# Homework 0
# Fall 2023 - CSCI 5622 - Machine Learning

Instructor:
Daniel Acuna
Associate Professor
Student:
Luna McBride
University of Colorado at Boulder

August 31, 2023

The programming question is worth 1 point and the rest of the questions are equally divided in 1 point.

- Kindly give brief responses with justifications for every query. When presenting experimental findings,use tables or figures efficiently to present the data. Ensure that all descriptions, tables, and illustrations related to a specific section are collectively placed.
- Upload your code as well as one Jupyter notebook file

Incomplete items marked with –INC . I spent far too long on this to keep worrying about this for what is just 1% of my grade and I have another assignment to get done for another class.

## 1 Probability and statistics

1. Consider a class with $n$ students. Assume that each of them comes randomly and independently from one of the 150 countries in the world (simplified) Assuming each person can come from one country (ie country of birth) rather than family lineage.

   (a) What is the probability that in a class with 20 students, at least two of them come from the samecountry? The probability of one student coming from a specified country is entirely independent.

   This is essentially the birthday problem, which is very well explained here: https://medium.com/i-math/the-birthday-problem-307f31a9ac6f (super helpful when you have to try to remember the logic backward when memories of actual probability class are choppy)

   Permutation is used because the order does matter. (oct 31, jul 4) is a very different lineup than (oct 31, dec 25). Since the regular equation the probability of not sharing a country, it needs to be reciprocated by the 1- at the front. All the denominators are the same, so it is just multiplied as many times as there students.

   $1 - \frac{Permutation(k,n)}{k^n}$

1 - $\frac{\frac{150!}{(150-20)!}}{150^{20}}$

1 - $\frac{150!}{(150-20)!150^{20}}$ <- (into a calculator, because there is no way I could calculate this by hand in any reasonable amount of time).

Probability = 0.73

(b) --INC How large does the class have to be so that there is 95% chance that two or more students comefrom the same country?

0.95 = 1 - $\frac{150!}{(150-n)!150^{n}}$

0.05 = $\frac{150!}{(150-n)!150^{n}}$

0.05((150-n)!$150^{n}$) = 150!

(150-n)! $150^{n}$ = 150!/0.05 --INC

2. For the following joint discrete PMF $p_{X,Y}(x,y)$

| | x=1 | x=2 | x=3 | x=4 |
|---|---|---|---|---|
| y=1 | 1/20 | 2/20 | 2/20 | 0 |
| y=2 | 2/20 | 4/20 | 1/20 | 2/20 |
| y=3 | 0 | 1/20 | 3/20 | 1/20 |
| y=4 | 0 | 1/20 | 0 | 0 |

(a) Show that $X$ and $Y$ are not independent by showing that for at least combination of $x$ and $y$, the rvs are not conditionally independent.

Rule for independence: p(x,y) = p(x)p(y)

p(x=1) = 3/20

p(y=1) = 5/20

(3/20)*(5/20) = 0.0375

P(1,1) = 1/20 = 0.05

0.0375 != 0.05

Therefore, x and y are not independent.

3. For the following continuous random variable with probability density function

$$f_X(x) = \begin{cases} \frac{1}{10} & 0 \le x \le 5 \\ c & 6 \le x \le 7 \\ 0 & otherwise \end{cases}$$

(a) What is the value of the constant $c$?

Sum of probability = 1

0 to 5 is six occurrences, so 6 * $\frac{1}{10}$ = $\frac{6}{10}$

Probability past x=7 is 0, so sum $0 \leq x \leq 7 = 1$

1 - $\frac{6}{10}$ = $\frac{4}{10}$

This $\frac{4}{10}$ is split between two occurrences (x=6,

x=7)

Therefore, c = $\frac{2}{10}$ = $\frac{1}{5}$

4. A continuous random variable $X$ is uniformly distributed between −2 and 2. There is a continuous noise process $N$ uniformly distributed between −1 and 1 that gets added to $X$ and we only observe the random variable $Y = X + N$. What is $E[Y]$?

   Equations:

   E[X] = $\int_n^m xf(x)dx$ given in the lecture, changing out ∞ for n and m, representing arbitrary upper and lower bounds (for my own brain's sake).

