

Full Name: Yi Ren

Number: 002269013

Signature: Yi Ren

Bishop's University

CS 509 - Pattern Recognition

Final exam

Fall 2020

*The exam is worth 100 points total and has five problems. Be sure to read the whole exam before attempting any of it. This exam is open book since it is online. **Please note that this exam is individual not in group.** You have 240 minutes to complete the exam. Use the provided white space to respond to each question. Please, write legibly.*

Submission: *All submissions must be performed through Moodle. The exam is from 9:00 am to 1:00 pm. Any submission with errors will get attributed the grade zero. No submission by email will be accepted. Please note that any late **submission or update** will involve a penalization of 10% for 30 minutes late submission, 20% of one hour, etc. **The penalization will be applied for submissions or updates after 1:30 pm.** It will be of your interest to be efficient in managing your time.*

Problem 1: “Recall” Questions

1. Suppose that you have a dataset that you do not know which probability density function to use, and none of the known probabilities densities functions fit to your dataset. What solution can you suggest for the choice of the pdf for your dataset?
 - a . As none of the known PDF fit to my data set, parameter models are not applicable, because they need to know the specific form of density function.
 - b . So it's time to use Non-parameter methods to solve this problem, such as KDE or KNN Estimation.
 - c . Non-parameter methods may have better performance on many practical problems, because practical problems are more likely to have multimodal densities.

2. What quantity is PCA maximizing during dimension reduction?

a . Principal Component Analysis (PCA), is mainly focus on maximizing variance when dimension reduction occurs.

b . When we are processing dimension reduction, there could be redundant features in our dataset. So we need to remove redundancy by comparing variance between these features, low variance means high redundancy.

3. What does ISOMAP focus on preserving during dimension reduction?

Isometric feature mapping(ISOMAP) is a non-linear dimension reduction method that focus on preserving the geodesic distances for the low dimension.

4. Highlight two limitations of K-Means clustering. Suggest a clustering method that overcome these limitations.

a . K-means need to specify the number of cluster, which can be difficult, because we do not know how many clusters according to the given dataset.

b . K-means has problem when we have different size or density of clusters.

c . K-means has problem dealing with outliers, because K-means is very sensitive to the outliers.

d . We can use hierarchical clustering to overcome these limitations.

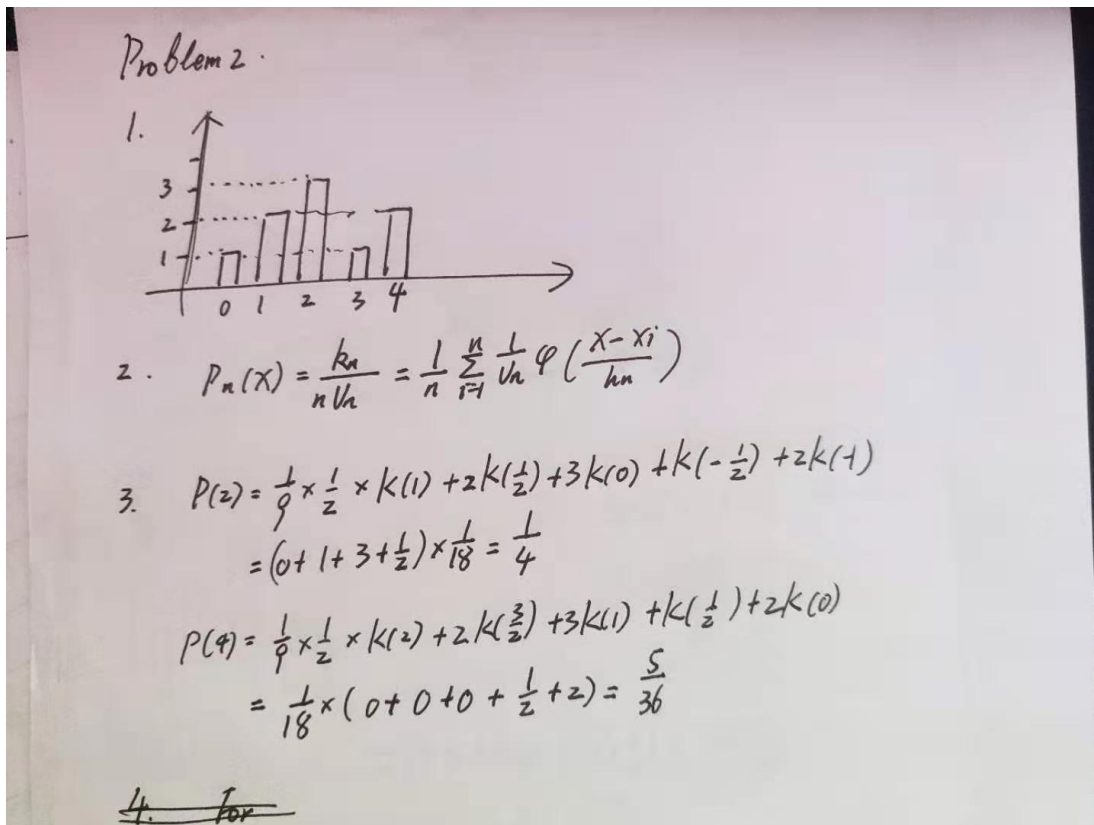
5. What are the limitations of neural networks that were overcome by convolutional neural networks (CNN)? Justify your answer.

