CARNEGIE MELLON UNIVERSITY 10-315
HOMEWORK 10

DUE: Saturday, April 22, 2023

https://www.cs.cmu.edu/~10315

INSTRUCTIONS

- **Format:** Use the provided LaTeX template to write your answers in the appropriate locations within the *.tex files and then compile a pdf for submission. We try to mark these areas with STUDENT SOLUTION HERE comments. Make sure that you don't change the size or location of any of the answer boxes and that your answers are within the dedicated regions for each question/part. If you do not follow this format, we may deduct points.

  You may also type you answer or write by hand on the digital or printed pdf. Illegible handwriting will lead to lost points. However, we suggest that try to do at least some of your work directly in LaTeX.

- **How to submit written component:** Submit to Gradescope a pdf with your answers. Again, make sure your answer boxes are aligned with the original pdf template.

- **How to submit programming component:** See section Programming Submission for details on how to submit to the Gradescope autograder.

- **Policy:** See Piazza for updated policy for this assignment.

Name	
Andrew ID	
Hours to complete all components (nearest hour)	

Project teammates. Required if collaborating on this assignment.

	Name(s)	AndrewID(s)
1)		
2)		

1 [5 pts] K-Means

1.1 Non-increasing with K

Given a set of N observations $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$, and cluster centers $\mu_1, ..., \mu_K$, assign each observation to its closest center. Each point has corresponding cluster assignment, $z_i \in \{1, \dots, K\}$ for i-th point.

Consider the K-means optimization as a function of the hyperparameter for the number of clusters, K:

$$J(K) = \min_{\mu_1,\dots,\mu_K,z_1,\dots,z_N} \sum_{i=1}^{N} ||\mathbf{x}^{(i)} - \mu_{z_i}||_2^2$$

Prove that $J(K)$ is a non-increasing function of $K \in \mathbb{Z}^+$. You write your proof in a series of sentences. To receive full credit, each statement in your proof should have sound reasoning and the proof must be complete, e.g. cover all cases for K.

Hint: Try to consider two cases: 1) $K \geq N$, and 2) $K \leq N$.

Your Answer

2 [21 pts] Gaussian Mixture Models

In this section, you will study the update rules for Gaussian Mixture Models (GMMs) using the expectation-maximization (EM) algorithm.

Remember that the EM algorithm has two steps:

$$\begin{aligned} E_{Z^{(i)}|X^{(i)},\theta^{(t)}}\left[Z_k^{(i)}\right] &= p\left(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \theta^{(t)}\right) \\ \theta^{(t+1)} &= \underset{\theta}{\operatorname{argmax}}\, E_{Z|X,\theta^{(t)}}\left[\ell_c(\theta \mid X, Z)\right] \end{aligned}$$

$\ell_c(\theta \mid X, Z)$ is the complete log likelihood given the dataset and the unobserved labels. X here is the dataset $\{\mathbf{x}^{(i)}\}_{i=1}^N$ where $\mathbf{x}^{(i)} \in \mathbb{R}^D$. We use one-hot vectors $\{\mathbf{z}^{(i)}\}_{i=1}^N$, where each element of $\mathbf{z}^{(i)}$, $z_k^{(i)}$, can take on values 0 or 1. We will consider when $z_k^{(i)} = 1$ to denote the event that that the i-th data point is generated by the k-th Gaussian mixture. We use θ to represent the mixture mean, covariance, and prior probability (μ_k, Σ_k, π_k) for all k. Assume the covariance matrices are invertible.

2.1 [6 pts] E-step

1. **[3 pts]** Derive the expression of $p\left(\mathbf{x}^{(i)}, z_k^{(i)} = 1 \mid \theta^{(t)}\right)$ and $p\left(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \theta^{(t)}\right)$ in terms of π_j, μ_j, Σ_j and $\mathbf{x}^{(i)}$.

Your Answer

2. **[3 pts]** In the following questions, use $\eta_{k,i}^{(t)}$ to represent $p\left(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \theta^{(t)}\right)$ for simplicity. Let K be the total number of classes.

   Derive the expression for the expected value of the complete log likelihood given X and the current parameters, $E_{Z|X,\theta^{(t)}}\left[\ell_c(\theta^{(t)} \mid X, Z)\right]$ in terms of $\eta_{k,i}^{(t)}$, π_j, Σ_j, $\mu_{\mathbf{k}}$ and $\mathbf{x}^{(i)}$.

Your Answer

2.2 [9 pts] M-step

1. **[3 pts]** Let $N_k^{(t)} = \sum_{i=1}^N \eta_{k,i}^{(t)}$ in the following questions. Show your work to derive:

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^N \eta_{k,i}^{(t)} \mathbf{x}^{(i)}}{N_k^{(t)}}$$

Your Answer

2. **[3 pts]** Show your work to derive:

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{n} \eta_{k,i}^{(t)}(x^{(i)} - \mu_k^{(t+1)})(x^{(i)} - \mu_k^{(t+1)})^T}{N_k^{(t)}}$$

You may want to use the fact that $\frac{d|A|}{dA} = |A|A^{-T}$, where $|A| = det(A)$ is the determinant of A.

