Machine Learning
CAP 5610

Instructor: Dr. Mengxin Zheng
TA : Shenyang Liu
Grader: Sai Preetham Damireddy
Department of Computer Science
University of Central Florida (UCF)
Spring, 2025

January 07, 2025

Assignment #0
Background Test for Machine Learning

Submission Instructions

This assignment is due **Friday, January 17, 2025, by 11:59 pm**. Please submit your solutions via Canvas (https://webcourses.ucf.edu/). You should submit your assignment as a PDF file. Please do not include blurry scanned/photographed equations, as they are difficult for us to grade.

Late Submission Policy

The late submission policy for assignments will be as follows unless otherwise specified:

1. 75% credit within 0-48 hours after the submission deadline.
2. 50% credit within 48-96 hours after the submission deadline.
3. 0% credit 96 hours after the submission deadline.

Introduction

The goal of this homework is to help you refresh the mathematical and programming background needed to take this class. Although most students find the machine learning class to be very rewarding, it does assume that you have a basic familiarity with several types of math: calculus, matrix and vector algebra, and basic probability. You don't need to be an expert in all these areas, but you will need to be conversant in each and understand:

1. **Basic probability and statistics** (at the level of a first undergraduate course). For example, we assume you know how to find the mean and variance of a set of data, and that you understand basic notions such as conditional probabilities and Bayes rule. During the class, you might be asked to calculate the probability of a data set with respect to a given probability distribution.

2. **Basic calculus** (at the level of a first undergraduate course). For example, we rely on you being able to take derivatives. During the class we will sometimes calculate derivatives (gradients) of functions with several variables.
3. **Linear algebra** (at the level of a first undergraduate course). For example, we assume you know how to multiply vectors and matrices and that you understand matrix inversion.
4. **Basic Programming skills with Python** including data loading and simple analytics. Python will be heavily used for assignments, and we will NOT teach how to code with Python in the course.

For each of these topics, this homework provides basic tests. If you pass all the tests, you are in good shape to take the class. If you cannot pass all the background tests, then you may struggle and should expect to devote some extra time to fill in the necessary math background as the course introduces it.

Background Test [100 Points]

1 Vectors and Matrices [20 points]

Consider the matrix X and the vectors y and z below:

$$X = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix} \; y = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \; z = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

1. [**5 points**] What is the inner product of the vectors y and z? (this is also sometimes called the dot product, and is sometimes written $y^T z$).
2. [**5 points**] What is the product Xy?
3. [**5 points**] Is X invertible? If so, give the inverse; if not, explain why.
4. [**5 points**] What is the rank of X? Explain your answer.

2 Calculus [10 Points]

1. [**5 points**] If $y = x^3 + x - 5$, then what is the derivative of y with respect to x?
2. [**5 points**] If $f(x_1, x_2) = x_1 sin(x_2) e^{-x_1}$, what is the gradient $\nabla f(x)$ of f ? Recall that, $\nabla f(x) = \binom{\delta_{x_1} f}{\delta_{x_2} f}$.

3 Probability and Statistics [30 Points]

Consider a sample of data $S = \{1, 1, 0, 1, 0\}$ created by flipping a coin x five times, where 0 denotes that the coin turned up heads and 1 denotes that it turned up tails.

1. [**4 points**] What is the sample mean for this data?
2. [**4 points**] What is the sample variance for this data?
3. [**4 points**] What is the probability of observing this data, assuming it was generated by flipping a coin with an equal probability of heads and tails (i.e. the probability distribution is $p(x = 1) = 0.5$, $p(x = 0) = 0.5$).
4. [**4 points**] Note that the probability of this data sample would be greater if the value of $p(x = 1)$ was not 0.5, but instead some other value. What is the value that maximizes the probability of the sample S. Please justify your answer.

5. **[4 points]** Consider the following joint probability table over variables y and z, where y takes a value from the set {a,b,c}, and z takes a value from the set {T,F}:

	y		
z	**a**	**b**	**c**
T	0.2	0.1	0.2
F	0.05	0.15	0.3

Table 1: Caption

- What is $p(z = T \text{ AND } y = b)$?
- What is $p(z = T|y = b)$?

6. **[5 points]** State True or False. Here A^c denotes the complement of the event A.
   - $P(A \cup B) = P(A \cap (B \cap A^c))$
   - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
   - $P(A) = P(A \cap B) + P(A^c \cap B)$
   - $P(A|B) = P(B|A)$
   - $P(A_1 \cap A_2 \cap A_3) = P(A_3|(A_2 \cap A_1))P(A_2|A_1)P(A_1)$
7. **[5 points]** Match the distribution name to the formula of its probability density function.

Multivariate Gaussian	$p^x(1-p)^{1-x}$
Bernoulli	$\frac{1}{b-a}$, when $a \leq x \leq b$; 0 otherwise
Uniform	$\binom{n}{x}p^x(1-p)^{n-x}$
Binomial	$\frac{1}{\sqrt{(2\pi)d\|\sum\|}} \exp\{(-\frac{1}{2} - (x-\mu)^T \sum^{-1}(x-\mu))\}$

Table 2: Caption

4 Linear Algebra [15 Points]

Vector Norms

[5 points] Draw the regions corresponding to vectors $x \in R^2$ with the following norms:

1. $||x||_2 \leq 1$ (Recall, $||x||_2 = \sqrt{\sum_i x_i^2}$)
2. $||x||_1 \leq 1$ (Recall, $||x||_1 = \sum_i |x_i|$)
3. $||x||_\infty \leq 1$ (Recall, $||x||_\infty = \max_i |x_i|$)

Geometry

1. **[5 points]** Show that, the vector w is orthogonal to the line $w^T x + b = 0$. (Hint: Consider two points x_1, x_2 that lie on the line. What is the inner product $w^T(x_1 - x_2)$?)
2. **[5 points]** Argue that the distance from the origin to the line $w^T x + b = 0$ is $\frac{b}{||w||}$.

