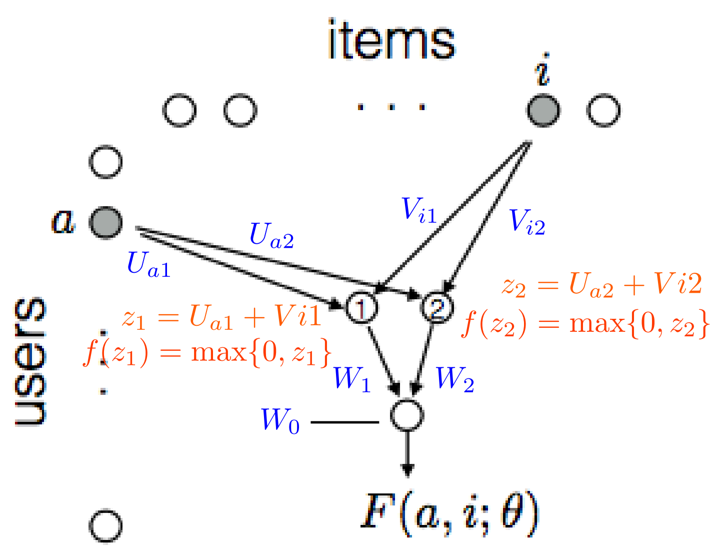
Problem 3

Suppose we have a recommender problem with n users $a \in \{1, \ldots, n\}$ and m items $i \in \{1, \ldots, m\}$. For simplicity, we will treat the target rating values as class labels, i.e., using $\{-1, 1\}$ ratings (dislikes,likes). Each user is likely to provide feedback for only a small subset of possible items and thus we must constrain the models so as not to overfit. Our goal here is to understand how a simple neural network model applies to this problem, and what its constraints are.



Schematic representation of the simple neural network model

We use the simple neural network depicted above. We introduce an input unit corresponding to each user and each item. In other words, there are $n + m$ input units. In the figure above, the n input units corresponding to the users are on the left and the m input units corresponding to the items are on top.

When querying about a selected entry (a, i), only the ath user input unit and ith item input unit are active (set to 1), the rest are equal to zero and will not affect the predictions. Put another way, only the outgoing weights from these two units matter for predicting the value (class label) for entry (a, i).

User a has two outgoing weights, U_{a1} and U_{a2}, and item i has two outgoing weights, V_{i1} and V_{i2}. These weights are fed as inputs to the two hidden units in the model. The hidden units evaluate

$$\begin{aligned} z_1 &= U_{a1} + V_{i1}, \quad f(z_1) = \max\{0, z_1\} \\ z_2 &= U_{a2} + V_{i2}, \quad f(z_2) = \max\{0, z_2\}. \end{aligned}$$

Thus, for the (a, i) entry, our network outputs

input

$$F(a, i; \theta) = W_1 f(z_1) + W_2 f(z_2) + W_0$$

where θ denotes all the weights U, V, and W. Finally, a sign function is applied to $F(a, i; \theta)$ for the classification.

for 3.(2)

In vector form, each user a has a two-dimensional vector of outgoing weights

$$\vec{u}_a = [U_{a1}, U_{a2}]^T$$

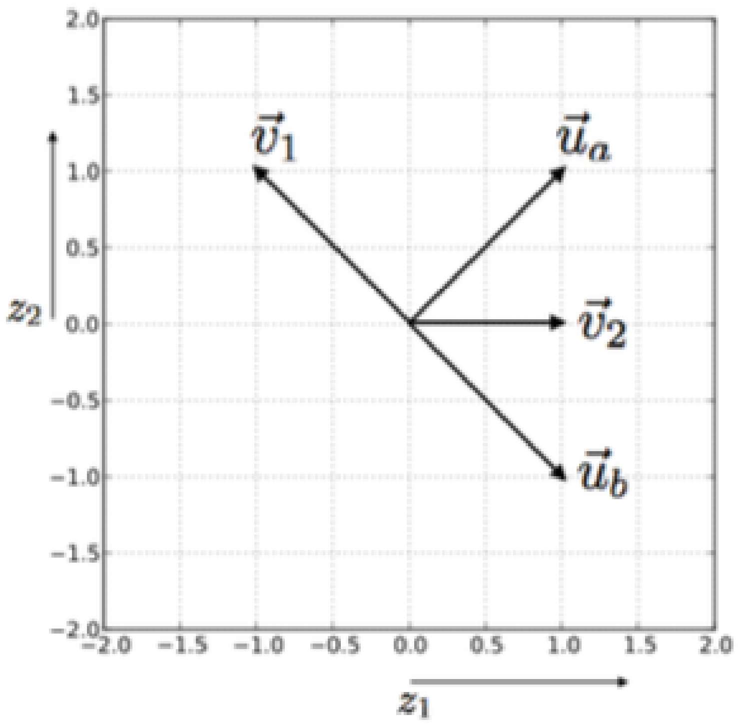
and similarly each item i has a two-dimensional vector of outgoing weights

