Professor: *Dr. Mohammed Ayoub Alaoui Mhamdi*

Full Name:

Number:

Signature:

Bishop's University

CS 509 - Pattern Recognition

Midterm exam

Fall 2020

The exam is worth 60 points total and has three 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 150 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 11:30 am and the submission system on Moodle will close at 11:45 am. Any submission with errors will get attributed the grade zero.* ***No submission by email will be accepted****.*

Problem 1: "Recall" Questions (20 points)

1. Describe Bayes Decision Rule. (**4 points**)

   Answer

   Decision making relies on both the priors and the likelihoods and Bayes Decision Rule combines them to achieve the minimum probability of error.

2. What is the difference between the likelihood and the posterior probability? **(4 points)**

   Answer

The likelihood is the probability of the observations given a state of nature. Meanwhile, the posterior probability is the probability of a state of nature given observations, and it is proportional to the production of the likelihood with the prior probability.

3. What is the fundamental difference between maximum Likelihood parameter estimation and Bayesian parameter estimation? **(4 points)**

   Answer

   - For infinite amounts of data, the solutions converge. However, limited data is always a problem.
   - If prior information is reliable, a Bayesian estimate can be superior.
   - Bayesian estimates for uniform priors are similar to an ML solution.
   - If $p(\theta|D)$ is broad or asymmetric around the true value, the approaches are likely to produce different solutions.

4. What is the difference between maximum Likelihood parameter estimation and maximum *a posteriori* (MAP) estimation? When they can meet? **(4 points)**

   Answer

The maximum *a posteriori* (MAP) is proportional to the product of the maximum likelihood with the prior of a given state of nature. That is the prior plays an important role to find the maximum *a posteriori* for a sample. The MAP and the maximum likelihood meet when the prior is uniform or flat.

5. Suppose we have built a classifier on multiple features. What do we do if one of the features is not measurable for a particular case? **(4 points)**

   Answer

   The possible answers are:

   - **Ignore the features attributing the value zero to it,**
   - **Replace it with mean value of the other features,**
   - **Marginalize,**
   - **Use EM algorithm.**

Problem 2: Parametric models (30 points)

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

$$f(x|\theta) = \begin{cases} \frac{4}{\sqrt{\pi}} \theta^{\frac{3}{2}} x^2 e^{-\theta x^2}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

1. What is the joint probability distribution $P(X|\theta)$ of the sample? **(3 points)**

$$P(X|\,\theta) = \prod_{i=1}^{n} P(x_i|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Solution

$$P(X|\,\theta) = \prod_{i=1}^{n} P(x_i|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \frac{4}{\sqrt{\pi}} \theta^{\frac{3}{2}} x_i^2 e^{-\theta x_i^2} = \left(\frac{4}{\sqrt{\pi}}\right)^n \theta^{\frac{3n}{2}} \prod_{i=1}^{n} x_i^2 e^{-\theta x_i^2}$$

$$= \left(\frac{4}{\sqrt{\pi}}\right)^n \theta^{\frac{3n}{2}} e^{-\theta \sum_{i=1}^{n} x_i^2} \prod_{i=1}^{n} x_i^2$$

2. What is the maximum likelihood (ML) estimation of the parameter θ? **(7.5 points)**

Solution

From 1) we have

$$L(\theta|X) = P(X|\,\theta) = P(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} P(x_i|\theta) = \left(\frac{4}{\sqrt{\pi}}\right)^n \theta^{\frac{3n}{2}} e^{-\theta \sum_{i=1}^{n} x_i^2} \prod_{i=1}^{n} x_i^2$$

We can put the terms that does not depend on θ in a constant.

$$L(\theta|X) = c\theta^{\frac{3n}{2}} e^{-\theta \sum_{i=1}^{n} x_i^2}$$

Since we have terms in power and exponent, it is better to compute the log of likelihood.

$$\ln L(\theta|X) = \ln\left(c\theta^{\frac{3n}{2}} e^{-\theta \sum_{i=1}^{n} x_i^2}\right) = \ln c + \ln \theta^{\frac{3}{2}} - \theta \sum_{i=1}^{n} x_i^2$$

$$= \ln c + \frac{3}{2}\ln\theta - \theta\sum_{i=1}^{n} x_i^2$$

$$\frac{\partial \ln L(\theta|X)}{\partial\theta} = 0 \Leftrightarrow \frac{\partial \ln\left(\ln c + \frac{3}{2}\ln\theta - \theta\sum_{i=1}^{n} x_i^2\right)}{\partial\theta} = 0$$

$$0 + \frac{3}{2\theta} - \sum_{i=1}^{n} x_i^2 = 0 \Leftrightarrow \frac{3}{2\theta} = \sum_{i=1}^{n} x_i^2$$

$$\Rightarrow \theta = \frac{3}{2\sum_{i=1}^{n} x_i^2}$$

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. **(4.5 points)**

Solution

$$P(\theta|X) = \frac{P(X|\,\theta)P(\theta)}{P(X)} = CP(X|\,\theta)P(\theta) = Cc\theta^{\frac{3n}{2}}e^{-\theta\sum_{i=1}^{n} x_i^2}\frac{\theta^{\alpha-1}e^{\frac{-\theta}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}$$

$$= c'\theta^{\frac{3n}{2}}e^{-\theta\sum_{i=1}^{n} x_i^2}\frac{\theta^{\alpha-1}e^{\frac{-\theta}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}$$

