links to the original image along with bounding boxes for the face.

Your Solution:

Encode, contain, or potentially exacerbate bias against people of a certain gender, race etc

Celebrities are not an accurate representation of the diversity in the general population in ethnicity, gender, race etc. Celebrities are mostly photographed at events where they may use make up or editing post click editing. People dress differently than celebrities and also look different. A simple google search of the word "celebrity" or "celebrities" shows results where almost everyone is from Hollywood (which is known to have issues with race etc) which is not a representative of the diverse population in USA,, let alone the world.

Since the images are available in a public domain to anyone in the world, I think it should be okay to use it as long as you are not harming or violating anyone's rights.

d. Researchers have recently created a machine learning model that can predict plant species automatically directly from a single photo [6]. The model was trained using photos uploaded to the iNaturalist app by users who consented to use of their photos for research purposes, and the model is only used within the app to help users identify plants they might come across in the wild.

Your Solution: Since they are also using date and location data, this information can be exploited to form a history of a user or to find out locations that have more activity with a user who likes nature. This information could be used by advertisement firms etc. But the users consent to the use of their photos and the accompanying information so it should be okay.

Problem 4: Programming

In this problem, you will implement a bunch of short functions. The main purpose of this exercise is to familiarize yourself with Python, but as a bonus, the functions that you will implement will come in handy in subsequent homeworks.

Do not import any outside libraries (e.g. numpy). Only standard python libraries and/or the libraries imported in the starter code are allowed.

See `submission.py`. No written submission.

Submission

Submission is done on Gradescope.

Written: When submitting the written parts, make sure to select **all** the pages that contain part of your answer for that problem, or else you will not get credit. To double check after submission, you can click on each problem link on the right side and it should show the pages that are selected for that problem.

Programming: After you submit, the autograder will take a few minutes to run. Check back after it runs to make sure that your submission succeeded. If your autograder crashes, you will receive a 0 on the programming part of the assignment. Note: the only file to be submitted to Gradescope is `submission.py`.

More details can be found in the Submission section on the course website.

References

[1] Obermeyer et al. Dissecting racial bias in an algorithm used to manage the health of populations. 2019.

[2] Institue of Medicine of the National Academies. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. 2003.

[3] Imperial College London. Loan Default Prediction Dataset. 2014.

[4] Caliskan-Islam et. al. De-anonymizing programmers via code stylometry. 2015.

[5] Parkhi et al. VGG Face Dataset. 2015.

[6] iNaturalist. A new vision model. 2020.