   E[X+N] = E[X] + E[N] as per the Linearity of Expectation Theorem found here: https://dlsun.github.io/probability/ev-joint-continuous.html

   Power rule of integrals: $\int x^n dx = \frac{x^{n+1}}{n+1}$ from https://www.mathbootcamps.com/power-rule-integrals/

   The noise only occurs between -1 and 1, so -2 to -1 and 1 to 2 is the expectation of X. Noise becomes 0 otherwise. This can thus be split into three intervals.

   $E[X]_{-2}^{-1}$ + $E[X+N]_{-1}^{1}$ + $E[X]_{1}^{2}$

   $E[X]_{-2}^{-1}$ + $E[X]_{-1}^{1}$ + $E[N]_{-1}^{1}$ + $E[X]_{1}^{2}$

   $\int_{-2}^{-1} X(X)dx + (\int_{-1}^{1} X(X)dx + \int_{-1}^{1} N(N)dn) + \int_{1}^{2} X(X)dx$

   $\int_{-2}^{-1} X^2 dx + \int_{-1}^{1} X^2 dx + \int_{-1}^{1} N^2 dn + \int_{1}^{2} X^2 dx$

   $(\frac{X^3}{3})_{-2}^{-1} + (\frac{X^3}{3})_{-1}^{1} + (\frac{N^3}{3})_{-1}^{1} + (\frac{X^3}{3})_{1}^{2}$ (+c)

   (-1/3 – (-8/3)) + ((1/3 – (-1/3)) + (1/3 – (-1/3))) + (8/3 – 1/3)

   7/3 + 2/3 + 2/3 +7/3 = 18/3 = 6

5. For the random variables $X$ and $Y$, the covariance is defined as $Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$.

(a) If $E[Y|X=x] = x$ show that $Cov(X,Y) = Var(X) = E[(X - E[X])^2]$

Both a and b using https://www.statlect.com/fundamentals-of-probability/conditional-expectation and https://www.statlect.com/fundamentals-of-probability/expected-value-properties

Assuming x (small x, not to be confused with big X X in E[X]) is a constant, and E[constant] = constant

$Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$

$Cov(X,Y) = E[XY - E[X]Y - E[Y]X + E[X]E[Y]]$

$Cov(X,Y) = E[X]E[Y] - E[E[X]]E[Y] - E[E[Y]]E[X] + E[E[X]]E[E[Y]]$

$Cov(X,Y) = E[X]E[E[Y|X]] - E[E[X]]E[E[Y|X]] - E[E[E[Y|X]]]E[X] + E[E[X]]E[E[E[Y|X]]]$

$Cov(X,Y) = E[X]E[X] - E[E[X]]E[X] - E[E[X]]E[X] + E[E[X]]E[E[X]]$

$Cov(X,Y) = E[XX] - E[E[X]]E[X] - E[E[X]]E[X] + E[E[X]E[X]]$

$Cov(X,Y)$ =E[X^2 – E[X]X – E[X]X + E[X]^2]

$Cov(X,Y) = E[(X - E[X])^2]$

(b) (b) If $X$ and $Y$ are independent, show that $Cov(X,Y) = 0$.

*Cov(X,Y) = $E[(X - E[X])(Y - E[Y])]$*

*Cov(X,Y) = $E[XY - E[X]Y - E[Y]X + E[X]E[Y]]$*

*Cov(X,Y) = E[XY] – E[E[X]Y] – E[E[Y]X] + E[E[X]E[Y]]*

*Cov(X,Y) = E[X]E[Y] – E[E[X]]E[Y] – E[E[Y]]E[X] + E[E[X]]E[E[Y]]*

---

*E[E[X]] = E[X], E[E[Y]] = E[Y] with conditional expectation, as no |X or |Y is required given independence*

---

*Cov(X,Y) = E[X]E[Y] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y]*

*Cov(X,Y) = 2 E[X]E[Y] - 2 E[X]E[Y]*

*Cov(X,Y) = 0*

Therefore, Cov(X,Y) = 0

6. Imagine we independently ask $n$ people about their vote for a candidate. Let $p$ be the proportion of people voting for a candidate, which is unknown. Assume that each person $X_i$'s answer is a Bernoulli trial with probability $p$ which we know has variance $p(1-p)$. Use $M_n = \frac{\sum x_i}{n}$ as the average of their answers

   (a) Show that $E[M_n] = p$ and show that $var(M_n) = p(1-p)/n$ .

