

PART 4

MACHINE LEARNING: PARAMETRIC APPROACH FOR UNCERTAIN PROCESSES

Mohammed Elmusrati, Professor
 School of Technology and Innovations
 University of Vaasa
 Finland
mohammed.elmusrati@univaasa.fi



University
of Vaasa



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

1

1

Uncertainty vs. Learning

- In machine learning, we mainly seek to extract information from raw data. Many applications could be built over machine learning, such as data-driven systems, new inferences, new services, etc.
- We do not know (at least precisely) in advance what we are looking for! Otherwise, no need for machine learning.
- This means that there are uncertainties associated with data.
- Furthermore, data is usually corrupted with noise, distortion and could be biased and incomplete.
- The main goal of machine learning is to resolve the uncertainties or at least to reduce them and enhance the retrieved information quantity and quality (cleaned from distortions).



University
of Vaasa



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

2

2

Outlines of Part 4

- In this part we will go through several fundamental concepts of estimation theory.
- Some important algorithms for parametric machine learning are presented.
- Estimation theory is a probabilistic approach to extracting the most likely information from noisy and distorted data (could be measured, tabulated, or observed).



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

3

3

Introduction

- In estimation theory, we are looking for the best estimate for hidden (unseen) value(s) of parameters based on the available data (measurements) or observations.
- Let's define the problem as:

$$\mathbf{y} = \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}) + \mathbf{n}$$

Where \mathbf{x} is the hidden input, \mathbf{n} is random unknown noise or biasing, and \mathbf{h} is a vector of mapping function of \mathbf{x} to the output based on parameters $\boldsymbol{\theta}$.
The mapping \mathbf{h} could represent the distortion done to the hidden information.

$$\mathbf{x} \in \mathbb{R}^{M \times 1}, \text{ and } \mathbf{y}; \mathbf{n} \in \mathbb{R}^{N \times 1}, \mathbf{h} = [h_1(\mathbf{x}), \dots, h_K(\mathbf{x})]$$

Where M is the length of vector \mathbf{x} , K is the number of functions, N is the number of outputs from each function (without loss of generality, it has been assumed with the same length).



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

4

4

Estimation Problems

- For the previous equation, we have at least the following three interesting cases in the theory of estimation:

1. We know (at least partially) \mathbf{x} and the observation \mathbf{y} , and we are looking for the best set of the mapping functions \mathbf{h} and their parameters $\boldsymbol{\theta}$. This is the central concept of the Regression problem discussed in Part 2.

We may know the general form of the mapping \mathbf{h} , for example, increasing exponentially in the population models. Still, we need to know (or estimate) the parameters of the exponential function.

In most cases, we do not know precisely the shape of the mapping, and we need to test/modify some general shapes to fit the available input data \mathbf{x} and the output \mathbf{y} . One option is to use the black box modeling as we have seen in the Artificial Neural Networks.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

5

5

Estimation Problems

2. The second case is when we have the noisy observations \mathbf{y} , and we know (or at least we can assume) \mathbf{h} with its parameters, and we are looking to estimate \mathbf{x} . Since \mathbf{x} vector is unknown to us (fully or partially) and corrupted by noise, then we should deal with it as a random process.
3. The third case is when we have only the observations \mathbf{y} , and we do not know the mapping functions \mathbf{h} nor the inputs \mathbf{x} . However, we may build statistical relations between the inputs and the outputs. Hence, we are looking for the best estimate of the statistical parameters based on the available information (e.g., \mathbf{y}) and some pre-assumptions about \mathbf{x} .



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

6

6

Introduction

- We may simplify the general form of our estimation problem as:

$$\mathbf{y} = h(\mathbf{x}) + \mathbf{n}$$

- Here, we consider only one mapping function, and we are interested in estimating vector \mathbf{x} based on the N observations of \mathbf{y} . We also assumed that we have K outputs of the mapping $h(\mathbf{x})$. In the case of time series, we assume one output at a time, and then for N samples, \mathbf{y} is a vector with length N .

$$\mathbf{x} \in \mathbb{R}^{M \times 1}, \text{ and } \mathbf{y}; \mathbf{n} \in \mathbb{R}^{N \times 1}$$



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

7

7

One Application

- For the mathematical relation

$$\mathbf{y} = h(\mathbf{x}) + \mathbf{n}$$

- One may think about different applications. One application assumes that the vector \mathbf{y} represents the attributes (features) of patients' records. For example, \mathbf{y} could contain the blood pressure, age, weight, symptoms, BMI, etc.
- We are looking for an \mathbf{x} vector that contains the hidden information, which contains, for example, a list of diseases that the patient has. Here, $h(\mathbf{x})$ is the unknown mapping between the causes and the attributes. Finally, \mathbf{n} is the noise part which represents the limited accuracy or biases in the attributes as well as any other errors in the model.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

8

8

Unbiased Estimation

- How to assess the estimation process? How to decide which estimate is the best? What are our evaluation criteria?
- Usually, we estimate the hidden parameter or inputs \mathbf{x}_{est} based on the observations $\{\mathbf{y}\}$.
- However, since the estimate \mathbf{x}_{est} is based on the random process \mathbf{y} , therefore, the estimate will be a random process as well.
- If the expected value of the estimate $E[\mathbf{x}_{est}]$ equals the actual estimated value, we call this estimate as an unbiased estimate.
- \mathbf{x}_{est} is **unbiased estimate** of \mathbf{x} if $E[\mathbf{x}_{est}] = \mathbf{x}$
- This means that the expected error value between the estimate and the actual value is **zero**.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

9

9

Consistent Estimator

- The estimator \mathbf{x}_{est} is called consistent if increasing the number of observations will reduce the error variance, i.e., when N is the number of observations (or samples) used in the estimation

$$\lim_{N \rightarrow \infty} E \left[\left(\mathbf{x}_{est} - E[\mathbf{x}_{est}] \right)^2 \right] = 0$$

- Hence, it is highly desirable for the estimator to be both **unbiased and consistent**. Because we will be close to the actual parameter when we have a large enough N .



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

10

10

Efficient Estimator

- It is not possible to have an infinite number of samples in order to make the estimate extremely close to the actual parameter.
- Furthermore, it is possible to have several different unbiased and consistent estimators. Hence, which is the best for a finite number of samples?
- We may define the **efficient estimator** as one with the minimum error variance for a certain finite number of samples N .



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

11

11

Efficient Estimator

- In order to decide if the estimator is efficient or not, we should be able to know the minimum variance for estimators of a particular problem.
- Fortunately, the minimum variance or, more precisely, the lower bound of the variance of the unbiased estimators is given by Cramer-Rao (CR) lower bound.
- Considering the same problem in Slide 7 that we are looking to estimate the vector $\mathbf{x}=[x_1, \dots, x_M]^T$ based on the observations $\mathbf{y}=[y_1, \dots, y_N]^T$,



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

12

12

Efficient Estimator

- Define the covariance matrix of the vectors of estimates as

$$\mathbf{R}_{\mathbf{x}_{est}} = E \left[(\mathbf{x}_{est} - \mathbf{x})(\mathbf{x}_{est} - \mathbf{x})^T \right]$$

- The element (i,j) of the Fisher Information matrix $\mathbf{I}(\mathbf{x})$ are given by

$$[\mathbf{I}(\mathbf{x})]_{ij} = -E \left[\frac{\partial^2}{\partial x_i \partial x_j} \log f(\mathbf{y}; \mathbf{x}) \right], \quad i, j = 1, \dots, M$$

- Where $f(\mathbf{y}; \mathbf{x})$ is the joint probability density function between the observations \mathbf{y} and the parameters \mathbf{x}



Cramer-Rao Lower Bound

- Under the following regularity condition

$$E \left[\frac{\partial}{\partial x_i} \log f(\mathbf{y}; \mathbf{x}) \right] = 0 \quad \forall i = 1, \dots, M$$

- We may define the lower bound of the estimate variance as

$$\text{Var} \left([\mathbf{x}_{est}]_i \right) \geq [\mathbf{I}(\mathbf{x})^{-1}]_{ii}$$

- This means that the variance of the i^{th} parameter estimate is lower bounded by the diagonal of the inverse of the Fisher information matrix *under* the regularity condition.



Cramer-Rao Lower Bound

- We can reduce the previous general form given for the Cramer-Rao Lower Bound (CRLB) in the previous slide for a single estimated parameter as:

$$\text{Var}(x_{\text{est}}) \geq \frac{-1}{E \left[\frac{\partial^2}{\partial x^2} \log f(\mathbf{y}; x) \right]}$$

- And the regularity condition as

$$E \left[\frac{\partial}{\partial x} \log f(\mathbf{y}; x) \right] = 0$$

- The proof is not difficult and can be found in many books on estimation theory. However, it is not given here. Nevertheless, we will use the CR bound to prove the efficiency of some estimators.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

15

15

Least Mean Square Error Estimate

- Let's derive our first estimation algorithm. We assume a single scalar parameter, i.e., we have N observations based on

$$y_i = h(x) + n_i, \quad i = 1, 2, \dots, N$$

- We don't know the function $h(\cdot)$, and n_i is an additive random measurement noise (usually zero mean Normal distributed).
- One possible criterion is to find the best estimate x_{est} that minimizes the mean square error as:

$$\min E \left[\left(x - x_{\text{est}}(\mathbf{y}) \right)^2 \right]$$



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

16

16

Least Mean Square Error Estimate



University
of Vaasa

- We may formulate the expected value as

$$E\left[(x - x_{est}(y))^2\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - x_{est}(y))^2 f_{XY}(x, y) dx dy$$

- Where $f_{XY}(x, y)$ is the joint distribution between the measurements and the parameter x . We may modify this formula, since $f_{XY}(x, y) = f_Y(y) f_{X|Y}(x|y)$

$$E\left[(x - x_{est}(y))^2\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - x_{est}(y))^2 f_Y(y) f_{X|Y}(x|y) dx dy$$

- We can find x_{est} that minimizes this function as:

$$\frac{d}{dx_{est}} E\left[(x - x_{est}(y))^2\right] = 0$$



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

17

17

Least Mean Square Error Estimate



University
of Vaasa

- Therefore, $\frac{d}{dx_{est}} \int_{-\infty}^{\infty} f_Y(y) \left[\int_{-\infty}^{\infty} (x - x_{est}(y))^2 f_{X|Y}(x|y) dx \right] dy = 0$

