Middle East Technical University
Department of Computer Engineering

CENG 564
PATTERN RECOGNITION
Fall 2018
TAKE-HOME EXAM 2

Part 1: Bayesian Decision Theory

1) Let X be a one-dimensional feature (i.e. a scalar). Suppose that X is to be used to decide between two states of nature ω_1 and ω_1. Both $P(X|\omega_1)$ and $P(X|\omega_2)$ have a Gaussian distribution, with mean 0 and variance 3, and mean 2 and variance 1 respectively.

(a) Sketch the two densities on the same graph as a function of X.

(b) What is the likelihood ratio ?

(c) Suppose that $P(\omega_1) = P(\omega_1) = 0.5$, and $\lambda_{11} = \lambda_{22} = 0$, $\lambda_{12} = \lambda_{21} = 1$. Find the Bayes decision rule and the probability of error. Show the boundary on the sketch in part (a).

2) Consider a two-class classification problem given by the following class conditional densities:

$$P(x \mid w_1) = 0.6P_1(x) + 0.4P_2(x)$$
$$P(x \mid w_2) = P_2(x)$$

where

$$P_1(x) = p^x(1-p)^{\{1-x\}}$$
$$P_2(x) = q^x(1-q)^{\{1-x\}}$$

are Bernoulli distributions with $p = 1 - q = 0.6$

a) Find and plot $P(w_1 \mid x)$ when $P(w_2) = 2P(w_1)$

b) Determine a decision strategy and classify the sample at $x = 0$.

c) What is the risk of deciding on w_1, $R(w_1 \mid x = 1)$ when you have zero-one loss, i.e. $\lambda_{11} = \lambda_{22} = 0$, $\lambda_{12} = \lambda_{21} = 1$.

3) Suppose we have three categories in two dimensions with the following distributions

$$P(X\,|\omega_1) \sim N(\mathbf{0},\ \boldsymbol{I})$$
$$P(X\,|\omega_2) \sim N\left(\begin{bmatrix}1\\1\end{bmatrix},\ \boldsymbol{I}\right)$$
$$P(X\,|\omega_3) \sim 0.5\,N\left(\begin{bmatrix}0.5\\0.5\end{bmatrix},\ \boldsymbol{I}\right) + 0.5\,N\left(\begin{bmatrix}-0.5\\0.5\end{bmatrix},\ \boldsymbol{I}\right)$$
$$P(\omega_i) = {}^{1}/_{3}\,, \quad i = 1,2,3$$

(a) By explicit calculating of the posterior probabilities, classify the point $\mathbf{x} = \begin{bmatrix}0.3\\0.3\end{bmatrix}$ for minimum probability error.

(b) Suppose that the for a particular test point the first feature is missing. Classify $\mathbf{x} = \begin{bmatrix}*\\0.3\end{bmatrix}$ i.e. how likely the given point belongs to the three categories.

4) Given a two-class 1D classification problem where the conditional densities are Cauchy distribution:

$$p(x|\omega_i) = \frac{1}{\pi b} \cdot \frac{1}{1 + (\frac{x - a_i}{b})^2} \quad , \quad i = 1, 2\,.$$

(a) Find the minimum error decision boundary and confirm that it doesn't depend on b.

(b) Prove that the minimum probability of error is determined by :

$$P(error) = \frac{1}{2} - \frac{1}{\pi}\, tan^{-1} \left|\frac{a_1 - a_2}{2b}\right|$$

Part 2: Maximum Likelihood & Bayesian Estimation

5) Error functions are key components of machine learning and pattern recognition algorithms, and sum-of-squares error function as an error function is frequently. In this question you will find out the relationship between maximum likelihood and implement it as classifier.

(a) What is the relationship between maximum likelihood and least squares?

(b) Implement least square for classification and compute the sum of squared errors by following instructions inside "*the2_leastSquares.m*" file.

6) Let $\boldsymbol{x}$ be a *d*-dimensional binary vector with a multivariate Bernoulli distribution given by where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)^t$ is an unknown parameter vector, θ_i being the probability that x_i = 1. Show the MLE for $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x}_k$$

Check your solution twice to not to find the likelihood fort he univariate case.