   a. $E[M_n] = \sum M_n f(M_n)$
   $E[M_n] = \sum \frac{\sum x_i}{n} f(M_n)$
   $E[M_n] = \sum \frac{p}{n} f(M_n)$
   $E[M_n] = \frac{p}{n} \sum f(M_n)$
   $E[M_n] = \frac{p}{n} n$
   $E[M_n] = p$

   b. --INC $Var(M_n) = E[(M_n - E[M_n])^2]$
   $Var(M_n) = E[(\frac{\sum x_i}{n} - p)^2]$
   $Var(M_n) = E[(\frac{p}{n} - p)^2]$
   $Var(M_n) = E[(\frac{p}{n})^2 - 2(p^2/n) + p^2)]$
   $Var(M_n) = E[p^2(\frac{1}{n^2} - 2/n + 1)]$
   $Var(M_n) = E[p^2(\frac{1}{n} - 1)^2]$ --INC

   (b) --INC Use a) to write the Chebyshev inequality governing the probability that the absolute differencebetween the real $p$ and our estimate $M_n$ is more than $\epsilon$ away. --INC

# 2 Linear algebra and calculus

It has been over half a decade since I have done Linear Algebra (Fall 2017) or Calculus (Spring 2017)

1. Let $A$=

| | | |
|---|---|---|
| *0* | *2* | *4* |
| *2* | *4* | *2* |
| *3* | *3* | *1* |

$b$ = 2 2 4 $^T$, and $c$ = 1 −1 1

   (a) What is $Ac$? Ac cannot be multiplied. A The number of columns of the first matrix/vector must be equal to the number or rows in the second. 3X**3 1**X3. Those are not the same, so it cannot be done.

(b) What is the solution to the linear system $Ax = b$? Following the rules presented at https://www.dummies.com/article/academics-the-arts/math/pre-calculus/how-to-use-gaussian-elimination-to-solve-systems-of-equations-167828/

21.

b

$$\begin{bmatrix} 0 & 2 & 4 \\ 2 & 4 & 2 \\ 3 & 3 & 1 \end{bmatrix} \bullet \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix}$$

$$\left[\begin{array}{ccc|c} 0 & 2 & 4 & 2 \\ 2 & 4 & 2 & 2 \\ 3 & 3 & 1 & 4 \end{array}\right] \xrightarrow{R_1 - 2R_2} \left[\begin{array}{ccc|c} -4 & -6 & 0 & -2 \\ 2 & 4 & 2 & 2 \\ 3 & 3 & 1 & 4 \end{array}\right]$$

$0-2(2)=-4$
$2-2(4)=-8$
$4-2(2)=0$
$2-2(2)=-2$

$$\left[\begin{array}{ccc|c} -4 & -6 & 0 & -2 \\ 2 & 4 & 2 & 2 \\ 3 & 3 & 1 & 4 \end{array}\right] \xrightarrow{-\frac{1}{2}R_1} \left[\begin{array}{ccc|c} 2 & 3 & 0 & 1 \\ 2 & 4 & 2 & 2 \\ 3 & 3 & 1 & 4 \end{array}\right]$$

$-\frac{-4}{-2}=2$
$-\frac{-6}{-2}=3$
$-\frac{0}{-2}=0$
$-\frac{-2}{-2}=1$

$$\left[\begin{array}{ccc|c} 2 & 3 & 0 & 1 \\ 2 & 4 & 2 & 2 \\ 3 & 3 & 1 & 4 \end{array}\right] \xrightarrow{R_2 - R_1} \left[\begin{array}{ccc|c} 2 & 3 & 0 & 1 \\ 2 & 4 & 2 & 2 \\ 1 & 0 & 1 & 3 \end{array}\right]$$

$3-2=1$
$3-3=0$
$1-0=1$
$4-1=3$

$$\left[\begin{array}{ccc|c} 2 & 3 & 0 & 1 \\ 2 & 4 & 2 & 2 \\ 1 & 0 & 1 & 3 \end{array}\right] \xrightarrow{R_2 - 2R_3} \left[\begin{array}{ccc|c} 2 & 3 & 0 & 1 \\ 0 & 4 & 0 & -4 \\ 1 & 0 & 1 & 3 \end{array}\right]$$

$2-2(1)=0$
$4-2(0)=4$
$2-2(1)=0$
$2-2(3)=-4$

$$\left[\begin{array}{ccc|c} 2 & 3 & 0 & 1 \\ 0 & 4 & 0 & -4 \\ 1 & 0 & 1 & 3 \end{array}\right] \xrightarrow{\frac{1}{4}R_2} \left[\begin{array}{ccc|c} 2 & 3 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & 1 & 3 \end{array}\right]$$

