Homework 1

Kyle Thornton, No collaborators and no outside sources besides notes provided and Wikiped

January 2025

1 Linear Algebra review (No need for work)

1. 14
2. 0
3. $\begin{bmatrix} 6 \\ 10 \\ 14 \end{bmatrix}$
4. $\sqrt{5}$
5. $\begin{bmatrix} 0 & 1 & 2 \end{bmatrix}$
6. $\begin{bmatrix} 6 \\ 5 \\ 7 \end{bmatrix}$
7. 19

2 Matrix Algebra Rules

1. True
2. True
3. False
4. True
5. False
6. True
7. False
8. True
9. True

3 Matrix Operations

To check whether the matrix is invertible, we will take the determinant and see if it is nonzero. We can use the formula for $3x3$ matrices, and we get the equation $1(2-1)+1(-1)+0 = 0$ so the matrix is singular and has no inverse. We know if B has independent eigenvalues it can be diagonalized so lets first start by finding those. $det(B - \lambda I) = 0$ finding the roots of this equation will find the eigenvalues. This yields the matrix $\begin{bmatrix} 1-\lambda & -1 & 0 \\ -1 & 2-\lambda & -1 \\ 0 & -1 & 1-\lambda \end{bmatrix}$ using the same formula we used for determinant above we get the equation $1-\lambda((2-\lambda)(1-\lambda)-1)+(-1(2-\lambda))+0 = 0$ leading to the following $-\lambda^3+4\lambda^3-3\lambda = 0$ which has 2 different eigenvalues of 3 and 1 and 0. Now we find its eigenvectors by solving the equation $(B-\lambda I)x = 0$. For 3 we will have the equation $\begin{bmatrix} -2 & -1 & 0 \\ -1 & -1 & -1 \\ 0 & -1 & -2 \end{bmatrix} x = 0$ for 1 we have the equation $\begin{bmatrix} 0 & -1 & 0 \\ -1 & 1 & -1 \\ 0 & -1 & 0 \end{bmatrix} x = 0$ and for 0 we have $Bx = 0$. For 0 there are many solutions but one of them is the transpose of $[1, 1, 1]$, for 1 the answer is $[-1, 0, 1]$, and for three the answer is $[1, -2, 1]$ next we find the adjoint and determinant of this matrix of eigenvectors to find its inverse. This yields the equation $\begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{-1}{2} & 0 & \frac{-1}{2} \\ \frac{1}{6} & \frac{-1}{3} & \frac{1}{6} \end{bmatrix} = B.$

4 Basic Probability: No need to show work

1. 5$
2. .55
3. $\frac{10}{29}$ or roughly .345

5 Expectation and Variance

1. The expectation formula is $E[X] = \sum x_i p(x_i)$, so for this problem we are counting heads, so each head is 1 and each tail is 0. So, since we have two coins and we pick them with equal probability we will have $E[X] = .5(E[X|C_1]) + .5(E[X|C_2])$ These can be quickly calculated as $E[X|C_1] = 3 * .3 = .9$ and $E[X|C_2] = 3 * .9 = 2.7$. Then using the equation from above you get the total expectation to be 1.8.

2. I believe that the correct way to do this is to use the law of total variance. So we have that $Var[X] = E[Var(X|C)] + Var(E[X|C])$ this calculation becomes easier because the variance of a binomial variable is $np(1-p)$

where n is the number of flips and p is the probability of heads which is what we are looking for. So the expected value of the variance given one of the coins is $\frac{1}{2}$ the variance of each coin which yields the value: .45. For the next we have to calculate the $E[E[X|C]^2] - E[E[X|C]]^2$ so this yields the equation $\frac{1}{2}E[X|C_1]^2 + \frac{1}{2}E[X|C_2]^2$ subtracted by our previous answer. This yields $4.05 - 3.42 = .81$ putting the two terms together we get $Var(X) = .45 + .81 = 1.26$

3. To start I will write out what the new equation for the expectation is $E[Y] = \sum_{x=0}^{3} P(X = x) * \frac{1}{2+x}$ since we have two coins that are equally likely we know that $P(X = x) = \frac{1}{2}P(X = x|C_1) + \frac{1}{2}P(X = x|C_1)$ next we can use the binomial rules to find these probabilities. Which are $\binom{n}{k}p^k(1-p)^k$ where n is the total number of flips, k is the number of heads, and p is the probability of the heads. Now using these two summations and a script we get the result for each probability $P(X = 0) = \frac{1}{2}(.343 + .001)$, $P(X = 1) = \frac{1}{2}(.441 + .027)$, $P(X = 2) = \frac{1}{2}(.189 + .243)$, $P(X = 0) = \frac{1}{2}(.027 + .729)$ next we plug these into the equation above and sum them and get the answer that $E[Y] = .294$. Sorry for not writing out the intermediate values and just writing the equations. I started to script some of the addition to make it easier.