CNN is a special kind of multi-layer neural network. Neural networks(NN) is too expensive when solving practical problems, while CNN can extract some features from certain layers, could provide better performance.

Problem 2: Non-Parametric Methods

You are given a dataset $D = \{0,1,1,2,2,2,3,4,4\}$. Using techniques from non-parametric density estimation, answer the following questions:

1 - 3:



4. Suppose that we are doing classification of d -dimensional data using k -nearest neighbor method, show that the “effective” number of parameters used by k -nearest neighbor is in the order of $N=k$, where N is the number of training examples.

Hint: Think of the cases where $k = 1$ and $k = N$.

For k -nearest neighbor classification, we partitioned the dataset to $N=k$ disjoint regions, and if our sample falls in one of the regions then we claim the class of the sample to the class of that region.

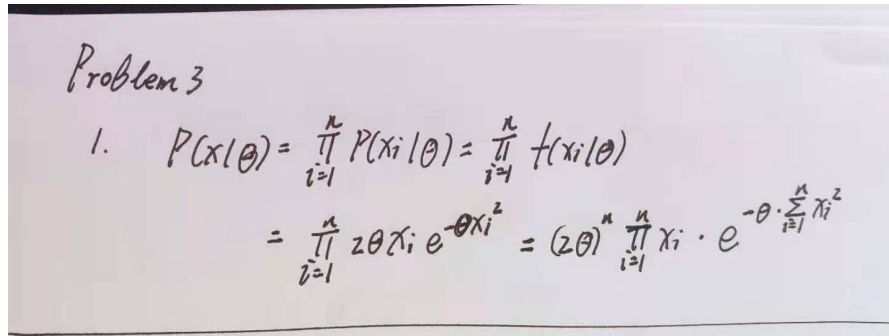
Therefore the “effective parameters” that affect the classification in this case is the number of regions, i.e. $N=k$.

Problem 3: MLE and MAP

Let $X = \{x_1, \dots, x_n\}$ be an identically, independent distributed (iid) sample drawn from a Rayleigh distribution with an unknown θ defined as follows:

$$f(x|\theta) = 2\theta x e^{-\theta x^2}$$

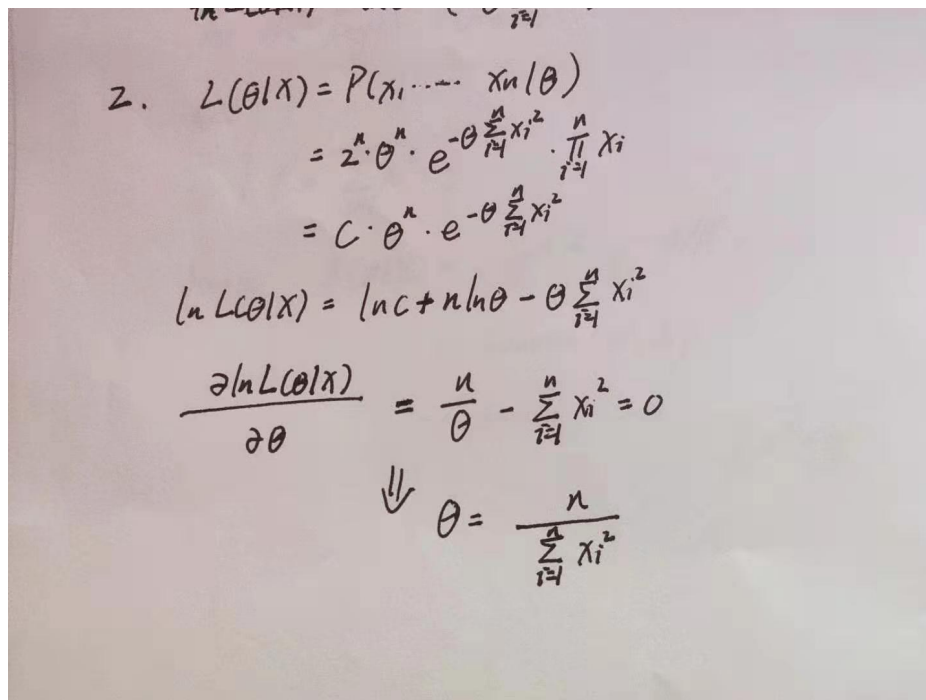
1. What is the joint probability distribution $P(X|\theta)$ of the sample?