$\frac{0}{4}=0$
$\frac{4}{4}=1$
$\frac{0}{4}=0$
$\frac{-4}{4}=-1$

$$\left[\begin{array}{ccc|c} 2 & 3 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & 1 & 3 \end{array}\right] \xrightarrow{R_1 - 3R_2} \left[\begin{array}{ccc|c} 2 & 0 & 0 & 4 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & 1 & 3 \end{array}\right]$$

$2-3(0)=2$
$3-3(1)=0$
$0-3(0)=0$
$1-3(-1)=4$

$$\left[\begin{array}{ccc|c} 2 & 0 & 0 & 4 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & 1 & 3 \end{array}\right] \xrightarrow{\frac{1}{2}R_1} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & 1 & 3 \end{array}\right]$$

$\frac{2}{2}=1$
$\frac{0}{2}=0$
$\frac{0}{2}=0$
$\frac{4}{2}=2$

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & 1 & 3 \end{array}\right] \xrightarrow{R_3 - R_1} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{array}\right]$$

$1-1=0$
$0-0=0$
$1-0=1$
$3-2=1$

$$\boxed{\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}}$$

2. For the following equation with matrices that all compatible and invertible

$$AXA + C = Y$$

(a) Solve for $X$ in terms of the other matrices $A$, $C$, and $Y$.

Going off rules found at https://byjus.com/maths/algebra-of-matrices/

$$AXA = Y - C$$
$$A(XA) = Y - C \text{ <- Associative Law of Multiplication}$$
$$A^{-1}A(XA) = A^{-1}(Y - C)$$
$$XA = A^{-1}(Y - C)$$
$$X(AA^{-1}) = A^{-1}(Y - C)A^{-1}$$

$$\underline{\boldsymbol{X = A^{-1}(Y - C)A^{-1}}}$$

3. For the following set of equations

$$L(p) = (y - p)^2$$
$$p(z) = \frac{1}{1 + e^{-z}}$$
$$z = b_0 + b_1 x$$

(a) Compute the derivative $\frac{dL}{db_1}$

Step 1: combine to similar terms (made big so it can actually be readable)

$$p(x) = \frac{1}{1+e^{-(b_0+b_1x)}}$$

$$L(x) = \left(y - \frac{1}{1+e^{-(b_0+b_1x)}}\right)^2$$

Step 2: Set up the equation in respect to L and $b_1$ (using the equation on the top answer of https://math.stackexchange.com/questions/340744/what-do-the-symbols-d-dx-and-dy-dx-mean)

$$\frac{dy}{dx} = \lim_{h \to 0}\left(\frac{f(x+h) - f(x)}{h}\right)$$

$$\frac{dL}{db_1} = \lim_{h\to 0}\left(\frac{\left(y-\frac{1}{1+e^{-(b_0+(b_1+h)x)}}\right)^2-\left(y-\frac{1}{1+e^{-(b_0+b_1x)}}\right)^2}{h}\right)$$

$$\frac{dL}{db_1} = \lim_{h\to 0}\left(\frac{\left(y-\frac{1}{1+e^{-(b_0+b_1x+hx)}}\right)^2-\left(y-\frac{1}{1+e^{-(b_0+b_1x)}}\right)^2}{h}\right)$$

---

$$\frac{dL}{db_1} = \lim_{h\to 0}\left(\frac{y^2-\frac{2y}{1+e^{-(b_0+b_1x+hx)}}+\left(\frac{1}{1+e^{-(b_0+b_1x+hx)}}\right)^2-y^2+\frac{2y}{1+e^{-(b_0+b_1x)}}-\left(\frac{1}{1+e^{-(b_0+b_1x)}}\right)^2}{h}\right)$$

---

$$\frac{dL}{db_1} = \lim_{h\to 0}\left(\frac{y^2-\frac{2y}{1+e^{-(b_0+b_1x+hx)}}+\frac{1}{(1+e^{-(b_0+b_1x+hx)})^2}-y^2+\frac{2y}{1+e^{-(b_0+b_1x)}}-\frac{1}{(1+e^{-(b_0+b_1x)})^2}}{h}\right)$$

