# CS 165B – Machine Learning, Spring 2023

## Assignment #4
## Due Thursday, June 08 by 11:59pm PST

---

**Notes:**

- *This assignment is to be done individually. You may discuss the problems at a general level with others in the class (e.g., about the concepts underlying the question, or what lecture or reading material may be relevant), but the work you turn in must be solely your own.*
- Zoey Song zhenqiao@ucsb.edu is the head TA for this HW. Please contact her for questions.
- Be sure to re-read the "Policy on Academic Integrity" on the course syllabus.
- Be aware of the late policy in the course syllabus – i.e., *late submissions will be accepted but 20% deduction per day*.
- Justify every answer you give – **show the work** that achieves the answer or **explain** your response.
- Any updates or corrections will be posted on Piazza, so check there occasionally.
- **All assignments must be clear and legible.** It is recommended to type the solutions on this Microsoft Word file directly and save it as PDF. If you'll be submitting a handwritten assignment, please ensure that it's readable and neat. If your writing is not easily readable, your solution is not easy to follow on the page, or your PDF is not of very high quality, your assignment will not be graded. **DO NOT submit picture of your written work.** (If you must scan your written work in, use a high-quality scanner. Plan in advance.)
- To turn in your assignment:
  - Submit a PDF version to Gradescope

---

**Problem 1 [10 points]**
Given 4 data points in 3-d space, (1, 1, 0), (2, 2, 0), (0, -1, -1), (0, -2, -2).
(1) [2pts] Please explain what PCA can be used for and give a geometrical interpretation of first principal component of PCA.
(2) [4pts] 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.
(3) [4pts] 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.

**Problem 2 [10 points]**
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. In each iteration: (1) Assign data points to a cluster center. (2) Update the cluster centers.

**Problem 3 [10 points]**
Suppose you have trained three classifiers, each of which returns either 1 or −1, and their accuracies are listed as below:

| Classifier | Accuracy |
|---|---|
| $C_1$ | 0.7 |
| $C_2$ | 0.6 |
| $C_3$ | 0.8 |

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?

**Problem 4 [10points]**
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?

**Problem 5 [10 points]**
Suppose you have run Adaboost on a training set for three boosting iterations. The results are classifiers h1, h2, and h3, with coefficients $\alpha 1 = 0.3$, $\alpha 2 = 0.1$, and $\alpha 3 = 0.6$. You find that the classifiers' results on a test example x are $h1(x) = 1$, $h2(x) = 1$, and $h3(x) = -1$, What is the class returned by the Adaboost ensemble classifier H on test example x?

**Problem 6 [10 points]**

Let's consider a simple two-layer neural network (MLP). It has input size 2, one hidden layer size 3, and output size 1.

The input $x=\begin{bmatrix}-2\\3\end{bmatrix}$, y = 1

Two sets of weights $W1=\begin{bmatrix}-1.6 & 0.8\\0.3 & 0.6\\1.6 & -0.2\end{bmatrix}$, W2=[0.2 0.8 -0.5]

Biases $b1=\begin{bmatrix}0.3\\0.5\\-0.7\end{bmatrix}$, b2=0.6

All the non-linear activation function is sigmoid function $\sigma(x)=\frac{1}{1+e^{-x}}$, and loss function:

$$\mathcal{L}=-y\log\hat{y}-(1-y)\log(1-\hat{y}).\ (\text{hint: }\sigma'(x)=\sigma(x)*(1-\sigma(x)))$$

For forward propagation,

(a) Calculate z1=W1x+b1
(b) Calculate h1= σ(z1)
(c) Calculate z2=W2h1+b2
(d) Calculate $\hat{y}=\sigma(z2)$

**Problem 7 [10 points]**

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

(a) Calculate $\partial L/\partial\hat{y}$
(b) Calculate $\partial L/\partial z2$
(c) Calculate $\partial z2/\partial W2$ (there are three numbers)
(d) Calculate $\partial L/\partial W2$ (there are three numbers)

**Problem 8 [10 points]**

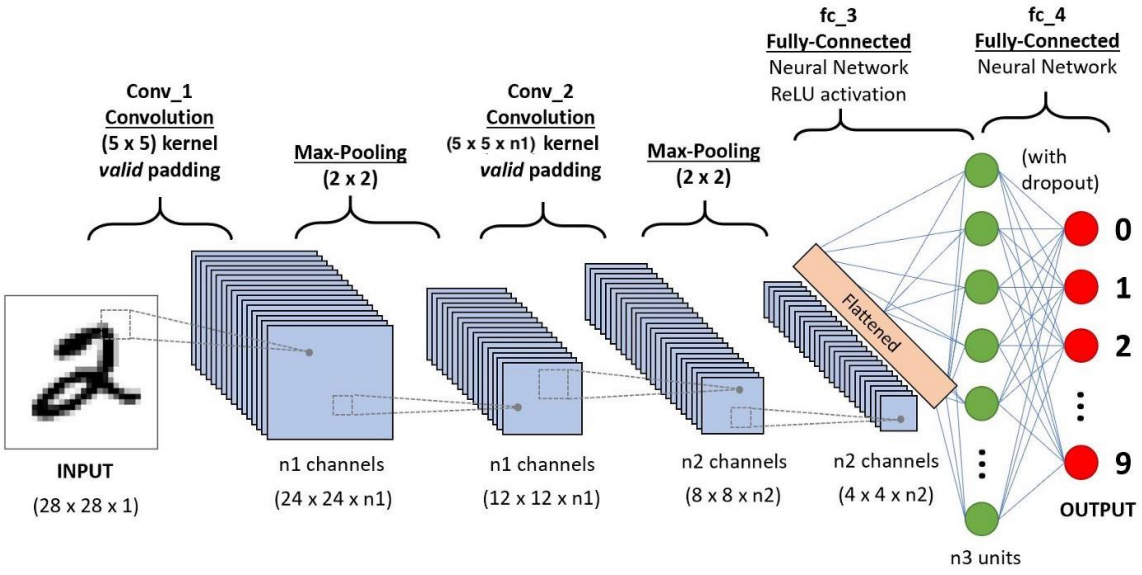
**(1)** Which assumption is word2vec based on?

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

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

**Problem 9 [10 points]**

The below CNN model is for digital recognition. The first layer is doing convolution where a kernel 5*5 is used. The second layer is doing pooling where the max-pooling is applied. The third layer is doing convolution where a kernel 5*5*n1 is used. The forth layer is doing pooling where the max-pooling is applied. The fifth layer is a fully connected hidden layer where the hidden size is n3. The sixth layer is a fully connected




output layer where the number of output is 10.

(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.

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