- But since $f_Y(y)$ is always positive, then, it is always correct to say the optimum is achieved at:

$$\frac{d}{dx_{est}} \left[\int_{-\infty}^{\infty} (x - x_{est}(y))^2 f_{X|Y}(x|y) dx \right] = 0 \Rightarrow -2 \int_{-\infty}^{\infty} (x - x_{est}(y)) f_{X|Y}(x|y) dx = 0$$

- Hence,

$$\int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx - \int_{-\infty}^{\infty} x_{est}(y) f_{X|Y}(x|y) dx = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx - x_{est}(y) \int_{-\infty}^{\infty} f_{X|Y}(x|y) dx = 0$$

$$x_{est}(y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx = E[x|y]$$



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

18

18

Least Mean Square Error Estimate



University
of Vaasa

- The previous result is significant since it indicates that the best estimate of the unknown parameter x is the mean of the conditional probability density function of the parameter given the observations, i.e., $f(x|y)$.
- In most practical cases, this PDF function is unavailable. 😞
- Are there other ways to estimate the unknown parameter x based on the observations?
- Yes, we can use several other norms like minimizing the absolute value of the error or minimizing the maximum value of the error as described in the following slides:



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

19

19

Minimum Absolute Error



University
of Vaasa

- What is the optimum parameter estimate x_{est} that minimizes

$$error = E[|x - x_{est}(y)|]?$$

- Let's do it in a similar way that we did to minimize the means square error:

$$E[|x - x_{est}(y)|] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |x - x_{est}(y)| f_Y(y) f_{X|Y}(x|y) dx dy = \int_{-\infty}^{\infty} f_Y(y) \left[\int_{-\infty}^{\infty} |x - x_{est}(y)| f_{X|Y}(x|y) dx \right] dy$$

$$\frac{d}{dx_{est}} E[|x - x_{est}(y)|] = 0 \Rightarrow \frac{d}{dx_{est}} \int_{-\infty}^{\infty} f_Y(y) \left[\int_{-\infty}^{\infty} |x - x_{est}(y)| f_{X|Y}(x|y) dx \right] dy = 0$$

$$\Rightarrow \frac{d}{dx_{est}} \int_{-\infty}^{\infty} |x - x_{est}(y)| f_{X|Y}(x|y) dx = 0 = \frac{d}{dx_{est}} \left[- \int_{-\infty}^{x_{est}(y)} (x - x_{est}(y)) f_{X|Y}(x|y) dx + \int_{x_{est}(y)}^{\infty} (x - x_{est}(y)) f_{X|Y}(x|y) dx \right] = 0$$



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

20

20

Minimum Absolute Error



University
of Vaasa

- Now we differentiate with respect to x_{est} using the following **Leibniz Integral Rule**

$$\frac{\partial}{\partial z} \int_{a(z)}^{b(z)} f(x, z) dx = \int_{a(z)}^{b(z)} \frac{\partial f(x, z)}{\partial z} dx + f(b(z), z) \frac{\partial b}{\partial z} - f(a(z), z) \frac{\partial a}{\partial z}$$

We obtain:

$$\int_{-\infty}^{x_{est}(y)} f_{x|y}(x|y) dx - \int_{x_{est}(y)}^{\infty} f_{x|y}(x|y) dx = 0 \Rightarrow \int_{-\infty}^{x_{est}(y)} f_{x|y}(x|y) dx = \int_{x_{est}(y)}^{\infty} f_{x|y}(x|y) dx$$

- This means that the optimum estimate is the **median** of the conditional probability density function of $f(x|y)$.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

21

21

MinMax Error Criteria



University
of Vaasa

- The third famous criterion for optimizing the estimated parameter is minimizing the maximum error.
- This could be formulated as $\min \left\{ \max E \left[\left| x - x_{est}(y) \right| \right] \right\}$
- Roughly speaking the maximum of x is achieved statistically at the maximum of $f(x|y)$. In other words, the best estimate, in this case, is the mode of $f(x|y)$.

$$\frac{d}{dx_{est}} \left\{ \max \int_{-\infty}^{\infty} \left| x - x_{est}(y) \right| f_{x|y}(x|y) dx \right\}$$



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

22

22

Different Estimators Criteria

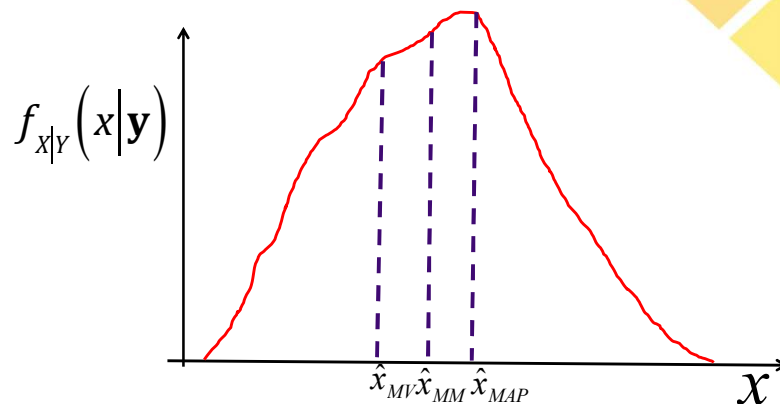
- We have seen so far three different estimators according to three different error criteria. However, all of them are based on the conditional probability density function $f_{x|y}(x|y)$ as:

- Minimizing the error variance (L-2 Norm):** the mean of $f_{x|y}(x|y)$
- Minimize the absolute value (L-1 Norm):** the median of $f_{x|y}(x|y)$
- Minimize the maximum of the error (L-inf Norm):** The mode of

$$f_{x|y}(x|y)$$

23

Different Estimators



It is interesting to know that if the conditional probability is symmetric like the Normal distribution, then all these three estimators are identical.

24

Maximum Likelihood Estimator



- We have seen that the best estimator should be based on the conditional probability of the parameter we are looking to estimate given the observations or measurements, i.e.,

$$f_{x|y}(x|y)$$

- However, unfortunately, it is a big challenge to have an accurate estimation of this posterior density. However, using the Bayes rule, we may express it as:

$$f_{x|y}(x|y) = \frac{f_{y|x}(y|x)f_x(x)}{f_y(y)}$$

25

Maximum Likelihood Estimator



- In terms of estimation theory, we may express the parameters of the previous Bayes' formula as:
 - The density $f_{x|y}(x|y)$ represents the distribution of the unknown parameter x after collecting the measurements y . Hence, it is called the **posterior probability density function**.
 - The density $f_x(x)$ represents our beliefs about the possible values of x before we watch any observations or collect any measurements. This could be based on assumptions and/or physical behavior. It represents the **priori statistical knowledge about x** .
 - The density $f_{y|x}(y|x)$ is called the **likelihood density**, which expresses how the measurement or observations should behave at a certain parameter x .
 - Finally, the density $f_y(y)$ represents the general distribution of the measurements regardless of the parameter x ; it is called the **model evidence** or **marginal likelihood**. It is not a function in the parameter x .

26

Maximum Likelihood Estimator



University
of Vaasa

- As we have seen from the different estimation techniques, one method is by taking the maximum value of the posterior probability density $f_{x|Y}(x|y)$
- Therefore, it is named as maximum a posteriori MAP estimation.
- However, since $f_{x|Y}(x|y)$ is generally very hard to know, let's see how to find some other equivalent estimator.
- Taking the logarithm of the posterior probability density, we obtain:

$$\log(f_{x|Y}(x|y)) = \log(f_{Y|X}(y|x)) + \log(f_X(x)) - \log(f_Y(y))$$

University of Vaasa

Mohammed S. Elmusrati – University of VAASA

27

27

Maximum Likelihood Estimator



University
of Vaasa

- It is clear that taking the logarithm makes the density function easier to handle. Moreover, the logarithm function is a monotonic increasing function, i.e., it will not change the location of the stationary (maximum or minimum) point.
- Generally speaking
if $g(x) > 0$ for all x , and $x_{max} = \text{Arg_Maximize } [g(x)]$
Then it is always true that $x_{max} = \text{Arg_Maximize } [\log(g(x))]$
- Therefore, MAP estimate could be formulated as:

$$\max \left\{ \log(f_{x|Y}(x|y)) \right\} = \max \left\{ \log(f_{Y|X}(y|x)) + \log(f_X(x)) \right\}$$

University of Vaasa

Mohammed S. Elmusrati – University of VAASA

28

28

Maximum Likelihood Estimator



University
of Vaasa

- In the last formulation, we have dropped $f_Y(\mathbf{y})$ because it is not a function in the parameter x , and hence, it does not have any effect in finding the point that maximizes the estimate.
- It is clear that to find the MAP point, we will need to know the likelihood density function as well as the priori statistical knowledge about the parameter $f_X(x)$.
- In case we ignore the a priori part and we maximize only the likelihood density, we call this estimate the **maximum likelihood estimation (ML or MLE)**.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

29

29

Maximum Likelihood Estimator



University
of Vaasa

- In other words, the maximum likelihood estimator is defined as:

$$x_{ML} = \arg \max \left\{ f_{Y|X}(\mathbf{y}|x) \right\} = \arg \max \left\{ \log \left[f_{Y|X}(\mathbf{y}|x) \right] \right\}$$

- We have seen that the MAP is an optimum estimator according to specific criteria.
- Is MLE (maximum likelihood estimator) optimum, and in what sense?
- ML estimator can be the optimum solution as the MAP in some cases and can be suboptimal estimators in other cases, as will be explained later.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

30

30

Maximum Likelihood Estimator



University
of Vaasa

- To see when the MLE=MAP, let's revisit the Bayes rule with the assumption that all measurements (y_1, y_2, \dots, y_N) are independent, then

$$\begin{aligned} \max \left\{ \log \left(f_{x|y}(x|y) \right) \right\} &= \max \left\{ \log \left(f_{y|x}(y|x) f_x(x) \right) \right\} = \max \left\{ \log \left(\prod_{k=1}^N f_{y|x}(y_k|x) f_x(x) \right) \right\} \\ &= \max \left\{ \sum_{k=1}^N \log \left[f_{y|x}(y_k|x) f_x(x) \right] \right\} \end{aligned}$$