---

$$\frac{dL}{db_1} = \lim_{h\to 0}\left(\frac{-\frac{2y}{1+e^{-(b_0+b_1x+hx)}}+\frac{1}{(1+e^{-(b_0+b_1x+hx)})^2}+\frac{2y}{1+e^{-(b_0+b_1x)}}-\frac{1}{(1+e^{-(b_0+b_1x)})^2}}{h}\right)$$

---

$\frac{dL}{db_1}$ = (formatting change to make readable)

$$\lim_{h\to 0}\left(\frac{-\frac{2y}{1+e^{-(b_0+b_1x+hx)}}+\frac{1}{1+2e^{-(b_0+b_1x+hx)}+e^{-2(b_0+b_1x+hx)}}}{h}+\frac{\frac{2y}{1+e^{-(b_0+b_1x)}}-\frac{1}{1+2e^{-(b_0+b_1x)}+e^{-2(b_0+b_1x)}}}{h}\right)$$

---

$$\frac{dL}{db_1}=\lim_{h\to 0}\left(\frac{-\frac{2y}{1+e^{(-b_0-b_1x-hx)}}+\frac{1}{1+2e^{(-b_0-b_1x-hx)}+e^{(-2b_0-2b_1x-2hx)}}}{h}+\frac{\frac{2y}{1+e^{(-b_0-b_1x)}}-\frac{1}{1+2e^{(-b_0-b_1x)}+e^{(-2b_0-2b_1x)}}}{h}\right)$$

---

$\frac{dL}{db_1}$ = (figuring out what the pieces are for the substitution)

$$\lim_{h\to 0}\left(\frac{-\frac{2y}{1+e^{(-b_0-b_1x-hx)}}+\frac{1}{1+e^{(-b_0-b_1x-hx)}(2+e^{(-b_0-b_1x-hx)})}}{h}+\frac{\frac{2y}{1+e^{(-b_0-b_1x)}}-\frac{1}{1+e^{(-b_0-b_1x)}(2+e^{(-b_0-b_1x)})}}{h}\right)$$

---

Sub n for $e^{(-b_0-b_1x)}$ and m for $e^{(-b_0-b_1x-hx)}$

$$\frac{dL}{db_1} = \lim_{h\to 0}\left(\frac{-\frac{2y}{1+m}+\frac{1}{1+m(2+m)}}{h} + \frac{\frac{2y}{1+n}-\frac{1}{1+n(2+n)}}{h}\right) \text{ -> } \lim_{h\to 0}\left(\frac{-\frac{2y}{1+m}+\frac{1}{1+2m+m^2}}{h} + \frac{\frac{2y}{1+n}-\frac{1}{1+2n+n^2}}{h}\right)$$

---

$$\frac{dL}{db_1} = \lim_{h\to 0}\left(\frac{-\frac{2y}{1+m}+\frac{1}{(m+1)(m+1)}}{h} + \frac{\frac{2y}{1+n}-\frac{1}{(n+1)(n+1)}}{h}\right)$$

---

$$\frac{dL}{db_1} = \lim_{h\to 0}\left(\frac{\frac{1}{m+1}(-2y+\frac{1}{(m+1)})}{h} + \frac{\frac{1}{n+1}(2y-\frac{1}{(n+1)})}{h}\right)$$

---

$m = e^{-xh} n$

$$\frac{dL}{db_1} =$$

$$\lim_{h\to 0} \left(\frac{\frac{1}{e^{-hx}{}_{n+1}}\left(-2y+\frac{1}{\left(e^{-hx}{}_{n+1}\right)}\right)}{h} + \frac{\frac{1}{n+1}\left(2y-\frac{1}{(n+1)}\right)}{h}\right) \rightarrow \lim_{h\to 0} \left(\frac{\frac{e^{hx}}{n+1}\left(-2y+\frac{e^{hx}}{(n+1)}\right)}{h} + \frac{\frac{1}{n+1}\left(2y-\frac{1}{(n+1)}\right)}{h}\right)$$

---

$$\frac{dL}{db_1} =$$

$$\lim_{h\to 0} \frac{\frac{1}{n+1}\left(e^{hx}\left(-2y+\frac{e^{hx}}{(n+1)}\right)+2y-\frac{1}{(n+1)}\right)}{h} \rightarrow$$

$$\lim_{h\to 0} \frac{\frac{1}{n+1}\left(-2ye^{hx}+\frac{e^{2hx}}{(n+1)}+2y-\frac{1}{(n+1)}\right)}{h} \rightarrow$$

$$\lim_{h\to 0} \frac{\frac{1}{n+1}\left(-2ye^{hx}+2y+\frac{e^{2hx}-1}{n+1}\right)}{h} \rightarrow \frac{1}{n+1}\lim_{h\to 0} \frac{-2ye^{hx}+2y+\frac{e^{2hx}-1}{n+1}}{h}$$

