CS 182: Introduction to Machine Learning, Fall 2022
Homework 5

(Due on Wednesday, Dec. 20 at 11:59pm (CST))

Notice:

- Please submit your assignments via Gradescope. The entry code is G2V63D.
- Please make sure you select your answer to the corresponding question when submitting your assignments.
- Each person has a total of five days to be late without penalty for all the assignments. Each late delivery less than one day will be counted as one day.

1. [10 points] *[Deep Learning Models]*

   (a) Consider a 3D convolution layer. Suppose the input size is $32 \times 32 \times 3$ (width, height, depth) and we use ten 5×5 (width, height) kernels to convolve with it. Set stride $= 1$ and pad $= 2$. What is the output size? Let the bias for each kernel be a scalar, how many parameters do we have in this layer? [5 points]

   (b) The convolution layer is followed by a max pooling layer with 2×2 (width, height) filter and stride $= 2$. What is the output size of the pooling layer? How many parameters do we have in the pooling layer? [5 points]

2. [30 points] *[Deep Learning Models]* Principal component analysis (PCA) and autoencoders are popular tools for dimension reduction in machine learning. In this problem, we look into the relation between PCA and the linear autoencoders.

   (a) Given a sample matrix $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_M] \in \mathbb{R}^{d\times M}$ where each column denotes a d-dimensional zero-mean sample. The goal of PCA is to find an orthogonal matrix (transformation) $\mathbf{W} \in \mathbb{R}^{d\times r}$ ($r \leq d$) which is the solution to

$$\underset{\mathbf{W}}{\text{maximize}} \;\; \text{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}) \quad \text{s.t.} \;\; \mathbf{W}^T\mathbf{W} = \mathbf{I}_r,$$

   where $\text{Tr}(\cdot)$ denotes the trace and $\mathbf{I}_r$ is an $r \times r$ identity matrix.
   Show that

      i. when $r = 1$, $\mathbf{W}$ is exactly the eigenvector of $\mathbf{X}\mathbf{X}^T$ corresponding to its largest eigenvalue. [7 points]

      ii. $\mathbf{W}$ is also the solution to

$$\underset{\mathbf{W}}{\text{minimize}} \;\; \|\mathbf{X} - \mathbf{W}\mathbf{W}^T\mathbf{X}\|_F^2 \quad \text{s.t.} \;\; \mathbf{W}^T\mathbf{W} = \mathbf{I}_r,$$

      where $\|\cdot\|_F$ is the Frobenius norm, i.e., $\|\mathbf{P}\|_F = \sqrt{\text{Tr}(\mathbf{P}^T\mathbf{P})}$. [7 points]

   (b) Consider a linear autoencoder (as a neural network) with a single hidden layer structure:

$$\mathbf{H} = \mathbf{A}_1\mathbf{X} + \mathbf{b}_1\mathbf{1}^T$$
$$\hat{\mathbf{X}} = \mathbf{A}_2\mathbf{H} + \mathbf{b}_2\mathbf{1}^T,$$

   where $\mathbf{A}_1 \in \mathbb{R}^{r\times d}$ ($\mathbf{A}_2 \in \mathbb{R}^{d\times r}$) and $\mathbf{b}_1 \in \mathbb{R}^{r\times 1}$ ($\mathbf{b}_2 \in \mathbb{R}^{d\times 1}$) are respectively the weight matrix and the bias vector of the layer in the encoder (decoder), and $\mathbf{1} = [1, ..., 1]^T \in \mathbb{R}^M$. One trains the linear autoencoder by minimizing the reconstruction error:

$$\{\mathbf{A}_1^\star, \mathbf{b}_1^\star, \mathbf{A}_2^\star, \mathbf{b}_2^\star\} = \underset{\mathbf{A}_1,\mathbf{b}_1,\mathbf{A}_2,\mathbf{b}_2}{\text{argmin}} \;\; \|\mathbf{X} - \hat{\mathbf{X}}\|_F^2.$$

   Show that

      i. $\mathbf{A}_2^\star$ can be solved from:

$$\underset{\mathbf{A}_2}{\text{minimize}} \;\; \|\mathbf{X} - \mathbf{A}_2\mathbf{A}_2^\dagger\mathbf{X}\|_F^2,$$

      where $\mathbf{A}_2^\dagger = (\mathbf{A}_2^T\mathbf{A}_2)^{-1}\mathbf{A}_2^T$ is the Moore-Penrose pseudoinverse of $\mathbf{A}_2$. [8 points]
      (*Hint:* use the derivative of Frobenius norm and the fact that $\mathbf{A}_2^\dagger = \underset{\mathbf{A}_1}{\text{argmin}}\|\mathbf{X} - \mathbf{A}_2\mathbf{A}_1\mathbf{X}\|_F^2$)

      ii. the solution $\mathbf{W}$ from (a) can be taken as the same as $\mathbf{A}_2^\star$. [8 points]
      (*Hint:* you may prove it by showing the equivalence of the column spaces spanned by these two matrices)

3. [15 points] *[Ensemble Learning]* Suppose there are L independent two-class classifiers used for simple voting and the output of classifier j ($j = 1, \ldots, L$) is denoted as d_j. From the point of view that the mean squared error of an estimator can be decomposed into the bias part and the variance part, explain why increasing L can lead to the increase of the classification accuracy.

4. [15 points] *[Model Assessment and Selection]* Suppose we carry out a K-fold cross-validation on a dataset and obtain the classification error rates $\{p_i\}_{i=1}^{K}$, describe the steps of a one-sided t test on testing the null hypothesis H_0 that the classifier has error percentage p_0 or less at a significance level α.

5. [30 points] *[Coding: CNN]* Complete "HW5-Coding.ipynb". After completion, you should convert your notebook to PDF, and concatenate the writing part and the coding part into one PDF which is the file to submit.