- It is clear that $f_x(x)$ is weighting the likelihood function. Hence, if $f_x(x)$ is **uniformly distributed** over the whole range of x , then it will not have any effects on the location of the optimum x . In that case, **MLE=MAP**.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

31

31

Maximum Likelihood Estimator



University
of Vaasa

- The previous status of MAP=MLE could be shown mathematically as:

$$\begin{aligned} \left. \frac{d}{dx} \left(\log \left(f_{x|y}(x|y) \right) \right) \right|_{x=x_{MAP}} &= 0 = \frac{d}{dx} \left(\sum_{k=1}^N \log \left[f_{y|x}(y_k|x) \right] + \log \left[f_x(x) \right] \right) \Big|_{x=x_{MAP}} = \\ \sum_{k=1}^N \frac{1}{f_{y|x}(y_k|x)} \frac{d}{dx} f_{y|x}(y_k|x) + \frac{1}{f_x(x)} \frac{d}{dx} f_x(x) &= 0 \end{aligned}$$

- Hence, when $f_x(x) = \text{constant}$, then its differentiation is zero, so that

$$\sum_{k=1}^N \frac{1}{f_{y|x}(y_k|x)} \frac{d}{dx} f_{y|x}(y_k|x) \Big|_{x=x_{ML}} = 0$$



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

32

32

Maximum Likelihood Estimator



University
of Vaasa

- The MLE could be considered in many cases as suboptimal. However, it can also be considered an optimal solution when no prior information is available about the parameter to be estimated.
- In this case, the best thing is to assume that the unknown parameter is uniformly distributed. In other words, our pre-knowledge uncertainty is the same as any value.
- What is your pre-knowledge about throwing a coin to be landed head or tail without looking at any observations? The best assumption is that each has a probability of 0.5 (uniform). Then the MLE estimation is the optimum, like the MAP. But if you know in advance (based on some pre-knowledge) that the probability of a head is, for example, larger than 0.7, hence, MLE becomes just a suboptimal estimator. One should use, for example, the MAP to get the best estimate.
- These concepts and more will be described through the next few examples.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

33

33

Machines Learn what Behind Uncertainty!



University
of Vaasa

- We have already seen some estimation techniques (e.g., MLE and MAP) to estimate some attributes buried in random data.
- If we are looking for specific information that affects or distorts the stream of random series in a certain manner. For example, it's mean, variance, or whatever.
- We collect samples of these random data, and we apply our algorithms in order to estimate the required parameters or attributes.
- Next, we have several examples to explain this concept.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

34

34

Example (1)

- Assume we are looking for a process of two outcomes (Success) or (Fail). It may express too many practical applications; a few examples:
 - Heading a target or mis-heading
 - Correct or incorrect receiving of a transmitted symbol or message.
 - Positive or negative revenue
 - Spam email or not
- Based on some historical independent observations, we like to estimate the process parameter (in this case, it will be the probability of success p). Let's first assume that we have no pre-knowledge about the process.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

35

35

Example (1)

- Assume that we have N observations, where M of them were Successes (S) with $y_k=1$ and $(N-M)$ were Fails (F) with $y_k=0$, as: SSFSFSSFFSFFFSSFFSSFFFFSSSFFS
- Since all observations are independent, hence,

$$P(\mathbf{y}|x=p) = p^M (1-p)^{N-M}$$
- Although it is easy to find the MLE estimation of this problem by taking the differentiation, however, in some other more complicated problems, it could be tedious and lengthy. However, convert the multiplications into summation by taking the logarithm makes it even easier to solve.
- We may call it the *Likelihood function* (not density!) as

$$l(\mathbf{y}; x=p) = \log [P(\mathbf{y}|x=p)] = M \log(p) + (N-M) \log(1-p)$$



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

36

36

Example (1)

- Now we find the estimate of p which maximizes the likelihood function (or density) as:

$$\left. \frac{d}{dp} l(\mathbf{y}; x = p) \right|_{p=\hat{p}} = 0 \Rightarrow \frac{M}{\hat{p}} - \frac{N-M}{1-\hat{p}} = 0$$

$$\Rightarrow \frac{M}{\hat{p}} = \frac{N-M}{1-\hat{p}} \Rightarrow N\hat{p} - M\hat{p} = M - M\hat{p} \Rightarrow \hat{p} = \frac{M}{N}$$

- Hence, the expected result to estimate the probability is the MLE estimation.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

37

37

Example (2)

- We may express what we have done in the previous slide mathematically as:

$$\hat{p} = \frac{1}{N} \sum_{k=1}^N y_k; P(y_k = 1) = p, \text{ and } P(y_k = 0) = 1 - p$$

- In some applications, we may call y_k as an indicator function.
- Is this MLE estimator biased or unbiased?**
- Is it consistent or not?**
- Is it an efficient estimator or not?**



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

38

38

Example (2)

- Based on the definition of unbiased estimators, we should find the expected value of estimation as:

$$E[y_k] = 1 \times P(y_k = 1) + 0 \times P(y_k = 0) = p$$

- Therefore,

$$E[\hat{p}] = \frac{1}{N} E\left[\sum_{k=1}^N y_k\right] = \frac{1}{N} \sum_{k=1}^N E[y_k] = \frac{1}{N} \sum_{k=1}^N p = \frac{Np}{N} = p$$

hence, this estimator is unbiased.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

39

39

Example (2)

- Now let's compute the variance of the estimated value to see if it is consistent or not! The derivation is given in step-by-step next:

$$\begin{aligned} \text{Var}(\hat{p}) &= E\left[(\hat{p} - E[\hat{p}])^2\right] = E\left[(\hat{p} - p)^2\right] = E[\hat{p}^2] - p^2 = E\left[\left(\frac{1}{N} \sum_{k=1}^N y_k\right)^2\right] - p^2 \\ &\Rightarrow E\left[\left(\frac{1}{N} \sum_{k=1}^N y_k\right)^2\right] = E\left[\frac{1}{N^2} \left(\sum_{k=1}^N y_k\right)^2\right] = E\left[\frac{1}{N^2} \sum_{k=1}^N \sum_{i=1}^N y_k y_i\right] = \frac{1}{N^2} \sum_{k=1}^N \sum_{i=1}^N E[y_k y_i] \\ E[y_k y_i] &= \begin{cases} E[y_k^2], & k=i \\ E[y_k]E[y_i], & k \neq i \end{cases}; E[y_k^2] = 1^2 P(y_k = 1) + 0^2 P(y_k = 0) = p \Rightarrow E[y_k y_i] = \begin{cases} p, & k=i \\ p^2, & k \neq i \end{cases} \\ &\Rightarrow \frac{1}{N^2} \sum_{k=1}^N \sum_{i=1}^N E[y_k y_i] = \frac{1}{N^2} (Np + (N^2 - N)p^2) = \frac{p}{N} + \frac{(N-1)}{N} p^2 \Rightarrow \text{Var}(\hat{p}) = \frac{p}{N} + \frac{(N-1)}{N} p^2 - p^2 = \frac{p(1-p)}{N} \end{aligned}$$



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

40

40

Example (2)

- From the previous slide result, it is clear that

$$\lim_{N \rightarrow \infty} \text{Var}(\hat{p}) = \lim_{N \rightarrow \infty} \frac{p(1-p)}{N} = 0$$

- Hence the estimator is also consistent** 😊
- Is it possible to have a better unbiased and consistent estimator than this one?
- To answer this, we should find the variance lower bound (CRLB)
- This is left as an exercise!

$$\text{Var}(x_{\text{est}}) \geq \frac{-1}{E \left[\frac{\partial^2}{\partial x^2} \log f(\mathbf{y}; x) \right]}$$

41

Example (3)

- In Example (2), assume that we could express our pre-uncertainty of the parameter x (i.e., the probability of success) as:

$$f_X(x) = 4e^{-\alpha x}, \quad 0 \leq x \leq 1$$

- Find the value of α
- How this pre-knowledge might affect our optimum estimator about the probability of x based on the observations.
- Compare both results.

42

Example (3)

- If we have maximum uncertainty about the parameter before looking at any measurements or observations, the optimum estimator is the MLE estimator.
- However, if there is some pre-knowledge about the parameter that we are looking at, then the MLE is not the optimum anymore.
- We should use any available information to improve the estimation.



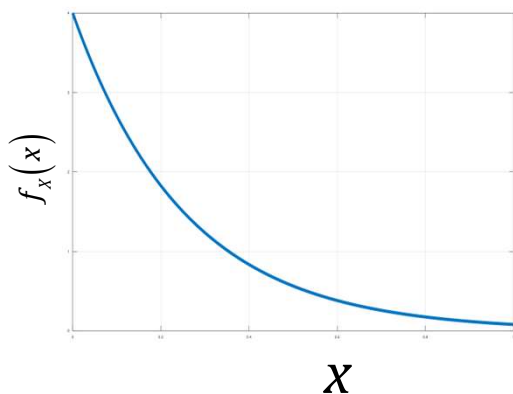
University of Vaasa

Mohammed S. Elmusrati – University of VAASA

43

43

Example (3)



$$\int_0^1 f_x(x) dx = 1 \Rightarrow 4 \int_0^1 e^{-\alpha x} dx = 1$$

$$\Rightarrow \alpha = 3.9207 \quad \text{Prove it?}$$

Looking to the priori distribution, the probability of being Success or Fail is not uniform. Now we have more accurate impression about the uncertainty. Actually, we know that the probability for the Success case is less than 0.3 with a chance of about 70%. This kind of information should have impact to improve our estimation about the parameter x .