Problem 3

1.
$$P(X|\theta) = \prod_{i=1}^n P(x_i|\theta) = \prod_{i=1}^n f(x_i|\theta)$$
$$= \prod_{i=1}^n 2\theta x_i e^{-\theta x_i^2} = (2\theta)^n \prod_{i=1}^n x_i \cdot e^{-\theta \sum_{i=1}^n x_i^2}$$

2. What is the maximum likelihood estimation (MLE) of the parameter θ ?



2.
$$L(\theta|X) = P(x_1, \dots, x_n|\theta)$$
$$= 2^n \cdot \theta^n \cdot e^{-\theta \sum_{i=1}^n x_i^2} \cdot \prod_{i=1}^n x_i$$
$$= C \cdot \theta^n \cdot e^{-\theta \sum_{i=1}^n x_i^2}$$
$$\ln L(\theta|X) = \ln C + n \ln \theta - \theta \sum_{i=1}^n x_i^2$$
$$\frac{\partial \ln L(\theta|X)}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^n x_i^2 = 0$$
$$\Downarrow \theta = \frac{n}{\sum_{i=1}^n x_i^2}$$

3. Let θ has a prior *Gamma* distribution given by

$$p(\theta) = \frac{\theta^{\alpha-1} e^{-\frac{\theta}{\beta}}}{\Gamma(\alpha) \beta^\alpha}$$

Where α and β are known.

- a. Please write down the objective function of maximum a posteriori (MAP) estimation of the parameter θ .

Handwritten derivation of the posterior distribution $p(\theta|x)$:

$$\begin{aligned}
 p(\theta|x) &= \frac{p(x|\theta) p(\theta)}{p(x)} = C \cdot p(x|\theta) \cdot p(\theta) \\
 &= C \cdot C' \theta^n \cdot e^{-\theta \sum_{i=1}^n x_i^2} \cdot p(\theta) \\
 &= C'' \theta^n \cdot e^{-\theta \sum_{i=1}^n x_i^2} \cdot p(\theta) \\
 &= C'' \theta^n \cdot e^{-\theta \sum_{i=1}^n x_i^2} \cdot \theta^{\alpha-1} \cdot e^{-\frac{\theta}{\beta}} \\
 &= C'' \theta^{n+\alpha-1} \cdot e^{-\theta (\sum_{i=1}^n x_i^2 + \frac{1}{\beta})}
 \end{aligned}$$

$$(b). p(\theta|x) = C'' \theta^{n+\alpha-1} \cdot e^{-\theta (\sum_{i=1}^n x_i^2 + \frac{1}{\beta})}$$

- b. Find the posterior and the Bayes Estimator.

Handwritten derivation of the Bayes Estimator:

For the Bayes Estimator:

Let $\alpha' = n + \alpha$

$$\beta' = \frac{1}{\sum_{i=1}^n x_i^2 + \frac{1}{\beta}}$$

Hence $p(\theta|x) = C'' \theta^{\alpha'-1} \cdot e^{-\theta/\beta'}$

$$\sim \text{Gamma}(\alpha', \beta')$$

$$\sim \text{Gamma}(n + \alpha, \frac{1}{\sum_{i=1}^n x_i^2 + \frac{1}{\beta}})$$

- c. Estimate θ_{MAP} that maximizes the objective function in a) using MAP estimation.

$$\begin{aligned}
 C. \quad P(\theta|x) &= C'' \theta^{n+d-1} \cdot e^{-\theta \left(\sum_{i=1}^n x_i^2 + \frac{1}{\beta} \right)} \\
 \ln(P(\theta|x)) &= \ln C'' + (n+d-1) \ln \theta - \theta \left(\sum_{i=1}^n x_i^2 + \frac{1}{\beta} \right) \\
 \frac{\partial \ln P(\theta|x)}{\partial \theta} &= \frac{n+d-1}{\theta} - \sum_{i=1}^n x_i^2 - \frac{1}{\beta} = 0 \\
 \Downarrow \\
 \theta &= \frac{n+d-1}{\sum_{i=1}^n x_i^2 + \frac{1}{\beta}}
 \end{aligned}$$

Problem 4: EM algorithm for a Gaussian mixture

Consider a special case of a Gaussian mixture model in which the covariance matrices Σ_k of the components are all constrained to be diagonals, such as $\Sigma_k = \sigma^2 I$, where σ^2 is the variance and I is a $d \times d$ identity matrix. Derive the EM equations for maximizing the likelihood function under such a model.