6 Variance Paradox

I don't believe so because X is not i.i.d. with itself, so there will be a covariance. The $Var(X + X) = Var(X) + Var(X) + 2Cov(X, X)$ equals $4Var(X) = 4\sigma^2$ because the covariance of something with itself is its variance. However, in the previous statement we know that $X_1...X_n$ are all i.i.d. so there is no covariance so it simplifies down to just $n\sigma^2$.

7 One variable Derivatives: No need to show work

1. $6x + 2$
2. $1 - 2x$
3. $p(x)$

8 Multi-Variable Derivatives

1. $[2x_1, e_2^x]$
2. $[e^{x_1+x_2x_3}, x_3e^{x_1+x_2x_3}, x_2e^{x_1+x_2x_3}]$
3. $[a_1, a_2]$

4. $[4x_1 - 2x_2, -2x_1 + 2x_2]$

5. $[x_1, x_2, ..., x_d]$

9 Algorithms and Data Structures Review

1. $O(nlog_2(n))$

2. $O(n)$

3. $O(log_2(n))$

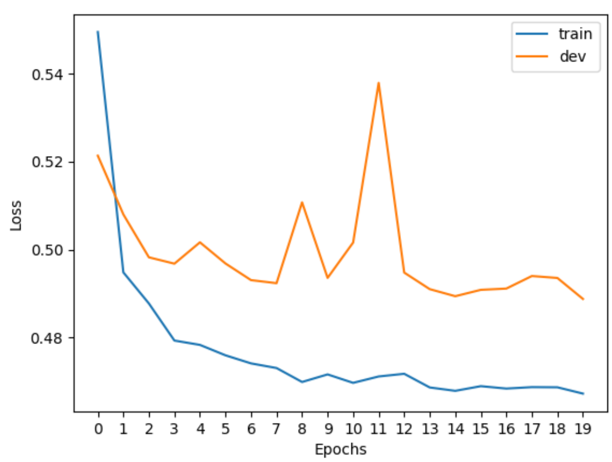
4. $O(1)$ amortized and $O(n)$ worst case

5. d^2

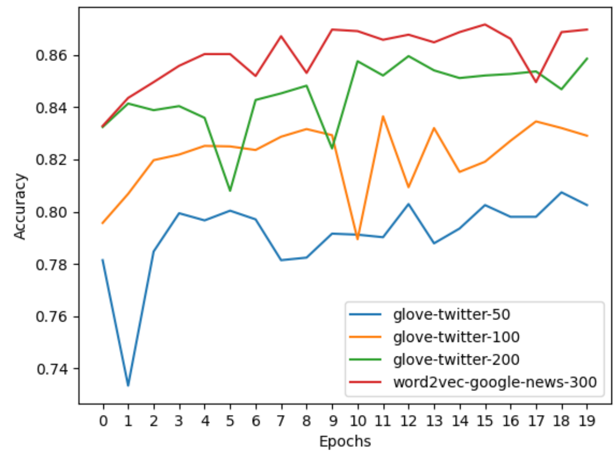
6. d^3

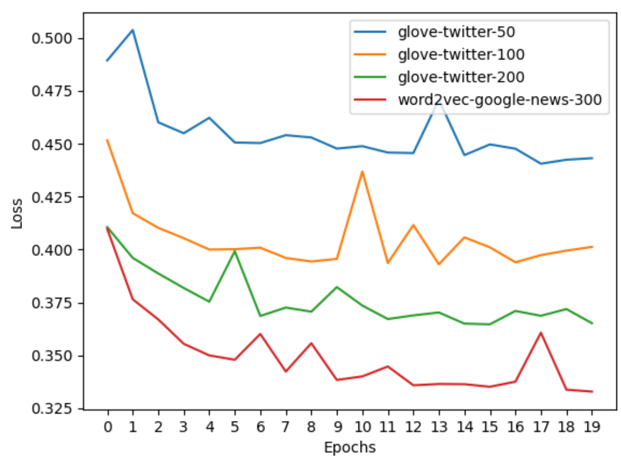
7. $m * n^2$

10 Programming Assignment



This is my loss graph for the 20 epochs designated by the single_loss function. There seems to be a spike in the dev dataset which could be due to over fitting to the training dataset but it seems the model was able to course correct and bring the loss to below any other epoch by the 19th epoch. Other than that I am surprised at how well the model performed. I am going to have to ask more questions regarding the averaging of the embeddings. Does this only work in sentiment analysis?





At a glance it looks like the embeddings that have the most vocab size ranked highest in accuracy and lowest in loss. Going from GloVe 50, GloVe 100, GloVe 200, word2vec 300 which if I understand correctly means that they have more embedding dimensions. This allows for it to have more vocab and represent the words in a higher vector space possibly increasing the accuracy for meaning behind the cosine similarity score of two words. However, this also means each one is more computationally expensive than the last. All this being said, it looks like the word2vec-google-news-300 performed the best out of the embeddings.