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

44

44

Example (2)

- From slide 32

$$\left. \frac{d}{dx} \left(\sum_{k=1}^N \log [f_{Y|X}(y_k|x)] + \log [f_X(x)] \right) \right|_{x=x_{MAP}} = 0$$

$$\left. \frac{d}{dx} (M \log(x) + (N-M) \log(1-x) + \log(4) - \alpha x) \right|_{x=x_{MAP}} = 0$$

$$\frac{M}{x_{MAP}} - \frac{N-M}{1-x_{MAP}} - \alpha = 0 \Rightarrow \alpha x_{MAP}^2 - (\alpha + N)x_{MAP} + M = 0$$

$$\Rightarrow x_{MAP} = \frac{(3.92 + N) \pm \sqrt{(3.92 + N)^2 - 15.68M}}{7.84}, \quad 0 \leq x_{MAP} \leq 1$$

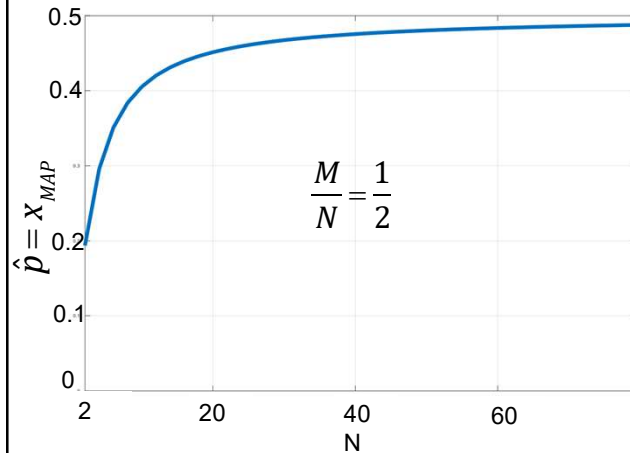
45

Example (2)

- In the previous equation, we have two results, however, we should always select the one which is between 0 and 1.
- Let's assume that in our observation we have $M/N=0.5$.
- In MLE estimation, we will decide the estimate as $p=0.5$. But how it will be in MAP estimate with the availability of a priori density function.
- Next figure shows the MAP estimate for $M/N=0.5$ for several N values (number of observations).

46

Example (2)



From this figure, we can easily see the impact of the pre-knowledge on the estimation of the parameter.

If we have only two observations, i.e., $N=2$, and we have $M=1$, in MLE, the best estimate of the probability $p=0.5$. But with MAP, we can see that the probability is just **0.2**. However, if we repeat the experiment many times, i.e., N is very large, and we have $M/N=0.5$, then we approach the belief that $p=0.5$.

In other words, we believe more in our measurements as N increases than in our pre-knowledge about the uncertainty of the parameter.

Mohammed S. Elmusrati – University of VAASA

47

Exercise

- In the previous estimation example, if as a priori information, we know that the probability is uniformly distributed from 0.4 to 0.8.
- Find the MAP estimation in this case.

University of Vaasa

Mohammed S. Elmusrati – University of VAASA

48

48

Multinomial Density



University
of Vaasa

- Consider the generalization of Bernoulli where instead of two states, the outcome of a random event is one of K mutually exclusive and exhaustive states, for example, classes, each of which has a probability of occurring p_i with:

$$P(x_1, x_2, \dots, x_K) = \prod_{i=1}^K p_i^{x_i}$$

$$\sum_{i=1}^K p_i = 1$$

- Let x_1, x_2, \dots, x_K are the indicator variables where x_i is 1 if the outcome is at state i and 0 otherwise. (Revise multinomial distribution in the previous Part 3)



University of Vaasa

Prof. Mohammed Elmusrati

49

49

Multinomial Density



University
of Vaasa

- Let us say we do N such independent experiments with outcomes:

$$\mathcal{X} = \{x^t\}_{t=1}^N$$

- Where,

$$x_i^t = \begin{cases} 1 & \text{if experiment } t \text{ chooses state } i \\ 0 & \text{otherwise} \end{cases}$$

with $\sum_i x_i^t = 1$. The MLE of p_i is

$$\hat{p}_i = \frac{\sum_t x_i^t}{N}$$

- The estimate for the probability of state i is the ratio of experiments with the outcome of the state i to the total number of experiments.



University of Vaasa

Prof. Mohammed Elmusrati

50

50

Multinomial Density: Example (4)



University
of Vaasa

- Assume that we are observing a mouse hit three balls, red, green, and black. Each time, the mouse selects one ball to touch it. We do not know what criteria it uses. However, we want to have some statistical inference about its behavior. One hit observed sequence was RBRRGBGBRRR, what is the maximum likelihood estimate of the probabilities assuming independence.
- Solution: we apply the obtained results directly as:
- Assume x_1 is the number of hitting the red ball = 6, x_2 is the number of hitting the Black ball = 3, and x_3 is the number of hitting the Green ball = 2, hence, $p_1=6/11$, $p_2=3/11$, and $p_3=2/11$.
- If we assume that the mouse outcomes are dependent, i.e., the mouse selects the next ball to hit based (somehow) on the previous hit. In this case, we may use the Markov process to model its behavior.



University of Vaasa

Prof. Mohammed Elmusrati

51

51

Example (5)



University
of Vaasa

- Assume that we are interested in estimating the actual value of a constant x . However, the observation (or measurement) is always corrupted by a zero mean Gaussian noise with known variance σ^2 .
- The mathematical model of this problem is

$$y_i = x + n_i$$

- Where $y_i=y_1, y_2, \dots, y_N$ are N different available measurements and $n_i=n_1, n_2, \dots, n_N$ are independent identical distributed zero mean Gaussian (Normal) distribution.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

52

52

Example (5)

- The problem presented in the previous slide is very important as it gives the foundation of several concepts in the *Estimation theory*.
- Since n_i samples are zero mean Normal distributed random process, hence, it is clear that the measurements y_i are also Normally distributed process but with mean equals the constant x and with the same variance as n_i . Therefore,

$$f_{y|x}(y_i|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i-x)^2/2\sigma^2}$$

53

Example (5)

- In this example, we assume that we have no prior knowledge about the parameter x we are looking to estimate. Therefore, the optimum estimator is the maximum likelihood estimation.
- Assume we have N measurements, y_1, y_2, \dots, y_N . Therefore, the likelihood density becomes (due to the independence assumptions of n_i)

$$f_{y|x}(y_1, y_2, \dots, y_N|x) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^N e^{-\sum_{i=1}^N (y_i-x)^2/2\sigma^2}$$

54

Example (5)

- Again, we compute the likelihood function as

$$l(\mathbf{y}; x) = \log \left[f_{\mathbf{y}|x}(\mathbf{y}|x) \right] = -\frac{N}{2} \log(2\pi) - N \log(\sigma) - \frac{\sum_{i=1}^N (y_i - x)^2}{2\sigma^2}$$

$$\Rightarrow \left. \frac{d}{dx} l(\mathbf{y}; x) \right|_{x=x_{ML}} = \frac{\sum_{i=1}^N (y_i - x)}{\sigma^2} = 0 \Rightarrow x_{ML} = \frac{\sum_{i=1}^N y_i}{N}$$

Hence, the well-known mean formula is the MLE estimation for the actual mean value in the case of Normal distribution.

55

Example (5)

- It is quite easy to prove that the previous MLE of the mean is unbiased and consistent estimator as:

$$E[x_{ML}] = E\left[\frac{\sum_{i=1}^N y_i}{N}\right] = \frac{\sum_{i=1}^N E[y_i]}{N} = \frac{\sum_{i=1}^N x}{N} = \frac{Nx}{N} = x \Rightarrow \text{unbiased}$$

$$\text{Var}(x_{ML}) = E\left[\left(x_{ML} - E[x_{ML}]\right)^2\right] = E\left[\left(\frac{\sum_{i=1}^N y_i}{N} - x\right)^2\right] = \frac{1}{N^2} E\left[\left(\sum_{i=1}^N y_i - Nx\right)^2\right]$$

$$= \frac{1}{N^2} E\left[\sum_{i=1}^N y_i^2 + \sum_{i=1}^N \sum_{k \neq i}^N y_i y_k - 2Nx \sum_{i=1}^N y_i + N^2 x^2\right]$$

$$= \frac{1}{N^2} \left[N(\sigma^2 + x^2) + N(N-1)x^2 - 2N^2 x^2 + N^2 x^2 \right] = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N} \Rightarrow \text{Consistent}$$

56

Example (5)

- Let's check if the estimator is efficient or not:

$$\text{Var}(x_{\text{est}}) \geq \frac{-1}{E\left[\frac{\partial^2}{\partial x^2} \log f(\mathbf{y}; x)\right]}$$

$$\begin{aligned} f_{y,x}(\mathbf{y}; x) &= f_{y|x}(\mathbf{y}|x) f_x(x) \Rightarrow \log[f_{y,x}(\mathbf{y}; x)] = \log[f_{y|x}(\mathbf{y}|x)] + \log[f_x(x)] \\ &= -\frac{N}{2} \log(2\pi) - N \log(\sigma) - \frac{\sum_{i=1}^N (y_i - x)^2}{2\sigma^2} + \alpha \quad (\text{constant represents } \log[f_x(x)]) \\ \Rightarrow \frac{\partial}{\partial x} \log[f_{y,x}(\mathbf{y}; x)] &= \frac{\sum_{i=1}^N (y_i - x)}{\sigma^2} \Rightarrow \frac{\partial^2}{\partial x^2} \log[f_{y,x}(\mathbf{y}; x)] = -\frac{N}{\sigma^2} \\ \Rightarrow \text{Var}(x_{\text{est}}) &\geq \frac{\sigma^2}{N}, \text{ Hence, } x_{ML} \text{ is an efficient estimator} \end{aligned}$$

It should be quite easy for you to prove the regularity condition:

$$E\left[\frac{\partial}{\partial x} \log f(\mathbf{y}; x)\right] = 0$$

57

Example (5)

- In the same example, assume that we are interested also to estimate the noise variance.

$$\begin{aligned} l(\mathbf{y}; x) &= \log[f_{y|x}(\mathbf{y}|x)] = -\frac{N}{2} \log(2\pi) - N \log(\sigma) - \frac{\sum_{i=1}^N (y_i - x)^2}{2\sigma^2} \\ \Rightarrow \frac{d}{d\sigma} l(\mathbf{y}; x) \Big|_{\sigma=\sigma_{ML}} &= -\frac{N}{\sigma} + \frac{\sum_{i=1}^N (y_i - x)^2}{\sigma^3} = 0 \Rightarrow \sigma_{ML}^2 = \frac{\sum_{i=1}^N (y_i - x_{ML})^2}{N} \end{aligned}$$