$$\vec{v}_i = [V_{i1}, V_{i2}]^T.$$

The input received by the hidden units, if represented as a vector, is

$$\vec{z} = [z_1, z_2]^T = \vec{u}_a + \vec{v}_i.$$

Consider a simple version of the problem where we have only two users, $\{a, b\}$, and two items $\{1, 2\}$. So the recommendation problem can be represented as a 2×2 matrix. We will initialize the first layer weights as shown in figure below.

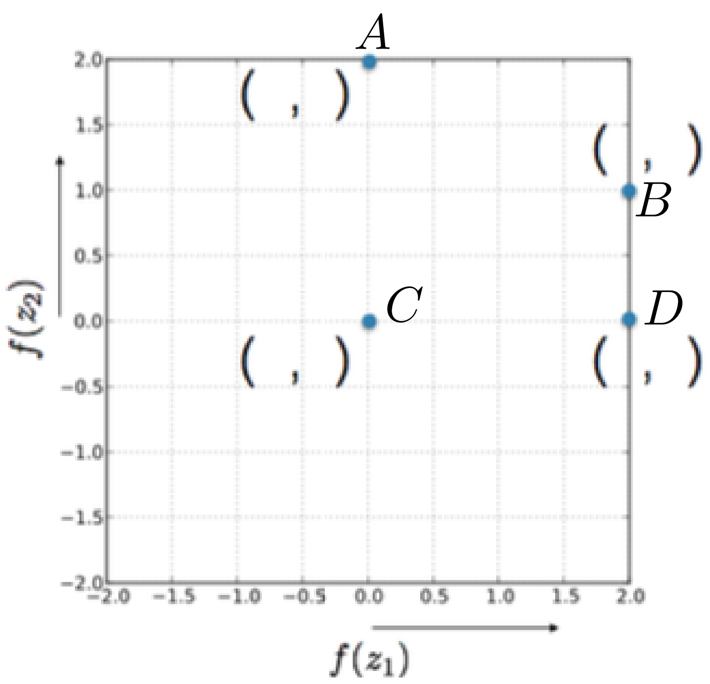


3. (1)

4/4 points (graded)

Using the initial input-to-hidden layer weights. each of the four user-item pairs in the 2×2 matrix, namely $(a, 1)$, $(a, 2)$, $(b, 1)$, $(b, 2)$, are mapped to a corresponding feature representation $[f(z_1), f(z_2)]^T$ (hidden unit activations).

For each of the points labeled A, B, C, D in the figure below, select the correct pair, e.g., $(a, 1)$, that it corresponds to.



(Choose one for each column below.)

$A:$	$B:$	$C:$	$D:$
◉ $(a, 1)$	○ $(a, 1)$	○ $(a, 1)$	○ $(a, 1)$
○ $(a, 2)$	◉ $(a, 2)$	○ $(a, 2)$	○ $(a, 2)$
○ $(b, 1)$	○ $(b, 1)$	◉ $(b, 1)$	○ $(b, 1)$
○ $(b, 2)$	○ $(b, 2)$	○ $(b, 2)$	◉ $(b, 2)$

Solution:

Recall

$$z_1 = U_{a1} + V_{i1}, f(z_1) = \max\{0, z_1\} \tag{9.1}$$
$$z_2 = U_{a2} + V_{i2}, f(z_2) = \max\{0, z_2\}. \tag{9.2}$$

In vector form:

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}_{a,i} = \vec{u}_a + \vec{v_i}$$

Hence, for each user-item pair, we have:

