CS 165B: Homework #4

Due on June 8, 2023

Xifeng Yan 07575

Qiyuan Zhuang

Problem 1

Given 4 data points in 3-d space, (1, 1, 0), (2, 2, 0), (0, -1, -1), (0, -2, -2).

(1) **Please explain what PCA can be used for and give a geometrical interpretation of first principal component of PCA.**

We can use PCA to rotate the data that maximizes the variance in the new axes, reduce the dimensionality by choosing only the first p eigenvector and reserve the greatest variance good for clustering. Geometrical interpretation: the first principal component of the dataset is the direction along which the data varies the most(maximize the variance). In the 3-d space, we can draw a straight line(direction) on which we can project the data points. The line passing mean of the data points has a direction of the highest variance of data.

(2) **Please show how to compute the first principal component given the data points. You only need to show how to compute it and do not need to give the exact number of the first principal component.**

Step 1 : transform the data into zero mean data. $x_i = x_i - \frac{1}{4}\sum_{j=1}^{4} x_j$
Step 2 : Calculate the covariancematrix (3×3).
Step 3 : Calculate the eigenvectors and eigenvalues of the covariance matrix.
Step 4 : Dimensionality Reduction. Order eigenvectors by their eigenvalue, highest to lowest. The eigenvector with the highest eigenvalue is the first principal component of the data set.

(3) **Please show how to compute the projected values for the data points using the first principal component. You can use w=(w1,w2,w3) as the first principal component.**

The projection of a data point x is $P(x) = \frac{\boldsymbol{w}\cdot\boldsymbol{x}^{\boldsymbol{T}}}{\|\boldsymbol{w}\|^{\boldsymbol{2}}}\boldsymbol{w}$. It contains coordinates of projected point.

Problem 2

There are four data points: x1 = (0, 2), x2 = (1, 1), x3 = (3, 2), x4 = (4, 0), and you are given two initial cluster centers: m1 = (0, 0), m2 = (3, 1). Please simulate one iteration of K-means clustering.

Assign data points: $x1 \to m1 \quad x2 \to m1 \quad x3 \to m2 \quad x4 \to m2$.
Update the cluster centers: m1 = (0.5, 1.5), m2 = (3.5, 1).

Problem 3

Let C be the classifier that returns a majority vote of the three classifiers. Assuming the errors of the C_i are independent, what is the probability that C(x) will be correct on a new test example x?

$$\begin{aligned} P(C = true) &= P(\sum(C_i = true) >= 2) \\ &= \sum_{a,b,c} P(C_a = C_b = true, C_c = false) + P(C_1 = C_2 = C_3 = true) \\ &= (0.7 \cdot 0.6 \cdot 0.2 + 0.7 \cdot 0.4 \cdot 0.8 + 0.3 \cdot 0.6 \cdot 0.8) + 0.7 \cdot 0.6 \cdot 0.8 \\ &= 0.788. \end{aligned}$$

Problem 4

Suppose there are 12 samples in total, among which 9 samples are correctly classified and 3 samples are incorrectly classified. The initial weight for each sample is 1/12. What is the updated weight for correctly classified samples and misclassified samples for the Boosting algorithm?

error rate: $\epsilon_1 = 3/12 = 0.25$.
updated weight for correctly classified samples: $w' = \frac{w}{2(1-\epsilon_1)} = \frac{1}{18}$.
updated weight for correctly classified samples: $w' = \frac{w}{2\epsilon_1} = \frac{1}{6}$.

Problem 5

What is the class returned by the Adaboost ensemble classifier H on test example x?

$H(x) = \sum_{t=1}^{3} \alpha_t h_t(x) = 0.3 + 0.1 - 0.6 = -0.2 < 0$ which means $\text{sign}(H(x)) = -1$.
So class -1 will be returned.

Problem 6

For forward propagation,

(a) **Calculate** $z1 = W1x + b1$

$$z1 = W1x + b1 = \begin{bmatrix} 5.6 \\ 1.2 \\ -3.8 \end{bmatrix} + \begin{bmatrix} 0.3 \\ 0.5 \\ -0.7 \end{bmatrix} = \begin{bmatrix} 5.9 \\ 1.7 \\ -4.5 \end{bmatrix}.$$

(b) **Calculate** $h1 = \sigma(z1)$

$$h1 = \sigma(z1) = \frac{1}{1+e^{-z1}} = \begin{bmatrix} 0.9973 \\ 0.8455 \\ 0.0110 \end{bmatrix}.$$

(c) **Calculate** $z2 = W2h1 + b2$

$$z2 = W2h1 + b2 = 0.2 * 0.9973 + 0.8 * 0.8455 - 0.5 * 0.0110 + 0.6 = 1.47037.$$

(d) **Calculate** $\hat{y} = \sigma(z2)$

$$\hat{y} = \sigma(z2) = \frac{1}{1+e^{-z2}} = 0.8131.$$

Problem 7

For the neural network in Problem 6, consider backward propagation.

(a) **Calculate** $\partial L/\partial \hat{y}$

$$\frac{\partial L}{\partial \hat{y}} = -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}} = -\frac{1}{0.8131} = -1.22986.$$

(b) **Calculate** $\partial L/\partial z2$

$$\frac{\partial L}{\partial z2} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z2} = \frac{\partial L}{\partial \hat{y}}\sigma(z2)(1-\sigma(z2)) = -0.1869.$$

(c) **Calculate** $\partial z2/\partial W2$ **(there are three numbers)**

$$\frac{\partial z2}{\partial W2} = h1 = \begin{bmatrix} 0.9973 \\ 0.8455 \\ 0.0110 \end{bmatrix}.$$

(d) **Calculate** $\partial L/\partial W2$ **(there are three numbers)**

$$\frac{\partial L}{\partial W2} = \frac{\partial L}{\partial z2}\frac{\partial z2}{\partial W2} = \frac{\partial L}{\partial z2}h1 = \begin{bmatrix} 0.1864 \\ 0.1580 \\ 0.0021 \end{bmatrix}.$$

Problem 8

(a) **Which assumption is word2vec based on?**

It's based on the Distributional Hypothesis in linguistics, which states that "a word shall be known by the company it keeps". Word2vec uses regularities between words. You can get a lot of value by representing a word by means of its neighbors(context).

(b) **What is the advantage of negative sampling in skip-gram?**

It will simplify the process of updating weights. If we use negative sampling, only a small percentage of weights will be updated. It will reduce the compute burden of the model, and improve the quality of their resulting word vectors as well.

(c) **Suppose the semantic space is well learned, and we know the embedding of king, queen, man. How can we calculate the vector of woman?**

King is similar to Queen as Man is similar to Woman:

$$\boldsymbol{V_{woman} = V_{queen} - V_{king} + V_{man}}$$

Problem 9

(a) **Please compute the number of parameters for this model when n1=6, n2=16, n3=128. Write out the number of weights and number of biases separately.**

First layer : $6 \times 5 \times 5 = 150$ weights, 6 bias.
Third layer : $5 \times 5 \times 6 \times 16 = 2400$ weights, 16 bias.
Fifth layer : $4 \times 4 \times 16 \times 128 = 32768$ weights, 128 bias.
Sixth layer : $128 \times 10 = 1280$ weights, 10 bias.
Sum : 36598 weights, 160 bias.

(b) **Based on your calculation, what is one significant advantage of using convolutional layer over fully-connected layers when handling image data?**

A fully-connected layer would have a huge set of weights to learn, while much less weights(bias as well) need to be learned in a convolutional layer. That means model will be more efficient and trained faster using convolutional layers.