58

Example (5)

- Is the previous estimate of the variance unbiased?

$$\begin{aligned}\because E[x_{ML}^2] &= \frac{1}{N^2} E\left[\left(\sum_{i=1}^N y_i\right)^2\right] = \frac{N(\sigma^2 + x^2) + N(N-1)x^2}{N^2} = \frac{\sigma^2 + Nx^2}{N} = \frac{\sigma^2}{N} + x^2 \\ E[\sigma_{ML}^2] &= \frac{1}{N} E\left[\sum_{i=1}^N (y_i - x_{ML})^2\right] = \frac{1}{N} \sum_{i=1}^N E[y_i^2 - 2y_i x_{ML} + x_{ML}^2] \\ &= \frac{N}{N} \left[(\sigma^2 + x^2) - 2x^2 + \frac{\sigma^2}{N} + x^2 \right] = \frac{N+1}{N} \sigma^2 \Rightarrow \text{Biased}\end{aligned}$$

- To have unbiased estimator:

Is it a big problem? No, especially for large N!

$$\sigma_{ML}^2 = \frac{\sum_{i=1}^N (y_i - x_{ML})^2}{N-1}$$

59

Example (6)

- In the same previous example, let's assume that we have a priori knowledge about the parameter to be estimated.
- For example, assume x itself has a Normal distribution with known mean μ_x and variance σ_x .
- It might be the same problem that x is not fixed but changing randomly. However, we may assume that it is fixed during the measurement period. One example is tracking a moving object in an unpredictable way. Hence, we will collect data to estimate its updated location
- Since in this problem, we have some extra knowledge even with high uncertainty, we should use MAP instead of MLE.

60

Example (6)



University
of Vaasa

- Using the MAP formulation given before in slide 32:

$$\begin{aligned} \frac{d}{dx} \left(\sum_{k=1}^N \log [f_{y_k|x}(y_k|x)] + \log [f_x(x)] \right) \Big|_{x=x_{MAP}} &= 0 \\ \frac{d}{dx} \left(-\frac{N}{2} \log(2\pi) - N \log(\sigma) - \frac{\sum_{i=1}^N (y_i - x)^2}{2\sigma^2} - \frac{\log(2\pi)}{2} - \log(\sigma_x) - \frac{(x - \mu_x)^2}{2\sigma_x^2} \right) \Big|_{x=x_{MAP}} &= 0 \\ \frac{\sum_{i=1}^N (y_i - x_{MAP})}{\sigma^2} - \frac{x_{MAP} - \mu_x}{\sigma_x^2} = 0 \Rightarrow \frac{\sum_{i=1}^N y_i}{\sigma^2} - \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_x^2} \right) x_{MAP} + \frac{\mu_x}{\sigma_x^2} = 0 \\ x_{MAP} = \frac{\sigma^2}{N\sigma_x^2 + \sigma^2} \mu_x + \frac{\sigma_x^2 \sum_{i=1}^N y_i}{N\sigma_x^2 + \sigma^2}, \beta \triangleq \frac{\sigma_x^2}{\sigma^2} \Rightarrow x_{MAP} = \frac{1}{1 + N\beta} \mu_x + \frac{1}{1 + (N\beta)^{-1}} x_{ML}, x_{ML} = \frac{\sum_{i=1}^N y_i}{N} \end{aligned}$$

61

61

Example (6)



University
of Vaasa

- The result in the previous slide is rather interesting.
- If σ_x is very small and close to zero, this means that the uncertainty of x is very small; and it should be very close to μ_x . Look at the MAP estimation with β is close to zero. You can see that $x_{MAP} \rightarrow \mu_x$ regardless of the number of samples N and the values of y_i .
- When σ_x is not tiny, but N is very large, our estimate will be closer to the MLE (summation of measurements divided by N).
- Actually, the MAP estimation is the optimum compromise between the information gained from the measurements and the prior information covered by $f_x(x)$.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

62

62

Naive Bayes

- We have seen in most of our previous analysis that we assume independent attributes or features.
- Mathematically, this could be expressed as

$$P(C_i | x_1, x_2, \dots, x_d) = \frac{f(x_1, x_2, \dots, x_d | C_i)}{f(x_1, x_2, \dots, x_d)} P(C_i) = \frac{\prod_{k=1}^d f(x_k | C_i)}{\prod_{k=1}^d f(x_k)} P(C_i)$$

- This is called the naive Bayes algorithm. For example, we do not take into account the cross-correlations between attributes. However, there could be strong relations.
- For example, in the fruits' classifier, every feature (color, size, shape, etc.) is assumed to be independent.
- Although it is a strong assumption, however, it has greatly simplified the analysis and has also been approved to work very well in many cases.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

63

63

Exercise

- Assume that we are interested to estimate a slowly changing unknown random process. However, we know that it follows normal distribution $N(1, 8)$. However, our measurements are corrupted with zero mean random noise as $N(0, 1)$.
- Write a simulation code to assist both MAP and MLE estimation methods for $N=1$ to 100. Compute the average error for 20 random values of the unknown parameter.
- Plot the results.



University of Vaasa

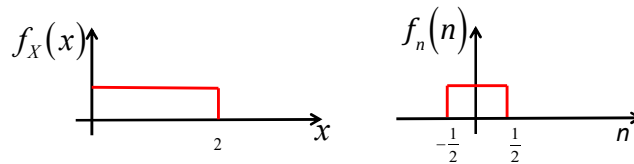
Mohammed S. Elmusrati – University of VAASA

64

64

Example (7)

- A random variable x is to be estimated on the basis of a priori information and the i^{th} noisy measurement is expressed as $y_i = x + n_i$, where n_i is the i^{th} noise sample
- Moreover, x and n_i are assumed to be independent. The distributions functions of x and n_i are shown next.
- Find the optimum estimate of x ?



65

Example (7)

- As we have done before, let's first construct $f_{x|y_i}$

$$f_{x|y_i}(x|y_i) = \frac{f_{y_i|x}(y_i|x)f_X(x)}{f_{y_i}(y_i)}$$

- Since f_{y_i} is not function in the parameter x , then we may ignore it, also f_X is fixed from 0 to 2, then it is useful only in the determination of the range of the admissible values of x .
- Therefore, as for the MLE estimation, we may find the optimum x by looking to one of moments of $f_{y_i|x}$
- It is clear that

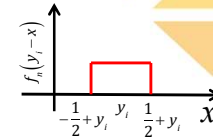
$$f_{y_i|x}(y_i|x) = f_n(y_i - x) = \begin{cases} 1 & -\frac{1}{2} \leq y_i - x \leq \frac{1}{2} \\ 0 & \text{elsewhere} \end{cases}$$

66

Example (7)

- From the previous equation, we have

$$f_{y_i|x}(y_i|x) = f_n(y_i - x) = \begin{cases} 1 & -\frac{1}{2} + y_i \leq x \leq \frac{1}{2} + y_i \\ 0 & \text{elsewhere} \end{cases}$$



- Therefore, the parameter x could be determined based on the measurement. For example, if we have a single measurement, say $y_i=1$, hence, the conditional probability will be uniform from 0.5 to 1.5. It is clear that this uniform does not have a single mode (maximum) value. Hence, we may consider the mean or the median which are equal, and $\hat{x} = 1$, i.e., $\hat{x} = y_i$. However, we should keep in mind that $0 \leq x \leq 2$ as well.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

67

67

Example (7)

- For example, what will be our estimate if $y_i=2.2$? Here, we know from our priori information that the maximum of x is 2. Hence it cannot be 2.2? Therefore, we should truncate the maximum to 2, and the minimum is $2.2-0.5=1.7$. Therefore, the average is $(1.7+2)/2=1.85$.
- We may construct the optimum estimate of the parameter based on the measurements as:

$$\hat{x} = \begin{cases} y_i, & 0.5 \leq y_i \leq 1.5 \\ \frac{y_i + 0.5}{2}, & y_i \leq 0.5 \\ \frac{y_i + 1.5}{2}, & y_i > 1.5 \end{cases}$$



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

68

68

Example (8)

- In Example (5), if the noise samples n_i are correlated. How will this dependence affect the estimation of the parameter x ?
- The conditional probability of the measurements based on the estimated parameter is:

$$f_{\mathbf{y}|x}(\mathbf{y}|x) = \frac{1}{(2\pi)^{N/2} |\mathbf{R}_{nn}|^{0.5}} e^{-\frac{1}{2}(\mathbf{y}-x\mathbf{1})^T \mathbf{R}_{nn}^{-1}(\mathbf{y}-x\mathbf{1})}; \quad \mathbf{1} = [1, 1, \dots, 1]^T$$

69

Example (8)

$$f_{\mathbf{y}|x}(\mathbf{y}|x) = \frac{1}{(2\pi)^{N/2} |\mathbf{R}_{nn}|^{0.5}} e^{-\frac{1}{2}(\mathbf{y}-x\mathbf{1})^T \mathbf{R}_{nn}^{-1}(\mathbf{y}-x\mathbf{1})}; \quad \mathbf{1} = [1, 1, \dots, 1]^T$$

$$l(\mathbf{y}; x) = \log \left[f_{\mathbf{y}|x}(\mathbf{y}|x) \right] = -\frac{N}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{R}_{nn}| - \frac{1}{2} (\mathbf{y} - x\mathbf{1})^T \mathbf{R}_{nn}^{-1} (\mathbf{y} - x\mathbf{1})$$

$$\left. \frac{d}{dx} l(\mathbf{y}; x) \right|_{x=x_{ML}} = 0 = \mathbf{1}^T \mathbf{R}_{nn}^{-1} (\mathbf{y} - x_{ML} \mathbf{1}) + (\mathbf{y} - x_{ML} \mathbf{1})^T \mathbf{R}_{nn}^{-1} \mathbf{1} \Rightarrow x_{ML} = \frac{\mathbf{1}^T \mathbf{R}_{nn}^{-1} \mathbf{y}}{\mathbf{1}^T \mathbf{R}_{nn}^{-1} \mathbf{1}}$$

70

Example (9)