---

$$\frac{dL}{db_1} = \frac{1}{n+1}\lim_{h\to 0} \frac{-2ye^{hx}+2y}{h} + \frac{\frac{e^{2hx}-1}{n+1}}{h} \rightarrow$$

$$\frac{1}{n+1}\lim_{h\to 0} \frac{-2ye^{hx}+2y}{h} + \frac{1}{n+1}\lim_{h\to 0} \frac{\frac{e^{2hx}-1}{n+1}}{h}$$

---

$$\frac{dL}{db_1} = \frac{1}{n+1}\lim_{h\to 0} \frac{-2y(e^{hx}-1)}{h} + \frac{1}{n+1}\lim_{h\to 0} \frac{\frac{e^{2hx}-1}{n+1}}{h} \rightarrow$$

$$\frac{-2y}{n+1}\lim_{h\to 0}\frac{e^{hx}-1}{h}+\frac{1}{n+1}\lim_{h\to 0}\frac{\frac{e^{2hx}-1}{n+1}}{h} \text{ ->}$$

$$\frac{-2y}{n+1}\lim_{h\to 0}\frac{(e^{x})^{h}-1}{h}+\frac{1}{n+1}\lim_{h\to 0}\frac{\frac{e^{2hx}-1}{n+1}}{h}$$

---

Let w = $e^{x}$, using the definition $\lim_{h\to 0}\frac{a^{h}-1}{h}$=ln(a) shown here: https://www.cuemath.com/calculus/derivative-of-exponential-function/

$$\frac{dL}{db_1}=\frac{-2y}{n+1}\lim_{h\to 0}\frac{w^{h}-1}{h}+\frac{1}{n+1}\lim_{h\to 0}\frac{\frac{e^{2hx}-1}{n+1}}{h} \text{ ->}$$

$$\frac{-2y}{n+1}\ln(w)+\frac{1}{n+1}\lim_{h\to 0}\frac{\frac{e^{2hx}-1}{n+1}}{h} \text{ ->}$$

$$\frac{-2y}{n+1}\ln(e^{x})+\frac{1}{n+1}\lim_{h\to 0}\frac{\frac{e^{2hx}-1}{n+1}}{h} \text{ ->}$$

$$\frac{-2y}{n+1}\mathrm{x}+\frac{1}{n+1}\lim_{h\to 0}\frac{\frac{e^{2hx}-1}{n+1}}{h}$$

---

$$\frac{dL}{db_1}=\frac{-2y}{n+1}x+\frac{1}{n+1}\lim_{h\to 0}\frac{\frac{1}{n+1}(e^{2hx}-1)}{h} \text{ ->}$$

$$\frac{-2y}{n+1}x + \frac{1}{n+1}\frac{1}{n+1}\lim_{h\to 0}\frac{e^{2hx}-1}{h} \text{ ->}$$

$$\frac{-2y}{n+1}x + \frac{1}{(n+1)^2}\lim_{h\to 0}\frac{e^{2hx}-1}{h}$$

---

Let s = $e^{2x}$

$$\frac{dL}{db_1} = \frac{-2y}{n+1}x + \frac{1}{(n+1)^2}\lim_{h\to 0}\frac{s^h-1}{h} \text{ ->}$$

$$\frac{-2y}{n+1}x + \frac{1}{(n+1)^2}\ln(s) \text{ -> } \frac{-2y}{n+1}x + \frac{1}{(n+1)^2}\ln(e^{2x}) \text{ ->}$$

$$\frac{-2y}{n+1}x + \frac{1}{(n+1)^2}2\text{x}$$

---

$$\frac{dL}{db_1} = \frac{-2y}{e^{(-b_0-b_1x)}+1}x + \frac{1}{(e^{(-b_0-b_1x)}+1)^2}2\text{x->}$$

$$\boldsymbol{\frac{dL}{db_1} = \frac{-2yx}{e^{(-b_0-b_1x)}+1} + \frac{2x}{(e^{(-b_0-b_1x)}+1)^2}}$$