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}_{a,1} = \vec{u}_a + \vec{v_1} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$
$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}_{a,2} = \vec{u}_a + \vec{v_2} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$
$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}_{b,1} = \vec{u}_b + \vec{v_1} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$
$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}_{b,2} = \vec{u}_b + \vec{v_1} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}.$$

Applying f to each component of each vector, we see that

$$\begin{bmatrix} f(z_1) \\ f(z_2) \end{bmatrix}_{a,i} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}_{a,i} \quad \text{for } i = 1, 2$$
$$\begin{bmatrix} f(z_1) \\ f(z_2) \end{bmatrix}_{b,1} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}_{b,1}$$
$$\begin{bmatrix} f(z_1) \\ f(z_2) \end{bmatrix}_{b,2} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

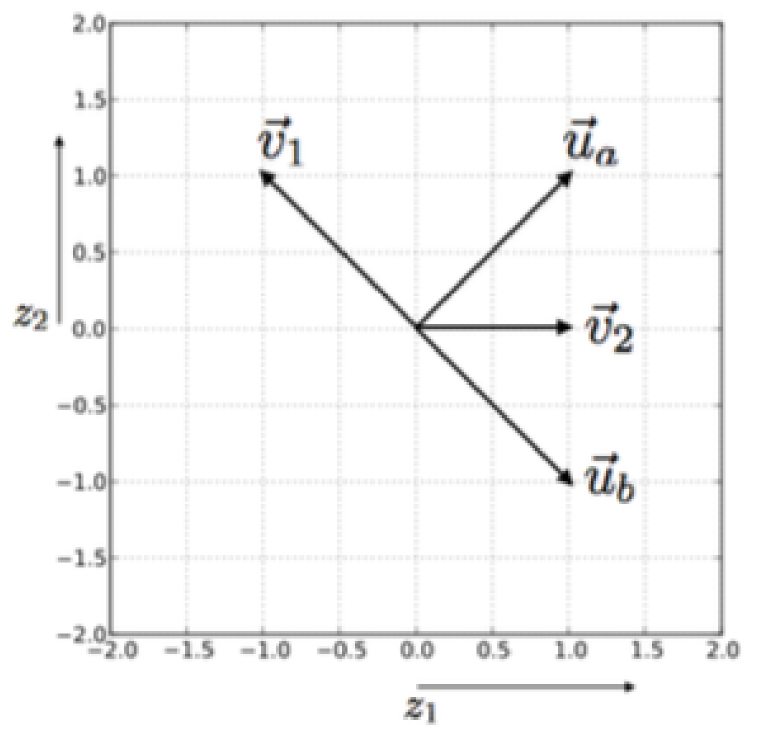
Submit You have used 1 of 3 attempts

Answers are displayed within the problem

3. (2)

0/1 point (graded)

Recall the initial values of the hidden layer weights are as in the figure below.



Suppose we keep the input to the hidden layer weights (U 's and V 's) at their initial values shown above, and only estimate the weights W corresponding to the output layer.

Different choices of the output layer weights will result in different predicted 2×2 matrices of $\{-1, 1\}$ labels. Which of the following matrices is one that the neural network **cannot** reproduce with any choice of the output layer weights W_1, W_2, and W_0? (Choose all that apply.)

output

☑

$$\begin{array}{c|cc} & 1 & 2 \\ \hline a & +1 & +1 \\ b & +1 & -1 \end{array}$$

$$\text{sign}\left(\begin{array}{l} 0\,W_1 + 2\,W_2 + W_0 \\ 2\,W_1 + 1\,W_2 + W_0 \\ 0\,W_1 + 0\,W_2 + W_0 \\ 2\,W_1 + 0\,W_2 + W_0 \end{array}\right) \begin{array}{l} = M_{1,1} \\ = M_{1,2} \\ = M_{2,1} \\ = M_{2,2} \end{array}$$

Sharing the solution using Sage (since there is none in the exam), it's the same system above but using inequalities (for the sign function):

```
In [4]: W_1, W_2, W_0 = var('W_1, W_2, W_0')

        solve([2*W_2 + W_0 > 0,
               2*W_1 + 1*W_2 + W_0 > 0,
               W_0 > 0,
               2*W_1 + W_0 < 0], W_1, W_2, W_0)

Out[4]: [[max(0, -2*W_1 - W_2) < W_0, W_0 < -2*W_1, W_1 < 0, 0 < W_2]]

In [5]: W_1, W_2, W_0 = var('W_1, W_2, W_0')

        solve([2*W_2 + W_0 < 0,
               2*W_1 + 1*W_2 + W_0 > 0,
               W_0 > 0,
               2*W_1 + W_0 < 0], W_1, W_2, W_0)

Out[5]: []
```

There is no solution for the second.