- In example (8) find analytically the ML estimation for two measurements and the following three cases:

$$\mathbf{R}_{nn} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \quad \mathbf{R}_{nn} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \quad \mathbf{R}_{nn} = \begin{bmatrix} \sigma_1^2 & \alpha \\ \alpha & \sigma_2^2 \end{bmatrix}$$

- Compare and comments the results

71

Example (9)

$$\begin{aligned} x_{ML} &= \frac{\mathbf{1}^T \mathbf{R}^{-1} \mathbf{y}}{\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1}}, \text{ for } \mathbf{R} = \begin{bmatrix} \delta^2 & 0 \\ 0 & \delta^2 \end{bmatrix} \\ \Rightarrow \mathbf{R}^{-1} &= \begin{bmatrix} 1/\delta^2 & 0 \\ 0 & 1/\delta^2 \end{bmatrix} \Rightarrow \mathbf{1}^T \mathbf{R}^{-1} = [1 \quad 1] \begin{bmatrix} 1/\delta^2 & 0 \\ 0 & 1/\delta^2 \end{bmatrix} \\ &= \begin{bmatrix} 1/\delta^2 & 1/\delta^2 \end{bmatrix} \Rightarrow \mathbf{1}^T \mathbf{R}^{-1} \mathbf{1} = \begin{bmatrix} 1/\delta^2 & 1/\delta^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{2}{\delta^2} \\ \mathbf{1}^T \mathbf{R}^{-1} \mathbf{y} &= \begin{bmatrix} 1/\delta^2 & 1/\delta^2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{y_1 + y_2}{\delta^2} \\ \Rightarrow x_{ML} &= \frac{(y_1 + y_2)/\delta^2}{2/\delta^2} = \frac{y_1 + y_2}{2} \end{aligned}$$

72

Example (9)

$$\text{For } \mathbf{R} = \begin{bmatrix} \delta_1^2 & 0 \\ 0 & \delta_2^2 \end{bmatrix} \Rightarrow \mathbf{R}^{-1} = \begin{bmatrix} 1/\delta_1^2 & 0 \\ 0 & 1/\delta_2^2 \end{bmatrix} \Rightarrow \mathbf{1}^T \mathbf{R}^{-1} = \begin{bmatrix} 1/\delta_1^2 & 1/\delta_2^2 \end{bmatrix}$$

$$\Rightarrow \mathbf{1}^T \mathbf{R}^{-1} \mathbf{1} = \frac{1}{\delta_1^2} + \frac{1}{\delta_2^2}$$

$$\mathbf{1}^T \mathbf{R}^{-1} \mathbf{y} = \frac{y_1}{\delta_1^2} + \frac{y_2}{\delta_2^2} \Rightarrow x_{ML} = \frac{\delta_2^2}{\delta_1^2 + \delta_2^2} y_1 + \frac{\delta_1^2}{\delta_1^2 + \delta_2^2} y_2$$

As the variance δ^2 increases, this means less reliable measure. Hence if $\delta_1^2 \gg \delta_2^2 \Rightarrow x_{ML} \cong y_2$. In different reliable measures, the optimum estimation is the weighting sum of the measurements according to their variances.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

73

73

Example (9)

$$\text{For } \mathbf{R} = \begin{bmatrix} \delta_1^2 & \alpha \\ \alpha & \delta_2^2 \end{bmatrix} \Rightarrow \mathbf{R}^{-1} = \frac{1}{\delta_1^2 \delta_2^2 - \alpha^2} \begin{bmatrix} \delta_2^2 & -\alpha \\ -\alpha & \delta_1^2 \end{bmatrix}, \beta \triangleq \delta_1^2 \delta_2^2 - \alpha^2$$

$$\Rightarrow \mathbf{1}^T \mathbf{R}^{-1} = \frac{1}{\beta} \begin{bmatrix} \delta_2^2 - \alpha & \delta_1^2 - \alpha \end{bmatrix}$$

$$\Rightarrow \mathbf{1}^T \mathbf{R}^{-1} \mathbf{1} = \frac{1}{\beta} (\delta_2^2 + \delta_1^2 - 2\alpha)$$

$$\mathbf{1}^T \mathbf{R}^{-1} \mathbf{y} = \frac{1}{\beta} ((\delta_2^2 - \alpha) y_1 + (\delta_1^2 - \alpha) y_2)$$

$$\Rightarrow x_{ML} = \frac{\delta_2^2 - \alpha}{\delta_1^2 + \delta_2^2 - 2\alpha} y_1 + \frac{\delta_1^2 - \alpha}{\delta_1^2 + \delta_2^2 - 2\alpha} y_2$$



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

74

74

Example (10)

- Assume a system consists of complex interconnected subsystems. Those subsystems may have faults independently with exponentially distributed time and parameter λ . The system is robust, so it will have general fault only if k of the subsystems has failed to operate.
- Therefore, the time until the system has a general failure is given by

$$y = \sum_{i=1}^k x_i$$

Where x_i is an exponentially distributed random variable with parameter λ . Hence, the distribution of y is Gamma with the following probability density function

$$f_Y(y) = \frac{y^{k-1} \lambda^k}{\Gamma(k)} e^{-\lambda y}$$



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

75

75

Example (10)

- We have a database history about a particular system with several failure times recorded as $\mathbf{y} = [y_1, y_2, \dots, y_N]$.
- Assume we do not know k nor λ . Based on the observations \mathbf{y} , find the MLE estimation of k and λ .
- Solution:**
- Since we assume that all records are independent, hence

$$f_Y(y_1, \dots, y_N | k, \lambda) = \prod_{i=1}^N f_Y(y_i | k, \lambda) = \left(\frac{\lambda^k}{\Gamma(k)} \right)^N e^{-\lambda \sum_{i=1}^N y_i} \prod_{i=1}^N y_i^{k-1}$$

- The log-likelihood function is given by

$$L(y_1, \dots, y_N; k, \lambda) = \log[f_Y(\mathbf{y} | k, \lambda)] = N \log \left(\frac{\lambda^k}{\Gamma(k)} \right) - \lambda \sum_{i=1}^N y_i + (k-1) \sum_{i=1}^N \log(y_i)$$



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

76

76

Example (10)

- Now we can find the optimum parameters that maximizes the log-likelihood function as

$$\frac{\partial}{\partial k} L(y_1, \dots, y_N; k, \lambda) = 0; \text{ and } \frac{\partial}{\partial \lambda} L(y_1, \dots, y_N; k, \lambda) = 0$$

$$\Rightarrow \frac{\partial}{\partial k} \left[Nk \log(\lambda) - N \log(\Gamma(k)) - \lambda \sum_{i=1}^N y_i + (k-1) \sum_{i=1}^N \log(y_i) \right] = N \log(\lambda) - \frac{N}{\Gamma(k)} \frac{d\Gamma(k)}{dk} + \sum_{i=1}^N \log(y_i) \Big|_{k=k_{ML}, \lambda_{ML}} = 0$$

$$\text{Also } \frac{\partial}{\partial \lambda} \left[Nk \log(\lambda) - N \log(\Gamma(k)) - \lambda \sum_{i=1}^N y_i + (k-1) \sum_{i=1}^N \log(y_i) \right] = \frac{Nk}{\lambda} - \sum_{i=1}^N y_i \Big|_{k=k_{ML}, \lambda_{ML}} = 0 \Rightarrow \lambda_{ML} = \frac{Nk_{ML}}{\sum_{i=1}^N y_i}$$

- Substituting the last result in the above equation we obtain

$$\log \left(\frac{Nk_{ML}}{\sum_{i=1}^N y_i} \right) - \frac{\Gamma'(k_{ML})}{\Gamma(k_{ML})} + \frac{1}{N} \sum_{i=1}^N \log(y_i) = 0 \Rightarrow \log(k_{ML}) - \frac{\Gamma'(k_{ML})}{\Gamma(k_{ML})} = \log \left(\frac{\sum_{i=1}^N y_i}{N} \right) - \frac{1}{N} \sum_{i=1}^N \log(y_i), \text{ where } \Gamma'(k_{ML}) = \frac{d\Gamma(k)}{dk} \Big|_{k=k_{ML}}$$

77

Example (10)

- It is clear from the previous result that estimating the parameters of the gamma distribution, we need to solve a non-linear equation. There are many efficient numerical methods that could be used to solve the previous equations.
- If you have the following database of $y=[2, 3, 7, 9, 3, 5]$, estimate the parameters k and λ .

78

Exercise

- You have database history about a certain system as $\mathbf{y}=[y_1, y_2, \dots, y_N]$. We believe that y_i represents Chi-Square random variables (revise Part 3).
- Find the MLE estimation of its number of freedom and its variance.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

79

79

Parametric Classification



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

80

80

Parametric Classification



University
of Vaasa

- In case we have classes with certain uncertainties of data, i.e., we are looking to estimate in which class the data belongs to? we can use the Bayesian relation as (C_i is the i^{th} class)

$$P(C_i|x) = \frac{f(x|C_i)P(C_i)}{f(x)} = \frac{f(x|C_i)P(C_i)}{\sum_{k=1}^K f(x|C_k)P(C_k)}$$

- We call this function a discriminant function:

$$g_i(x) = f(x|C_i)P(C_i)$$

- Or equivalently, it can be redefined with the logarithm as:

$$g_i(x) = \log[f(x|C_i)] + \log[P(C_i)]$$



University of Vaasa

Prof. Mohammed Elmusrati

81

81

Parametric Classification



University
of Vaasa

- In the case of Normal distribution, $f(x|C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$

- The previous discrimination function for Normal pdf becomes:

$$g_i(x) = -\frac{1}{2}\log(2\pi) - \log(\sigma_i) - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log(P(C_i))$$

- Example: Assume you have a car company selling K different cars, and for simplicity, let us say that the sole factor that affects a customer's choice is his or her yearly income, which we denote by x . Then $P(C_i)$ is the proportion of customers who buy car type i . If the yearly income distributions of such customers can be approximated with a Gaussian, then $f(x|C_i)$, the pdf of customers with income x that buy cars in class i , this could be taken $N(\mu_i, \sigma_i^2)$, where μ_i is the mean income of such customers and σ_i^2 is their income variance.