Additionally, we have $\frac{dx^T Ax}{dA} = xx^T$ and $\frac{dx^T A^{-1}x}{dA} = -A^{-T}xx^T A^{-T}$.

Your Answer

3. **[3 pts]** Show your work to derive:

$$\pi_k^{(t+1)} = \frac{N_k^{(t)}}{N}$$

Hint: You may want to use Lagrange multipliers with the constraint that all π_k must sum to one.

Your Answer

3 [9 pts] Kernels

3.1 Kernel Computation Cost

1. **[3 pts]** Suppose we have a two-dimensional input space such that the input vector is $\mathbf{x} = [x_1, x_2]^T$. Define the feature mapping $\phi(\mathbf{x}) = [x_1^2, \sqrt{2}x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, 1]^T$. What is the corresponding kernel function, i.e. $k(\mathbf{x}, \mathbf{z})$? Do not leave $\phi(\cdot)$ in your final answer. Simplify your answer to write it using input vectors $\mathbf{x}, \mathbf{z}$. In order to receive full credit, you must show your work.

Your Answer

2. **[3 pts]** Suppose we want to compute the value of the kernel function $k(\mathbf{x}, \mathbf{z})$ from the previous question, on two vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$. How many operations (additions, multiplications, powers) are needed if you map the input vector to the feature space and then perform the dot product on the mapped features? Show your work.

Num

Your Work

3. **[3 pts]** How many operations (additions, multiplications, powers) are needed if you compute through the kernel function you derived in question 1? Show your work.

Num

Your Work

4 [0 pts] (Optional) SVM Decision Boundaries

In this example we will use the following definition of SVM:

$$\begin{aligned}\min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}\cdot\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ & \text{s.t.} \quad \xi_i \geq 0 \quad \forall i = 1,\ldots,n \\ & (\mathbf{w}\cdot\mathbf{x}_i + b)y_i \geq (1-\xi_i) \quad \forall i = 1,\ldots,n.\end{aligned}$$

Notice this is different from the version covered in class since now our regularization hyperparameter λ is no longer multiplied with the $\frac{1}{2}w \cdot w$ but is now C and is multiplied with the $\sum \xi_i$. These are the same it just is the difference of where you would like to focus the SVM's minimization. For example increasing λ in the equations in class would be like decreasing C in the equation above.

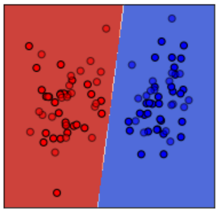
RBF kernel:

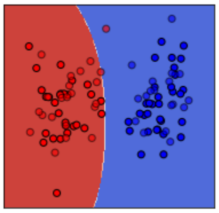
$$K(\mathbf{x},\mathbf{z}) = \exp(-\gamma\|\mathbf{x}-\mathbf{z}\|_2^2)$$

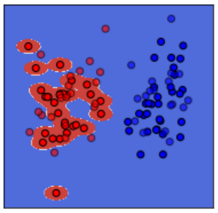
The figure below shows SVM decision boundaries resulting from different combinations of kernels, slack penalties, and RBF bandwidths. Match each plot from the figure below with one of the SVM settings.

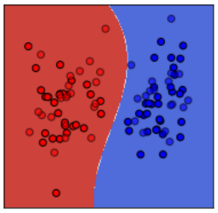
To obtain full credit, you do not need to provide any justification, but only provide the correct pairing.

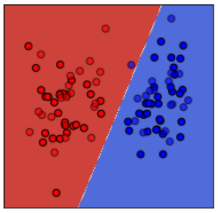
1. Linear kernel, C = 0.1
2. Linear kernel, C = 100
3. RBF kernel, C = 1, $\gamma = 10$
4. RBF kernel, C = 1, $\gamma = 50$
5. RBF kernel, C = 0.1, $\gamma = 0.1$
6. RBF kernel, C = 20, $\gamma = 0.1$

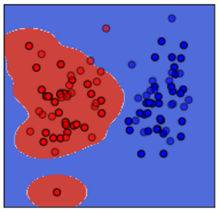

a


b


c


d


e


f

1

2

3

4

5

6

Explanation

5 Programming

The programming portion of this assignment will consist of working on 5 small problems that all pertain to the topics that you will have learned throughout the later half of the semester. The topics are as follows

1. Recommender Systems
2. K-Means Algorithm
3. PCA & Training Gaussian Mixture Models through Expectation-Maximization
4. Kernel Regression
5. (Optional) SVM

Complete the necessary programming portions of the assignment in the handout notebook, `hw10.ipynb`, in either Google Colab (preferred) or Jupyter Notebook. There will additionally be some related written questions that you will have to provide analyses or plots in the questions below.