5 Programming Skills [25 Points]

1. **[10 points] Sampling from a Distribution** Use the Python libraries Numpy and Matplotlib.

   - Draw 100 samples $x = [x_1 x_2]$ from a 2-dimensional Gaussian distribution with mean $[0, 0]$. and identity covariance matrix, i.e. $p(x) = \frac{1}{\sqrt{(2\pi)^d}} \exp(-\frac{||x||^2}{2})$. Plot them on a scatter plot (x_1 vs. x_2).
   - How does the scatter plot change if the mean is [1, 1]? (for the questions below, change the mean back to [0, 0])
   - How does the scatter plot change if you double the variance of each component (x_1 & x_2)?
   - How does the scatter plot change if the covariance matrix changes to the following?

   $$X = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

   - How does the scatter plot change if the covariance matrix changes to the following?

   $$X = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

2. **[15 points] Load and Analyze a Real-word Dataset using Python**

   In this task, you will load a real-world data set called Census-Income Data Set, available publicly for downloading at Dataset. This data set contains weighted census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. The data contains 41 demographic and employment-related variables. Basic statistics for this data set are provided below.

   - Number of instances data = 199523
   - Duplicate or conflicting instances: 46716
   - Number of instances in test = 99762
   - Duplicate or conflicting instances: 20936
   - Class probabilities for income-projected.test file
   - Probability for the label '- 50000' : 93.80
   - Probability for the label '50000+' : 6.20
   - Majority accuracy: 93.80% on value - 50000
   - Number of attributes = 40 (continuous : 7 nominal : 33)
   - Information about .data file :
     - 91 distinct values for attribute #0 (age) continuous
     - 9 distinct values for attribute #1 (class of worker) nominal
     - 52 distinct values for attribute #2 (detailed industry recode) nominal
     - 47 distinct values for attribute #3 (detailed occupation recode) nominal

  - 17 distinct values for attribute #4 (education) nominal
  - 1240 distinct values for attribute #5 (wage per hour) continuous
  - 3 distinct values for attribute #6 (enroll in edu inst last wk) nominal
  - 7 distinct values for attribute #7 (marital stat) nominal
  - 24 distinct values for attribute #8 (major industry code) nominal
  - 15 distinct values for attribute #9 (major occupation code) nominal
  - 5 distinct values for attribute #10 (race) nominal
  - 10 distinct values for attribute #11 (hispanic origin) nominal
  - 2 distinct values for attribute #12 (sex) nominal
  - 3 distinct values for attribute #13 (member of a labor union) nominal
  - 6 distinct values for attribute #14 (reason for unemployment) nominal
  - 8 distinct values for attribute #15 (full or part time employment stat) nominal
  - 132 distinct values for attribute #16 (capital gains) continuous
  - 113 distinct values for attribute #17 (capital losses) continuous
  - 1478 distinct values for attribute #18 (dividends from stocks) continuous
  - 6 distinct values for attribute #19 (tax filer stat) nominal
  - 6 distinct values for attribute #20 (region of previous residence) nominal
  - 51 distinct values for attribute #21 (state of previous residence) nominal
  - 38 distinct values for attribute #22 (detailed household and family stat) nominal
  - 8 distinct values for attribute #23 (detailed household summary in household) nominal
  - 10 distinct values for attribute #24 (migration code-change in msa) nominal
  - 9 distinct values for attribute #25 (migration code-change in reg) nominal
  - 10 distinct values for attribute #26 (migration code-move within reg) nominal
  - 3 distinct values for attribute #27 (live in this house 1 year ago) nominal
  - 4 distinct values for attribute #28 (migration prev res in sunbelt) nominal
  - 7 distinct values for attribute #29 (num persons worked for employer) continuous
  - 5 distinct values for attribute #30 (family members under 18) nominal
  - 43 distinct values for attribute #31 (country of birth father) nominal
  - 43 distinct values for attribute #32 (country of birth mother) nominal
  - 43 distinct values for attribute #33 (country of birth self) nominal
  - 5 distinct values for attribute #34 (citizenship) nominal
  - 3 distinct values for attribute #35 (own business or self employed) nominal
  - 3 distinct values for attribute #36 (fill inc questionnaire for veteran's admin) nominal
  - 3 distinct values for attribute #37 (veterans benefits) nominal
  - 53 distinct values for attribute #38 (weeks worked in year) continuous
  - 2 distinct values for attribute #39 (year) nominal
- classes: - 50000, 50000+.

One instance per line with comma-delimited fields. There are 199,523 instances in the data file and 99,762 in the test file.

The data was split into train/test in approximately $\frac{2}{3}$, $\frac{1}{3}$ proportions using MineSet's MIndUtil mineset-to-mlc. Below are your tasks:

(a) Based on the training data, how many people have an income of more than $50K$ per year?

(b) Based on the testing data, how many people have an income of more than $50K$ per year?

(c) Based on the testing data, how many people are Asian or Pacific Islander?

(d) Based on the training data, what is the average age of people with more than $50K$ income per year?

(e) Based on the testing data, what is the average age of people with more than $50K$ income per year?

Disclaimers: This assignment re-uses some materials from the publicly available website: CMU Introduction to Machine Learning Course, 10-315, Spring 2019. I personally thank Prof. Maria-Florina Balcan for sharing her teaching materials publicly. This assignment is exclusively used for instructional purposes.