University of Vaasa

Prof. Mohammed Elmusrati

82

82

Parametric Classification

- When we do not know $P(C_i)$ and $f(x|C_i)$, we estimate them from historical data and plug in their estimates to get the estimate for the discriminant function. We are given a sample

$$X = \{x^t, r^t\}_{t=1}^N$$

- Where $x \in \mathcal{X}$ is one-dimensional and $r \in \{0, 1\}^K$ such that

$$r_i^t = \begin{cases} 1 & \text{if } x^t \in C_i \\ 0 & \text{if } x^t \in C_k, k \neq i \end{cases}$$

- For each class separately, the estimates for the means and variances from data are (MLE estimates)

$$m_i = \frac{\sum_{t=1}^N x^t r_i^t}{\sum_{t=1}^N r_i^t}, \quad s_i^2 = \frac{\sum_{t=1}^N (x^t - m_i)^2 r_i^t}{\sum_{t=1}^N r_i^t}$$



University of Vaasa

Prof. Mohammed Elmusrati

83

83

Parametric Classification

- Moreover, the MLE estimates for the priors are

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N}$$

- Plugging these estimates into the previous discrimination function, we obtain

$$g_i(x) = -\frac{1}{2} \log(2\pi) - \log(s_i) - \frac{(x - m_i)^2}{2s_i^2} + \log(\hat{P}(C_i))$$

- We should select the class that has the maximum $g_i(x)$ value



University of Vaasa

Prof. Mohammed Elmusrati

84

84

Parametric Classification

- The first term is a constant and can be dropped because it is common in all $g_i(x)$. If the priors are equal, the last term can be also dropped.
- If we can further assume that variances are equal, then we can write

$$g_i(x) = -(x - m_i)^2$$

and thus we assign x to the class with the nearest mean:

Choose C_i if $|x - m_i| = \min_k |x - m_k|$



University of Vaasa

Mohammed S. Elmusrati – University of VAASA



85

Exercise

- Download the **Data1** file from the Moodle.
- It contains a one-dimensional data with 1000 entry with two classes. Estimate the mean and variance of the data and construct the discrimination function.
- Use only portion of the data for building the discrimination function and use the rest of the data to check the validity of your function.



University of Vaasa

Prof. Mohammed Elmusrati

86

86

Exercise

- Write a code consists of two parts:
- Part 1. Emulate the previous example of historical data about people that bought cars and their yearly income. We may assume 4 kinds of cars: KIA, TOYOTA, VOLVO, Ferrari.. You can generate list of people who bought KIA with $N(40,000, 30,000)$, i.e., their salary is normally distributed with an average of 40,000 Euro/year and variance of 30,000. TOYOTA $\rightarrow N(65,000, 30,000)$, VOLVO $\rightarrow N(85,000, 40,000)$, Ferrari. $\rightarrow N(280,000, 50,000)$. We may generate 100 customer of each kind.
- Part 2. Based on the available data from the above part, compute the discrimination function as in slide 82. Then generate another 500 random customers with random annual income and then classify them according to the function.
- Draw figure of the classification results.



University of Vaasa

Prof. Mohammed Elmusrati

87

87

Regression

- We have already studied the regression and derived some useful formulas about it (revise Part 2).
- Here, we show the relation between Regression based on minimizing the mean square error and Bayes' formula. However, we will not go through the derivations again.
- As we stated before: In regression, we would like to write the numeric output, called the *dependent variable*, as a function of the input, called the *independent variable*. We assume that the numeric output is the sum of a deterministic function of the input and random noise (represents modeling error + measurement noise+..):

$$\mathbf{y} = h(\mathbf{x}) + \mathbf{n}$$



University of Vaasa

Prof. Mohammed Elmusrati

88

88

Regression

- In the last equation, $h(x)$ is an unknown function, which we would like to approximate by our estimator, $g(x|\theta)$, defined up to a set of parameters θ . If we assume that n is zero-mean Gaussian with constant variance σ^2 , namely, $n \sim N(0, \sigma^2)$, and placing our estimator $g(\cdot)$ in place of the unknown function $h(\cdot)$, we have:

$$f(y|x) \sim N(g(x|\theta), \sigma^2)$$

- We again use maximum likelihood to learn the parameters θ . The pairs (x^t, y^t) in the training set are drawn from an unknown joint probability density $f(x, y)$, which we can write as

$$f(x, y) = f(y|x)f(x)$$

- Where $f(y|x)$ is the probability of the output given the input, and $f(x)$ is the input density



University of Vaasa

Prof. Mohammed Elmusrati

89

89

Regression

- Given an *iid* sample $X = \{x^t, y^t\}_{t=1}^N$, the log-likelihood is

$$L(\theta, X) = \log \prod_{t=1}^N f(x^t, y^t) = \log \prod_{t=1}^N f(y^t|x^t) + \log \prod_{t=1}^N f(x^t)$$

- We can ignore the second term since it does not depend on our estimator, and we have

$$L(\theta|X) = \log \prod_{t=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{[y^t - g(x^t|\theta)]^2}{2\sigma^2}\right) = -N \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{t=1}^N [y^t - g(x^t|\theta)]^2$$

- The first term is independent of the parameters θ and hence, could be dropped. Also, we can ignore the factor $1/\sigma^2$. Maximizing this likelihood function is equivalent to minimizing the function in the next slide:



University of Vaasa

Prof. Mohammed Elmusrati

90

90

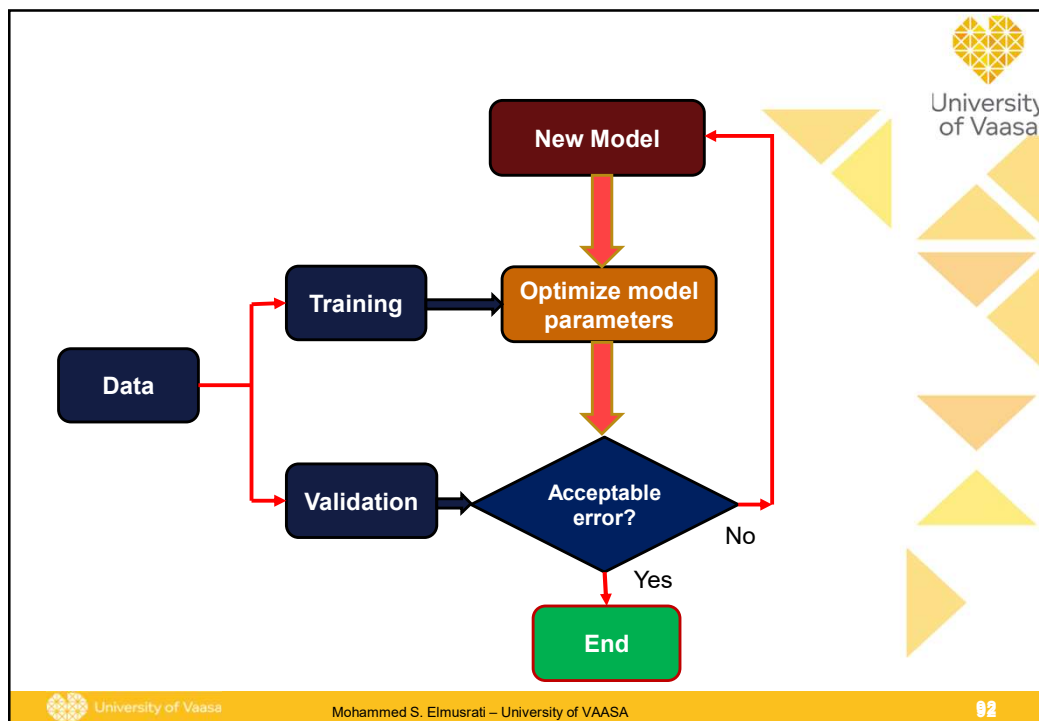
Regression

$$E(\theta|X) = \frac{1}{2} \sum_{t=1}^N [y^t - g(x^t|\theta)]^2$$

which is the most frequently used error function, and parameters θ that minimize it are called the *least squares estimates*.

- This is a transformation frequently done in statistics: When the likelihood L contains exponents, instead of maximizing L , we define an *error function*, $E = -\log L$, and minimize it.
- Now, we have precisely the same objective function to minimize as we have seen in Part 2 (Regression part).
- In the case of zero mean and independent noise samples, we will have the exact solution as those obtained in Part 2.

91



92

Tuning Model Complexity



University
of Vaasa

- Let us say that a sample $X = \{x^t, y^t\}$ is drawn from some unknown joint probability density $f(x, y)$. Using this sample, we construct our estimate $g(\cdot)$. The expected square error (over the joint density) at x can be written as:

$$E\left[\left(y - g(x)\right)^2 | x\right] = E\left[\left(y - E[y|x]\right)^2 | x\right] + \left(E[y|x] - g(x)\right)^2$$

- The first term on the right is the variance of r given x ; it does not depend on $g(\cdot)$ or X . It is the variance of noise added, σ^2 . This is part of the error that can never be removed, no matter what estimator we use. The second term quantifies how much $g(x)$ deviates from the regression function, $E[y|x]$. This does depend on the estimator and the training set.



University of Vaasa

Prof. Mohammed Elmusrati

93

93

Tuning Model Complexity



University
of Vaasa

- It may be the case that for one sample, $g(x)$ may be a very good fit; and for some other sample, it may make a bad fit. To quantify how well an estimator $g(\cdot)$ is, we average over possible datasets.
- The expected value {average over samples X , all of size N and drawn from the same joint density $f(y, x)$ } is

$$E_X\left[\left(E[y|x] - g(x)\right)^2\right] = \underbrace{\left(E[y|x] - g(x)\right)^2}_{\text{Bias}} + \underbrace{E_X\left[\left(E_X[g(x)] - g(x)\right)^2\right]}_{\text{Variance}}$$

- As we discussed before, bias measures how much $g(x)$ is wrong, disregarding the effect of varying samples, and variance measures how much $g(x)$ fluctuates around the expected value, $E[g(x)]$, as the sample varies. We want both to be small.



