Homework 2

Part 1

PCA

1. PCA can be explained from two different perspectives. What are the two perspectives explained in class?

2. The first principal direction is the direction in which the projections of the data points have the largest variance in the input space. We use λ_1 to represent the first/largest eigenvalue of the covariance matrix, w_1 to denote the corresponding principal vector/direction (w_1 has unit length i.e., its L2 norm is 1), μ to represent the sample mean, and x to represent a data point. The deviation of x from the mean μ is $x - \mu$.

The forward transform, $y = PCA(x)$, is implemented in sk-learn with "whiten=True".

(1) write down the scalar-projection of the deviation $x - \mu$ in the direction of w_1?

(2) what is the first component of y ?

note: compute it using w_1 , x, μ, and λ_1

(3) assuming y only has one component, then we do inverse transform to recover the input

$$\tilde{x} = PCA^{-1}(y)$$

compute $\tilde{x}$ using μ, y, λ_1 and w_1

(4) assuming x and y have the same number of elements, and we do inverse transform to recover the input

$$\tilde{x} = PCA^{-1}(y)$$

what is the value of $x - \tilde{x}$?

Note: the question asks for a value/number, not equations

(5) For face image generation applications shown in class, what is the major difference between the two methods: eigenface vs. statistical shape model ?

Maximum Likelihood Estimation and NLL loss
(This is a general method to estimate parameters of a PDF using data samples)

3. Maximum Likelihood Estimation when the PDF is an exponential distribution.
We have N i.i.d. (independently and identically distributed) data samples $\{x_1, x_2, x_3, \ldots, x_N\}$ generated from a PDF that is assumed to be an exponential distribution. $x_n \in \mathcal{R}^+ \ for\ n = 1\ to\ N$, which means they are positive scalars. This is the PDF:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & for\ x \geq 0 \\ 0 & otherwise \end{cases}$$

Your task is to build an NLL (negative log likelihood) loss function to estimate the parameter λ of the PDF from the data samples.
(1) write the NLL loss function: it is a function of the parameter λ
(2) take the derivative of the loss with respect to λ, and set the result to 0.
After some calculations, you will obtain an equation about $\lambda =$******
Hint: read NLL in the lecture of GMM

4. Maximum Likelihood Estimation when the PDF is histogram-like.

A histogram-like PDF $f(x)$ is defined on a 1-dimensional (1D) space that is divided into fixed regions/intervals. So, $f(x)$ takes constant value h_i in the i-th region. There are K regions. Thus, $\{h_1, h_2, \dots, h_K\}$ is the set of (unknown) parameters of the PDF. Also, $\sum_{i=1}^{K} h_i \Delta_i = 1$, where Δ_i is the width of the i-th region.

Now, we have a dataset of N samples $\{x_1, x_2, x_3, \dots, x_N\}$, and N_i is the number of samples in the i-th region. The task is to find the best parameters of the PDF using the samples.

(1) write the loss function: it is a function of the parameters
Note: it is a constrained optimization problem, so we need to use the Lagrange multiplier method to convert constrained optimization to unconstrained optimization. Thus, we add $\lambda(\sum_{i=1}^{K} h_i \Delta_i - 1)$ and the NLL together to get the complete loss function, where λ is the Lagrange multiplier.

(2) take the derivative of the loss with respect to h_i, set it to 0, and obtain the best parameters along with the value of λ.

Is Bayes optimal ?

5. Bayes classifier has the minimum classification error assuming we know the true $p(x|y)$ and $p(y)$. However, for many applications, reaching the minimum classification error may not be the best objective. Now, let's consider the application explained in the lecture: there are two classes, class-0 and class-1.

In class-0, patients have aneurysms, but the aneurysms will not rupture

In class-1, patients have aneurysms, and the aneurysms will rupture almost immediately if left untreated, and therefore surgeries will be performed to prolong the life of the patients.

Assume these:

(a) The patients in class-0 will live until the age of 100.
(b) The patients in class-1 will live until the age of 100 after receiving surgeries but will die immediately if left untreated.
(c) Ther risk of the surgery is ε between 0 and 1, e.g., ε=0.01 means there is a 1% chance that a patient may die during surgery.

Consider a patient at the age of 60, if the true class label of a patient is class-0, but this patient is misclassified to class-1, thus, this patient will get an unnecessary surgery and may die with the chance of ε. The average cost for this patient is 40×ε

Consider another patient at the age of 60, if the true class label of a patient is class-1, but this patient is misclassified to class-0, thus, this patient will not get surgery and die almost immediately. The cost of this misclassification is 40 years for this patient.

Now, we have data points $\{x_1, x_2, x_3, \dots, x_N\}$ with true labels $\{y_1, y_2, y_3, \dots, y_N\}$, and x_n is the aneurysm feature of the patient-n. The current age of the patient-n is t_n. We have this cost table for each patient:

True label y_n	Predicted Label $\hat{y}_n$	Cost for the patient-n
0	0	0
1	1	0
0	1	(100-t_n)×ε
1	0	100-t_n

$\hat{y}_n = f(x_n; w)$ is a classification model with internal parameter w

The value of $\hat{y}_n$ is a real number between 0 and 1.

Your task: design a differentiable loss $L_n(w)$ that is the cost of making a wrong classification on x_n.

“differentiable” means $\frac{\partial L_n}{\partial \hat{y}_n}$ exists, so that $\frac{\partial L_n}{\partial w}$ exists.

Part 2

Complete the task in H2P2T1.ipynb and H2P2T2.ipynb

Note: It is very time consuming to fit a GMM to high dimensional data, and therefore PCA + GMM is the "standard" approach.

Grading: the number of points

	Undergraduate Student	Graduate Student
1 (PCA)	1	1
2 (PCA)	5	5
3 (NLL)	4	4
4 (NLL)	N.A.	5 bonus points
5 (loss)	10	10
H1P2T1	15	15
H2P2T2	15	15
Total number of points	50 +5	50 + 5

Extra Reading

PCA is widely used in many applications. Do a google scholar search with PCA + some field, e.g., PCA +bioinformatics or PCA + finance, you will find relevant papers.

https://www.nature.com/articles/s41467-018-04608-8

There are many variants of PCA, such as sparse PCA and kernel PCA that are implemented in sk-learn.

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.72.7798&rep=rep1&type=pdf

https://www.di.ens.fr/sierra/pdfs/icml09.pdf

https://www.di.ens.fr/~fbach/sspca_AISTATS2010.pdf

Which one is good for your application? Test different algorithms and find the best. Remember that machine learning is more like an experimental science: you need to run lots of experiments.