(b) If you have a data point ($x,y$) = (1,1) and current parameters $b_0$ = 0 and $b_1$ = −1, what would the new values look like after applying the following update to $b_1$

$$b_1 \leftarrow b_1 - \lambda\frac{dL}{db_1}$$

$$\text{L(x)} = (y - \frac{1}{1+e^{-(b_0+b_1x)}})^2 \text{ ->}$$

$$L(x) = \left(y - \frac{1}{1+e^{-(b_0+(b_1-\lambda\frac{dL}{db_1})x)}}\right)^2 \rightarrow$$

$$L(x) = \left(y - \frac{1}{1+e^{-(b_0+(b_1-\lambda\frac{-2yx}{e^{(-b_0-b_1x)}+1}+\frac{2x}{(e^{(-b_0-b_1x)}+1)^2})x)}}\right)^2$$

---

With x=1, y=1, $b_0 = 0$, and $b_1 = -1$

$$L(1) = \left(1 - \frac{1}{1+e^{-(-1-\lambda\frac{-2}{e+1}+\frac{2}{(e+1)^2})}}\right)^2 \rightarrow$$

$$L(1) = \left(1 - \frac{1}{1+e^{1+\lambda(\frac{-2}{e+1}+\frac{2}{(e+1)^2})}}\right)^2 \rightarrow$$

$$L(1) = \left(1 - \frac{1}{1+e^{1+\lambda(\frac{2}{e+1}\left(-1+\frac{1}{e+1}\right))}}\right)^2 \rightarrow$$

$$L(1) = \left(1 - \frac{1}{1+e^{1+\lambda(\frac{2}{e+1}\left(\frac{1}{e+1}-1\right))}}\right)^2$$

From here, the value depends on the given lambda.

4. --INC For a general, twice-differentiable loss function $L(b)$ that depends on one parameter $b$, show that finding the lowest point of its Taylor series approximation of order two around the current solution $b\hat{}$ is equivalent to the gradient descent update rule (I know these "\" are not really the symbol for prime, but it is hard to see in word).

$$b \leftarrow b\hat{} - \lambda L\backslash(b\hat{}),$$

with $\lambda = 1/L\backslash\backslash(b\hat{})$.

$b \leftarrow b\hat{} - \lambda L\backslash(b\hat{})$ *with* $\lambda = 1/L\backslash\backslash(b\hat{})$ -> $b \leftarrow b\hat{} - (1/L\backslash\backslash(b\hat{}))L\backslash(b\hat{})$ ->

$b \leftarrow b\hat{} - L\backslash(b\hat{})/L\backslash\backslash(b\hat{})$

Taylor Approximation from https://www.expii.com/t/what-is-quadratic-nd-order-taylor-approximation-323#:~:text=The%202nd%20Taylor%20approximation%20of%20f(x)%20at%20a%20point,%2Ddifferentiable%20at%20x%3Da. : f(a) + f\(a)(x-a) + f\\(a)(x-a)^2