☑

$$\begin{array}{c|cc} & 1 & 2 \\ \hline a & -1 & +1 \\ b & +1 & -1 \end{array}$$

✔

✖

Solution:

Don't understand, how is it related to linear separable?

- The problem is equivalent to asking whether a linear classifier can classify the points in last problem correctly.
- The left can be separated by a linear classifier, and the right figure cannot.

Submit You have used 2 of 3 attempts

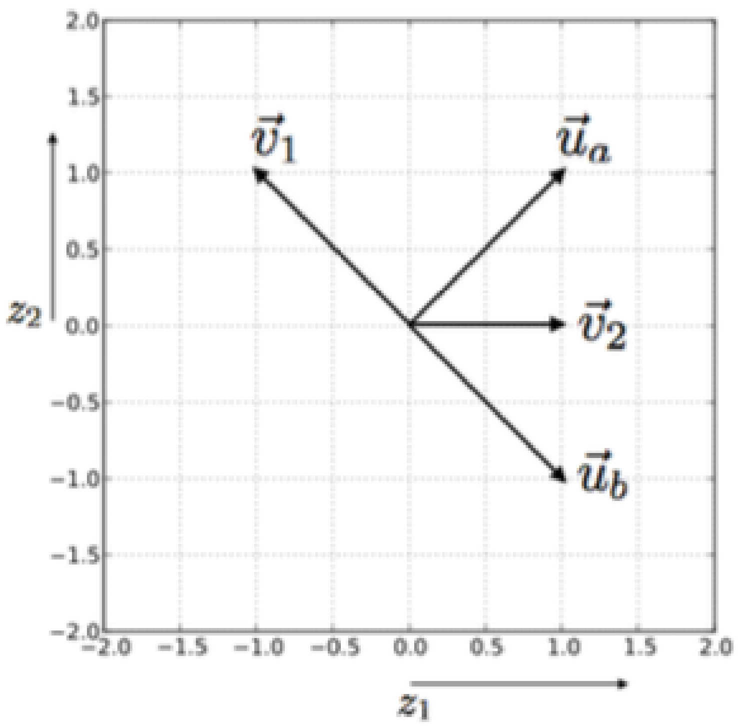
Answers are displayed within the problem

3. (3)

1/1 point (graded)

Learning a new representation for examples (hidden layer activations) is always harder than learning the linear classifier operating on that representation. In neural networks, the representation is learned together with the end classifier using stochastic gradient descent.

We initialize the output layer weights as $W_1 = W_2 = 1$ and $W_0 = -1$. Assume that all the weights are initialized as previously:



What is the class label $(+1/-1)$ that the network would predict in response to $(b, 2)$ (user b, item 2)?

1 ✔ Answer: 1

Solution:

$1 \times 2 + 1 \times 0 - 1 = 1$

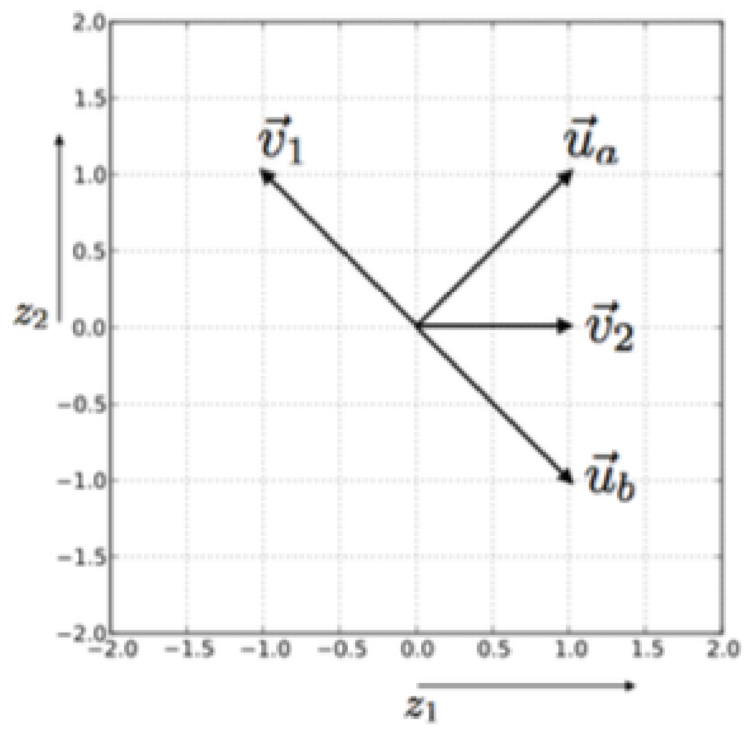
Submit You have used 1 of 3 attempts

Answers are displayed within the problem

3. (4)

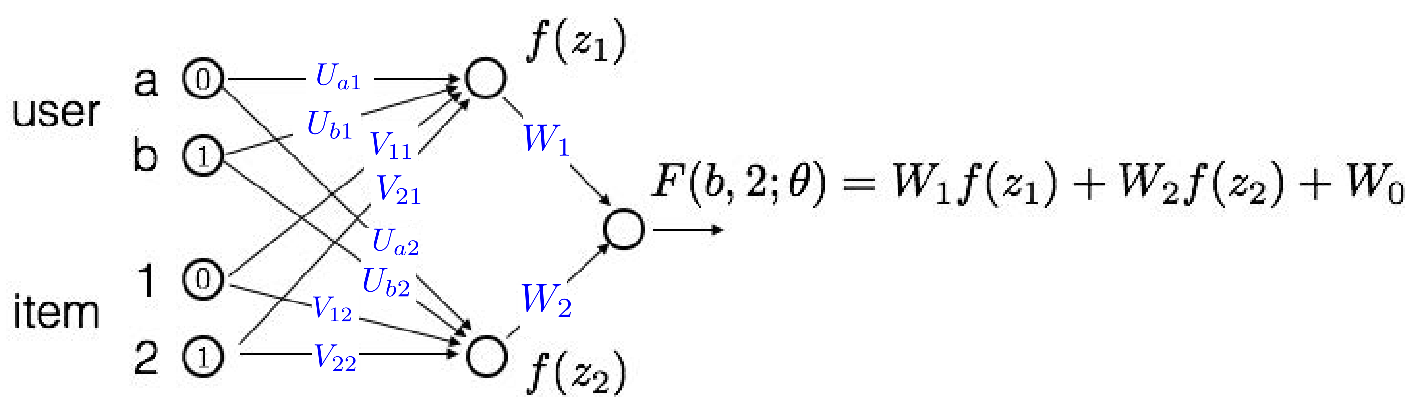
0/1 point (graded)

Assume that we observe the opposite label from your answer to the previous question. In other words, there is a training signal at the network output. All the weights are initialized as previously, i.e. $W_1 = W_2 = 1$ and $W_0 = -1$and



All the weights are initialized as previously, i.e. $W_1 = W_2 = 1$ and $W_0 = -1$ and the Us and V's are given by the figure above.

Which of the weights, depicted in blue in the schematic diagram below, would change (have non-zero update) based on a single stochastic gradient descent step in response to $(b, 2)$ with our specific weight initialization and the target label?



Note that the input units a, b and $1, 2$ are activated with 0's and 1's as shown inside the circles. You are not asked whether W_0 would change.

(Choose all that apply.)

- [ ] U_{a1}
- [x] U_{b1} ✔

- [ ] V_{11}
- [x] V_{21} ✔
- [ ] U_{a2}
- [x] U_{b2}
- [ ] V_{12}
- [x] V_{22}
- [x] W_1 ✔
- [x] W_2

✖

$\vec{u}_b = [U_{b1}, U_{b2}]$ and $\vec{v}_2 = [V_{21}, V_{22}]$ represent first layer weights in our NN model. They don't represent the data.

Solution:

The weight will be updated if it is involved in the path to generate prediction. Hence, all points that interact with $f_2(z_2)$ will not be updated, since its value is zero caused by ReLU.

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Error and Bug Reports/Technical Issues

Show Discussion

Topic: Final exam (1 week):Final Exam / Problem 3