It is not mandatory to develop it, but it will help you in b)

b. Find the posterior and the Bayes Estimator. **(7.5 points)**

Solution

For the posterior

$$P(\theta|X) = c'\theta^{\frac{3n}{2}}e^{-\theta\sum_{i=1}^{n} x_i^2}\frac{\theta^{\alpha-1}e^{\frac{-\theta}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}$$

$$P(\theta|X) = c''\theta^{\frac{3n}{2}} e^{-\theta\sum_{i=1}^{n} x_i^2}\theta^{\alpha-1}e^{\frac{-\theta}{\beta}} = c''\theta^{\frac{3n}{2}+\alpha-1}e^{-\theta\left(\frac{1}{\beta}+\sum_{i=1}^{n} x_i^2\right)}$$

The bayes Estimator

Suppose $\alpha' = \frac{3n}{2} + \alpha$ and $\beta' = \frac{1}{\frac{1}{\beta}+\sum_{i=1}^{n} x_i^2}$. Thus,

$$P(\theta|X) = c''\theta^{\alpha'-1}e^{-\frac{\theta}{\beta'}} \sim Gamma(\alpha', \beta') \sim Gamma\left(\frac{3n}{2} + \alpha, \frac{1}{\frac{1}{\beta} + \sum_{i=1}^{n} x_i^2}\right)$$

c. Estimate θ_{MAP} that maximizes the objective function in a) using MAP estimation. **(7.5 points)**

Solution

$$P(\theta|X) = c''\theta^{\frac{3n}{2}+\alpha-1}e^{-\theta\left(\frac{1}{\beta}+\sum_{i=1}^{n} x_i^2\right)}$$

$$\ln P(\theta|X) = \ln\left(c''\theta^{\frac{3n}{2}+\alpha-1}e^{-\theta\left(\frac{1}{\beta}+\sum_{i=1}^{n} x_i^2\right)}\right)$$

$$= \ln c'' + \ln\theta^{\frac{3n}{2}+\alpha-1} - \theta\left(\frac{1}{\beta} + \sum_{i=1}^{n} x_i^2\right)$$

$$\ln P(\theta|X) = \ln c'' + \left(\frac{3n}{2} + \alpha - 1\right)\ln\theta - \theta\left(\frac{1}{\beta} + \sum_{i=1}^{n} x_i^2\right)$$

$$\frac{\partial \ln P(\theta|X)}{\partial\theta} = 0 \Leftrightarrow \frac{\partial \ln\left(\ln c'' + \left(\frac{3n}{2} + \alpha - 1\right)\ln\theta - \theta\left(\frac{1}{\beta} + \sum_{i=1}^{n} x_i^2\right)\right)}{\partial\theta} = 0$$

$$0 + \frac{\frac{3n}{2} + \alpha - 1}{\theta} - \left(\frac{1}{\beta} + \sum_{i=1}^{n} x_i^2\right) = 0$$

$$\Rightarrow \theta = \frac{\frac{3n}{2} + \alpha - 1}{\frac{1}{\beta} + \sum_{i=1}^{n} x_i^2}$$

Problem 5: EM algorithm for a Gaussian mixture (10 points)

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 $\Sigma_k = \sigma_k^2 I$, where σ_k^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. Detail E-Step and M-Step.

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 $\Sigma_k = \sigma_k^2 I$ we have:

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

Solution

For the case of $\Sigma_k = \sigma_k^2 I$, we have

$$\mathcal{L}(\Theta) = L(\Theta|\mathrm{X}) = \sum_{i=1}^{N}\sum_{k=1}^{K}\gamma(z_{ik})[\ln \alpha_k + \ln p(x_i|\mu_k, \sigma_k^2 I)]$$

Where

$$\gamma(z_{ik}) = \frac{\alpha_k p(x_i|\mu_k, \sigma_k^2 I)}{\sum_{j=1}^{K}\alpha_j p\left(x_i|\mu_j, \sigma_j^2 I\right)}$$

$$\Rightarrow \mathcal{L}(\Theta) = \sum_{i=1}^{N}\sum_{k=1}^{K}\gamma(z_{ik})\left[\ln \alpha_k + \ln\left(\frac{1}{(2\pi)^{\frac{d}{2}}\sigma_k^d} exp\left[-\frac{1}{2}\frac{(x_i-\mu_k)^T(x_i-\mu_k)}{\sigma_k^2}\right]\right)\right]$$

By expanding the previous equation of $\mathcal{L}(\Theta)$ we obtain:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N}\sum_{k=1}^{K}\gamma(z_{ik})\left[\ln \alpha_k - \frac{d}{2}\ln 2\pi - \ln \sigma_k^d - \frac{1}{2}\frac{(x_i-\mu_k)^T(x_i-\mu_k)}{\sigma_k^2}\right]$$