L(b^) + L\( b^)(b - b^) + L\\( b^)(b - b^)^2

-L(b^) = L\( b^)(b - b^) + L\\( b^)(b - b^)^2 --INC

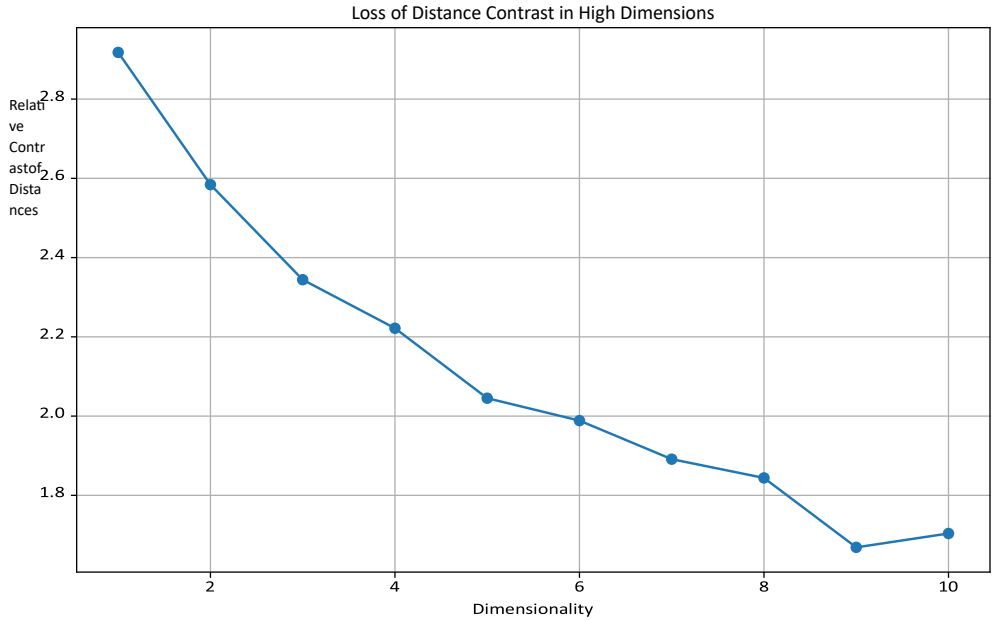


Figure 1: Programming plot

# 3 Statistical learning

1. (UML, Ch. 2, Exercise 1), For a binary classification problem with risk $L_{(D,f)}(h) = P_{x\sim D}[h(x) \neq f(x)]$, show that for a dataset of size $m$, $S = \{(x_1, y_1),\ldots, (x_m, y_m)\}$ you can build a polynomial regression formulation with thresholding that would achieve zero empirical risk

$$L_S(h) = \frac{|\{i \in \{1,\ldots,m\} : h(x) \neq f(x)\}|}{m}$$

and therefore overfit. The polynomial $p_S$ is such that $h_S(x) = 1$ if and only if $p_S(x) \geq 0$. Assume $y_i \in \{0,1\}$ and $x_i \in R^d$.

To minimize loss, $\lim_{m \to \infty} \frac{|\{i \in \{1,\ldots,m\}: h(x) \neq f(x)\}|}{m}$ must have the numerator go to 0, which is where the polynomial stipulations given by the following sentence comes into play.

If this is true for all dimensions of $x_i$ such that $x_i \in R^d$, d = 1 (thus the equation is in 1 dimension) should fit the same space. (Big shout out to the piazza post for question 3 for laying this out).

($x_i$, $y_i$) is a single point in this layout, and the polynomial can be represented with something like y=x (or something equally simple. A piecewise function would make this easier due to the pesky less than or **EQUAL TO** stipulation in the $p_S(x) \geq 0$ restriction.)

For this, I will pick y + (1/(|x!|+0.1)) -1. The -1 acts as a bias term, bringing the y to 0 if y=1 and pulling the value negative if y = 0. The 1/|x!| acts to make whatever x is chosen less than the -1 bias for all values of x, with the factorial forcing x=0 to 1 instead of undefined and the absolute value removing negative terms, as y=1 – 1 is already carefully at that 0 limit. The addition of 0.1 to the denominator also acts to prevent the y=0 case from going to 0, as 1/1 – 1 = 0. This makes the biggest cases of x=0 and x=1 still less than the -1, thus keeping the value negative when y = 0. Adding this x value place instead of subtracting also pushes the value when y=0 always above 0 and never below.

In other words, (0) - 1 + (1/(|x!|+0.1)) < 0 and (1) – 1 + (1/(|x!|+0.1)) $\geq$ 0 for $-\infty \leq x \leq \infty$ with specific cases for the two possible y values. This makes h(x) = 1 when y = 1 and h(x) = 0 when y = 0. Thus, h(x) = f(x) for all values $S = \{(x_1, y_1),\ldots, (x_m, y_m)\}$. There is thus not a case in $|\{i \in \{1,\ldots,m\} : h(x) \neq f(x)\}|$ where h(x) != f(x), and thus this value goes to 0 as m goes to infinity

Therefore, there exists at least one case where $\lim_{m \to \infty} \frac{|\{i \in \{1,\ldots,m\}: h(x) \neq f(x)\}|}{m} = 0$, making for a bad empirical hypothesis and an extremely overfit model. Heck, my equation was incredibly overfit to the specific data with all these crazy considerations for edge cases. That has to say something about empirical loss.

## Programming

Please see hw0.ipynb file starter code. Understand the “curse of dimensionality”: As dimensionality increases, the relative contrast tends to decrease, indicating that the range of distances between points becomes more uniform. This demonstrates that in high dimensions, the concept of "near" and "far" becomes less distinct, leading to challenges in methods that rely on distance metrics.

Done, turned in 9/8/2023