7) Prove that the MLE for multivariate Gaussian distribution for the parameter $\boldsymbol{\Sigma}$ is given by

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n} (\boldsymbol{x}_k - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_k - \widehat{\boldsymbol{\mu}})^T$$

Show and explain your steps in a detailed manner where necessary.

8) Assume we have training data from a Gaussian distribution of known covariance $\boldsymbol{\Sigma}$ but unknown $\boldsymbol{\mu}$. Now suppose that, this mean itself is random and characterized by Gaussian density having mean $\boldsymbol{\mu_0}$ and covariance $\boldsymbol{\Sigma_0}$.

(a) What is the MAP estimator for $\boldsymbol{\mu}$ **?**

(b) Suppose we transform our coordinates by a linear transform $\boldsymbol{x}' = \boldsymbol{Ax}$, for non-singular matrix $\boldsymbol{A}$. Determine whether your MAP estimator gives the appropriate estimate fort he transformed mean $\boldsymbol{\mu}'$. Explain.

Part 3: Bias & Estimators

9) Suppose we employ a novel method for estimating the mean of a data set $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$:

(a) Suppose we assign $\hat{\mu} = \boldsymbol{x}_1$. Show that this estimation is unbiased.

(b) State why the estimate in part (a) is undesirable by finding the variance of the ML estimate of the $\hat{\mu}$?

(c) Suppose in this part our samples in *D* are from 1D, then we assign;

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \mu)^2$$

where $\mu = \frac{1}{n}\sum_{k=1}^{n} x_k$. Show that this estimator is unbiased. *(Hint: var(x) = E[x²] – [E[x]]²)*

10) Consider that we have two datasets with;

- Gaussian Distribution $\sim N(\mu,\ \sigma^2)$

- Uniform Distribution $\sim \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$
- Cauchy Distribution $\sim \frac{1}{\pi[1+(x-\mu)^2]} \quad , -\infty < x < \infty$

All three distributions are symmetric about μ , and in fact the Cauchy distribution is bell-shaped but with much heavier tails than the normal curve. The uniform has no tails. The very important thing here is that the best estimator for μ depends crucially on which distrubtion is being sampled.

Compare each the following estimators for the mean i.e. μ, by stating whether they are a good or bad estimator for each of the given distributions ;

- Random sample : x_i
- Average of all samples : $\frac{1}{n}\sum_{k=1}^{n} x_k$
- Average of two extremes : $\frac{1}{2}\,(\,x_{\min} + x_{max}\,)$
- Trimmed mean, $tr(10)$: discard the smallest and largest 10% of the samples then average the rest

Give a 4x3 table where rows representing 4 estimators given above, columns representing the distributions, and give your explanations for each of the correspondence.

Part 4: Bayesian Estimation & Dogmatism

11) Consider the equations (34), (35) and their formulations in your book(DHS) under chapter 3.4.1.

$$\mu_n = \left(\frac{n{\sigma_0}^2}{n{\sigma_0}^2+\sigma^2}\right)\hat{\mu}_n + \left(\frac{\sigma^2}{n{\sigma_0}^2+\sigma^2}\right)\mu_0$$

$${\sigma_n}^2 = \left(\frac{\sigma^2{\sigma_0}^2}{n{\sigma_0}^2+\sigma^2}\right)$$

(a) What happens if the sample size grows to infinity for both of the equations ? Prove mathematically then explain your results.

(b) Prove that μ_n always lies between $\hat{\mu}_n$ and μ_0.

(c) What happens when the variance of the prior distribution of the unknown parameter is 0 for both of the equations? Prove mathematically then explain your results.

(d) What happens if the variance of the prior distribution of the unknown parameter is considerably larger (>>) than the variance of the likelihood function of the unknown parameter, for both of the equations? Prove mathematically then explain your results.

Part 5 : Dimensionality Reduction & Component Analysis

12) "Curse of dimensionality" is one of most important barrier in front of us to build the generalized model in some cases. This problem happens when our feature space is very large compared to the number of examples. Therefore, we should find a way to reduce the feature space in these situations. In this question we are going to implement 2 classical and still efficient dimension reduction algorithms: Principal Component Analysis (PCA) and Fisher's linear discriminant (FLD).

(a) Implement PCA by following instructions inside "*the2_PCA.m*" file.

(b) Implement FDA by following instructions inside "*the2_ FLD.m*" file.