Since $\sigma_k^d = (\sigma_k^2)^{\frac{d}{2}}$, we obtain

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N}\sum_{k=1}^{K}\gamma(z_{ik})\left[\ln \alpha_k - \frac{d}{2}\ln 2\pi - \frac{d}{2}\ln \sigma_k^2 - \frac{1}{2}\frac{(x_i-\mu_k)^T(x_i-\mu_k)}{\sigma_k^2}\right]$$

To find the maximum likelihood, we differentiate $\mathcal{L}(\Theta)$ with respect to parameters of interest and set the result to zero. Firstly, we differentiate $\mathcal{L}(\Theta)$ w.r.t μ_k and relying on the hint 1)

$$\frac{\partial \mathcal{L}(\Theta)}{\partial \mu_k} = \sum_{i=1}^{N}\gamma(z_{ik})\left[0 - 0 - 0 - \frac{1}{2\sigma_k^2}2(x_i-\mu_k)\right] = 0$$

$$\Rightarrow \sum_{i=1}^{N}\gamma(z_{ik})\left[\frac{(x_i-\mu_k)}{\sigma_k^2}\right] = 0$$

$$\Rightarrow \frac{1}{\sigma_k^2}\sum_{i=1}^{N}\gamma(z_{ik})(x_i-\mu_k) = 0$$

$$\Rightarrow \sum_{i=1}^{N}\gamma(z_{ik})(x_i-\mu_k) = 0$$

$$\Rightarrow \sum_{i=1}^{N}\gamma(z_{ik})x_i - \sum_{i=1}^{N}\gamma(z_{ik})\mu_k = 0$$

$$\Rightarrow \sum_{i=1}^{N}\gamma(z_{ik})x_i - \mu_k\sum_{i=1}^{N}\gamma(z_{ik}) = 0$$

$$\mu_k = \frac{\sum_{i=0}^{N}\gamma(z_{ik})x_i}{\sum_{i=1}^{N}\gamma(z_{ik})}$$

We differentiate $\mathcal{L}(\Theta)$ w.r.t σ_k^2

$$\frac{\partial \mathcal{L}(\Theta)}{\partial \sigma_k^2} = \sum_{i=1}^{N}\gamma(z_{ik})\left[0 - 0 - \frac{d}{2\sigma_k^2} + \frac{1}{2(\sigma_k^2)^2}(x_i-\mu_k)^T(x_i-\mu_k)\right] = 0$$

$$\Rightarrow \sum_{i=1}^{N} \gamma(z_{ik}) \left[-\sigma_k^2 + \frac{1}{d}(x_i - \mu_k)^T (x_i - \mu_k) \right] = 0$$

$$\Rightarrow \sigma_k^2 \sum_{i=1}^{N} \gamma(z_{ik}) = \frac{1}{d} \sum_{i=1}^{N} \gamma(z_{ik})(x_i - \mu_k)^T (x_i - \mu_k)$$

Since $(x_i - \mu_k)^T (x_i - \mu_k) = \|x_i - \mu_k\|^2$, we obtain:

$$\sigma_k^2 = \frac{\sum_{i=1}^{N} \gamma(z_{ik}) \|x_i - \mu_k\|^2}{d \sum_{i=1}^{N} \gamma(z_{ik})}$$

We also need to maximize $\mathcal{L}(\Theta)$ with respect to the mixing coefficients α_k. For that purpose, we introduce a Lagrange multiplier to enforce the constraint $\sum_k \alpha_k = 1$.

$$\mathcal{L}(\Theta) + \lambda \left(\sum_{k=1}^{K} \alpha_k - 1 \right)$$

Maximizing it yields:

$$\frac{\partial \left(\mathcal{L}(\Theta) + \lambda (\sum_{k=1}^{K} \alpha_k - 1) \right)}{\partial \alpha_k} = 0$$

$$0 = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(z_{ik}) + \lambda$$

After multiplying both sides by α_k and summing over k, we get

$$\lambda = -n$$

Eliminate λ and rearrange to obtain:

$$\alpha_k = \frac{N_k}{N}$$

Given a GMM, the goal is to maximize the likelihood function with respect to the parameters (μ_k, σ_k^2, and α_k).

1. Initialize μ_k, σ_k^2, and α_k. Evaluate the initial value of the log-likelihood.

2. E-Step: Evaluate the responsibilities using the current parameter values:

$$\gamma(z_{ik}) = \frac{\alpha_k p(x_i|\mu_k, \sigma_k^2 I)}{\sum_{j=1}^{K} \alpha_j p\left(x_i|\mu_j, \sigma_j^2 I\right)}$$

3. M-Step Update the parameters using the current responsibilities

$$\mu_k^{new} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(z_{ik}) x_i$$

$$\sigma_k^{2\,new} = \frac{\sum_{i=1}^{N} \gamma(z_{ik}) \|x_i - \mu_k^{new}\|^2}{d N_k}$$

$$\alpha_k^{new} = \frac{N_k}{N}$$

Where

$$N_k = \sum_{i=1}^{N} \gamma(z_{ik})$$