5.1 [3 pts] Recommender Systems

Complete the recommender systems programming. Afterwards, answer the following question: after running `matrix_factorization_alt_min` with `K=20`, `alpha=0.001`, and `num_epoch=200`, which of the first 10 users (user indices 0 through 9) do you predict would rate *The Lightning Thief* (book index 6) the highest? Which would rate it the lowest? What are these respective ratings? Round them to the 2nd decimal place.

User index with **highest** rating:

Index	Rating

User index with **lowest** rating:

Index	Rating

5.2 [4 pts] K-Means

Complete the K-Means programming problem. Afterwards, provide the plots of the centers that are produced by the K-Means algorithm below for the various K.

K=2

K=5

K=10

5.3 [6 pts] PCA & GMM

1. **[3 pts]** Complete the PCA programming problem. Afterwards, provide **both before and after** plots of applying PCA to the Toy Dataset. Then, include the plot of MNIST zeros and ones after performing PCA on it with $K = 2$.

Toy Dataset Before and After PCA

MNIST Zeros and Ones after PCA (*K*=2)

2. **[3 pts]** Complete the EM programming problem in which you use EM to train a Gaussian Mixture Model. Afterwards, include plots of learning GMM parameter for $K = 2$ on toy datasets one and two. Additionally, include the plots of learning GMM parameters for $K = 2$ and $K = 5$ on the MNIST zeros and ones dataset.

GMM on Toy Datasets 1 and 2 (*K=2* for both)

MNIST Zeros and Ones after PCA (*K=2* for both)

5.4 [15 pts] Kernel Regression

Complete the kernel implementation programming problem and then answer the following questions below.

1. **[4 pts]** Include surface plots for the boxcar kernel with width=2, the RBF kernel with gamma = 0.1, the linear kernel, and the polynomial kernel with d=2, all with $\mathbf{z} = [2, -3]$. Your surface plot should represent the values the function takes on around z with the given hyperparameter settings.

Plot boxcar, width=2

Plot RBF, gamma=0.1

Plot linear

Plot polynomial, d=2

2. **[5 pts]** Include surface plots for the kernel regression with N=2 training points with the boxcar kernel with width=2, the RBF kernel with gamma = 0.1, the linear kernel, the polynomial kernel with d=2, and the polynomial kernel with d=3.

Plot N=2 boxcar, width=2

Plot N=2 RBF, gamma=0.1

Plot N=2 linear

Plot N=2 polynomial, d=2

Plot N=2 polynomial, d=3

3. **[2 pts]** Include surface plots for the kernel regression with N=200 training points with the RBF kernel with gamma = 0.01, 0.1, and 1.

Plot N=200 RBF, gamma=0.01

Plot N=200 RBF, gamma=0.1

Plot N=200 RBF, gamma=1

4. **[1 pts]** Using the empirical results that we see when N=200 and in the plots in question 3, explain the relationship between settings of gamma in the RBF filter and over/underfitting.

Answer

5. **[1 pts]** Include surface plots for the kernel regression with N=200 training points with the linear kernel.

Plot N=200 linear

6. **[1 pts]** For the linear kernel with N=200 training points, why is the prediction surface significantly below the training points?

Answer

7. **[1 pts]** Of the various kernel regression models that were run by the test cases, which model do you believe is the best model for performing regression? How do you know this? Specifically, provide the kernel used and the hyperparameter settings.

Answer

5.5 [0 pts] (Optional) SVM

Complete the SVM programming section be generating plots for the various kernels and hyperparameters that are listed below. Feel free to use helper functions that we've provided and SVM solvers for generating these plots with the appropriate kernel.

1. **[0 pts]** Linear kernel, $K(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T\mathbf{z}$

Num. SV

y=-1:
y=1:

Plot

2. **[0 pts]** Quadratic kernel, $K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T\mathbf{z} + 1)^2$

Num. SV

y=-1:
y=1:

Plot

3. **[0 pts]** RBF kernel with $\gamma = 3$, $K(\mathbf{x}, \mathbf{z}) = \exp(-\gamma\|\mathbf{x} - \mathbf{z}\|_2^2)$

Num. SV	Plot

y=-1:
y=1:	

4. **[0 pts]** RBF kernel with $\gamma = 0.5$, $K(\mathbf{x}, \mathbf{z}) = \exp(-\gamma\|\mathbf{x} - \mathbf{z}\|_2^2)$

Num. SV	Plot

y=-1:
y=1:	

5. **[0 pts]** RBF kernel with $\gamma = 0.1$, $K(\mathbf{x}, \mathbf{z}) = \exp(-\gamma \|\mathbf{x} - \mathbf{z}\|_2^2)$

Num. SV

y=-1:
y=1:

Plot

6 Collaboration Policy

After you have completed all other components of this assignment, report your answers to the following collaboration questions.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details including names of people who helped you and the exact nature of help you received.

2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details including names of people you helped and the exact nature of help you offered.

3. Did you find or come across code that implements any part of this assignment? If so, include full details including the source of the code and how you used it in the assignment.