The Gaussian distribution is defined for $x \in \mathbb{R}^d$ as follows:

$$p(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right]$$

Hints:

1. For a symmetric matrix $A \in \mathbb{R}^{n \times n}$ and a vector $x \in \mathbb{R}^n$,

$$\frac{\partial}{\partial x} [x^T A x] = 2Ax$$

2. For $\Sigma_k = \sigma^2 I$ we have:

$$p(x|\mu_k, \sigma^2 I) = \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^d} \exp \left[-\frac{1}{2} \frac{(x - \mu_k)^T (x - \mu_k)}{\sigma^2} \right]$$

Problem 4.

For this case $\Sigma_k = \sigma^2 I$

$$L(\theta) = \ln L(\theta|x) = \sum_{i=1}^n \sum_{k=1}^K r(z_{ik}) [\ln d_k + \ln p(x_i | \mu_k, \sigma^2 I)]$$

where

$$r(z_{ik}) = \frac{d_k p(x_i | \mu_k, \sigma^2 I)}{\sum_{k=1}^K d_k p(x_i | \mu_k, \sigma^2 I)}$$

From Hint 2:

$$\begin{aligned} L(\theta) &= \sum_{i=1}^n \sum_{k=1}^K r(z_{ik}) [\ln d_k + \ln \left(\frac{1}{(2\pi)^{d/2} \sigma^d} \exp \left[-\frac{1}{2\sigma^2} (x_i - \mu_k)^T (x_i - \mu_k) \right] \right)] \\ &= \sum_{i=1}^n \sum_{k=1}^K r(z_{ik}) \left[\ln d_k - \frac{d}{2} \ln 2\pi - \frac{d}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (x_i - \mu_k)^T (x_i - \mu_k) \right] \end{aligned}$$

$$\frac{\partial L(\theta)}{\partial \mu_k} = 0$$

$$\Downarrow \sum_{i=1}^n r(z_{ik}) \left[0 - 0 - 0 - \frac{1}{2\sigma^2} 2 (x_i - \mu_k) \right] = 0$$

$$\boxed{\mu_k = \frac{\sum_{i=1}^n r(z_{ik}) x_i}{\sum_{i=1}^n r(z_{ik})}}$$

$$\frac{\partial L(\theta)}{\partial \sigma^2} = 0$$

$$\Downarrow \sum_{i=1}^n \sum_{k=1}^K r(z_{ik}) \left[0 - 0 - \frac{d}{2\sigma^2} + \frac{1}{2\sigma^4} (x_i - \mu_k)^T (x_i - \mu_k) \right] = 0$$

$$\boxed{\sigma^2 = \frac{1}{nd} \cdot \sum_{i=1}^n \sum_{k=1}^K r(z_{ik}) (x_i - \mu_k)^T (x_i - \mu_k)}$$

$$\frac{\partial (L(\theta) + \lambda (\sum_{k=1}^K d_k - 1))}{\partial d_k} = 0$$

$$\Downarrow \boxed{d_k = \frac{\sum_{i=1}^n r(z_{ik})}{n}}$$

Problem 5: Application of Maximum Likelihood Estimation

An astronomer measures the photon flux from a distant star using a very sensitive instrument that counts single photons. After one minute of observation, the instrument has collected \hat{r} photons. One can assume that the photon counts, \hat{r} , are distributed according to the Poisson distribution. The astronomer wishes to determine λ , the emission rate of the source.

The Poisson distribution is defined as follows:

$$f(x|\lambda) = P_{\lambda}(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Problem 5.

① Unknown Parameter: λ

Data: the photon counts \hat{r}

likelihood: $L(\lambda) = P(\hat{r}|\lambda) = e^{-\lambda} \cdot \frac{\lambda^{\hat{r}}}{\hat{r}!}$

② Let $\begin{cases} \lambda = 10 \\ \hat{r} = 15 \end{cases} \quad P(\hat{r}=15 | \lambda=10) = \frac{e^{-10} \times 10^{15}}{15!} = 0.0347$

③ $\ln L(\lambda) = \ln e^{-\lambda} \frac{\lambda^{\hat{r}}}{\hat{r}!} = -\lambda + \hat{r} \ln \lambda - \ln(\hat{r}!)$

Let $\frac{\partial \ln L(\lambda)}{\partial \lambda} = 0$

that is: $-1 + \hat{r} \frac{1}{\lambda} = 0$

\Downarrow
 $\hat{\lambda} = \hat{r} = 15$

So $\hat{\lambda}$ is 15 when given $\hat{r}=15$