University of Vaasa

Prof. Mohammed Elmusrati

94

94

Tuning Model Complexity

- Usually, reducing the bias error could be achieved (if possible) at the cost of higher variance. In the same way, reducing variance will be achieved (if possible) at the cost of higher bias.
- **The optimal model is the one that has the best trade-off between the bias and the variance.**
- If there is bias, this indicates that our model class does not contain the solution; this is *underfitting*. If there is variance, the model class is too general and also learns the noise samples; this is *overfitting*.
- In practice, the method we use to find the optimal complexity is *cross-validation*.
- With validation, we obtain the total error; however, we do not directly know if it is due to biasing or variance!



University of Vaasa

Prof. Mohammed Elmusrati

95

95

Model Selection Procedures

- Usually, we cannot calculate bias and variance for a model separately, but we can calculate the total error. Given a dataset, we divide it into two parts as training and validation sets, train candidate models of different complexities. We test their error on the validation set left out during training.
- As the model complexity increases, training error keeps decreasing. The error on the validation set decreases up to a certain level of complexity, then stops decreasing or does not decrease further significantly, or even increases if there is significant noise. We have already seen these effects in Part 2.



University of Vaasa

Prof. Mohammed Elmusrati

96

96

Model Selection Procedures



University
of Vaasa

- Another approach that is used frequently is **regularization**.
- In this approach, we write an *augmented error function*

$$E' = (\text{model complexity}) \times \lambda + (\text{error in data}) \times (1 - \lambda), 0 \leq \lambda \leq 1$$

- This has a second term that penalizes complex models with significant variance, where λ gives the weight of this penalty. When we minimize the augmented error function instead of the error on data only, we penalize complex models and thus decrease variance.
- If λ is closer to 1, only very simple models are allowed. There will be a high risk of large bias. On the other hand, when λ is closer to zero, we would have a higher risk of the overfitting problem (i.e., memorization rather than generalization).
- Therefore, λ should be optimized using cross-validation.



University of Vaasa

Prof. Mohammed Elmusrati

97

97

Model Selection Procedures



University
of Vaasa

- Another way is by regarding E' in the previous slide as the error on new test data. The second term on the right is the training error, and the first is an *optimism* term estimating the discrepancy between training and test error.
- Methods such as *Structural risk minimization* (SRM) use a set of models ordered in terms of their complexities. An example is polynomials of increasing order.
- The number of free parameters generally gives the complexity. In E' , we can have a set of decreasing λ_i to get a set of models ordered in increasing complexity. Model selection by SRM then corresponds to finding the model simplest in terms of order and best in terms of an empirical error on the data.



University of Vaasa

Prof. Mohammed Elmusrati

98

98

Model Selection Procedures



University
of Vaasa

- *Minimum Description Length* (MDL) uses an information-theoretic measure. *Kolmogorov complexity* of a dataset is defined as the shortest description of the data.
- If the data is simple, it has a short complexity; for example, if it is a sequence of '0's, we can just write '0' and the length of the sequence. For example, if data consists of a thousand "a" followed by a thousand "z", the shortest description could be "1000 (a) followed by 1000 (z)".
- If the data is completely random (uncorrelated), then we cannot have any description of the data shorter than the data itself. If a model is appropriate for the data, then it has a good fit for the data, and instead of the data, we can send/store the model description. Out of all the models that describe the data, we want to have the most straightforward model for lending itself to the shortest description. So we again have a trade-off between how simple the model is and how well it explains the data.



University of Vaasa

Prof. Mohammed Elmusrati

99

99

Bayesian Model Selection



University
of Vaasa

- *Bayesian model selection* is used when we have some prior knowledge about the appropriate class of approximating functions. This prior knowledge is defined as a prior distribution over models, $P(\text{model})$.
- Given the data and assuming a model, we can calculate $P(\text{model}|\text{data})$ using Bayes' rule:

$$P(\text{model}|\text{data}) = \frac{P(\text{data}|\text{model})P(\text{model})}{P(\text{data})}$$

- Where $P(\text{model}|\text{data})$ is the posterior probability of the model given our prior subjective knowledge about models, namely, $P(\text{model})$, and the objective support provided by the data, namely, $P(\text{data}|\text{model})$. We can then choose the model with the highest posterior probability or take average overall models weighted by their posterior probabilities.



University of Vaasa

Prof. Mohammed Elmusrati

100

100



University of Vaasa

Multivariate Data




University of Vaasa

Mohammed S. Elmusrati – University of VAASA

101

101



University of Vaasa


Multivariate Data

- Usually, we deal with many attributes as well as observations. We want to build our parameter estimation of each observation and to study the relations (can be hidden) between them.
- The sample may be viewed as a *data matrix*

$$\mathbf{X} = \begin{bmatrix} X_1^1 & X_2^1 & \dots & X_d^1 \\ X_1^2 & X_2^2 & \dots & X_d^2 \\ \vdots & \vdots & \ddots & \vdots \\ X_1^N & X_2^N & \dots & X_d^N \end{bmatrix}$$

where d columns correspond to d *variables* denoting the result of measurements made on an individual or event. These are also called *inputs*, *features*, or *attributes*. The N rows correspond to independent and identically distributed *observations*, *examples*, or *instances* of N individuals or events.

- If rows are dependent, how could we deal with it?



University of Vaasa

Prof. Mohammed Elmusrati

102

102

Multivariate Data

- Assume a Bank wants to evaluate the reliability (or trust) of customers applying for a loan. The bank has a record of previous customers, and they may use this record to build some reliable tests that could be applied to new customers.
- Apply this example in the previous matrix, we have N different recorded cases, and " d " is the number of attributes. For example, X_1^i can be the "age" of the i^{th} customer, X_2^i can be the "marital status", X_3^i can be "yearly income", X_4^i can be the "nationality" and so on. All available information that we think has an impact on the trust should be used (there might be ethical issues here!).
- These measurements may be of different scales, for example, age in years and yearly income in monetary units. Some like age may be numeric, and some like marital status or nationality may be discrete.



University of Vaasa

Prof. Mohammed Elmusrati

103

103

Multivariate Data

- Typically, these variables are correlated. If they are not, there is no need for multivariate analysis. If data are independent, we can treat them separately.
- Our aim may be a *simplification*, that is, summarizing this large body of data by means of relatively few parameters (for example, mean, correlation factors, etc.).
- Or our aim may be *exploratory*, and we may be interested in generating hypotheses about data. In some applications, we are interested in predicting the value of one variable from the values of other variables. If the predicted variable is discrete, this is multivariate classification, and if it is numeric, this is a multivariate regression problem.



University of Vaasa

Prof. Mohammed Elmusrati

104

104

Multivariate Parameter Estimation



University
of Vaasa

- The *mean vector* $\boldsymbol{\mu}$ is defined such that each of its elements is the mean of one column of \mathbf{X} :

$$E[\mathbf{x}] = \boldsymbol{\mu} = [\mu_1, \dots, \mu_d]^T$$

- In the loan example, X_1 corresponds to the ages of all old customers. Therefore, μ_1 is the average age of bank customers.
- The variance of X_i is denoted as σ_i^2 , and the covariance of two variables X_i and X_j is defined as

$$\sigma_{ij} \equiv \text{Cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = E[X_i X_j] - \mu_i \mu_j$$

- Where $\sigma_{ij} = \sigma_{ji}$, and when $i = j$, $\sigma_{ii} = \sigma_i^2$.



University of Vaasa

Prof. Mohammed Elmusrati

105

105

Multivariate Parameter Estimation



University
of Vaasa

- With d variables, there are d variances and $d(d-1)/2$ covariances, which are generally represented as a $d \times d$ matrix, named the *covariance-matrix*, denoted as $\boldsymbol{\Sigma}$, whose $(i,j)^{\text{th}}$ element is σ_{ij} :

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

- The diagonal terms are the variances, the off-diagonal terms are the covariances, and the matrix is symmetric.

$$\boldsymbol{\Sigma} \equiv \text{Cov}(\mathbf{X}) = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = E[\mathbf{X}\mathbf{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T$$



University of Vaasa

Prof. Mohammed Elmusrati

106

106

Multivariate Parameter Estimation



University
of Vaasa

- If two variables are related linearly, then the covariance will be positive or negative depending on whether the relationship has a positive or negative slope.
- But the size of the relationship is difficult to interpret because it depends on the units in which the two variables are measured.
- The *correlation* between variables X_i and X_j is a statistic normalized between -1 and $+1$, defined as

$$\text{Corr}(X_i, X_j) \equiv \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

- If two variables are independent, then their covariance, and hence their correlation, is **0**. However, the converse is not true: the variables may be dependent (in a nonlinear way), and their correlation maybe 0 (can you give an example?).



University of Vaasa

Prof. Mohammed Elmusrati

107

107

Multivariate Parameter Estimation



University
of Vaasa

- Given a multivariate sample, estimates for these parameters can be calculated: The maximum likelihood estimator for the mean is the *sample mean*, \mathbf{m} . Its i^{th} dimension is the average of the i^{th} column of \mathbf{X} :

$$\mathbf{m} = \frac{\sum_{t=1}^N \mathbf{x}^t}{N} \quad \text{with} \quad m_i = \frac{\sum_{t=1}^N x_i^t}{N}, \quad i = 1, \dots, d$$

- The estimator of $\mathbf{\Sigma}$ is \mathbf{S} , the *sample covariance matrix*, with entries

$$s_i^2 = \frac{\sum_{t=1}^N (x_i^t - m_i)^2}{N}, \quad \text{and} \quad s_{ij} = \frac{\sum_{t=1}^N (x_i^t - m_i)(x_j^t - m_j)}{N}$$

- Actually, we should have divided by $(N-1)$ to obtain the unbiased covariance estimation. However, for large N , results will be pretty similar.



University of Vaasa

Prof. Mohammed Elmusrati

108

108

Multivariate Parameter Estimation

- The *sample correlation* coefficients are

$$r_{ij} = \frac{s_{ij}}{s_i s_j}$$

and the sample correlation matrix **R** contains r_{ij} .



University of Vaasa

Prof. Mohammed Elmusrati

109

109

Example

- Assume the following table of students' grades in different courses and their gender

Student	Gender	Math	Physics	Biology
S1	M	87	75	65
S2	F	74	63	96
S3	M	89	90	88
S4	F	72	60	92
S5	M	78	74	85

- Compute the mean and Covariance matrix by hand (to understand how they are calculated). What is the correlation between gender and courses, and also the correlation between math and physics?



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

110

110

Example

- Let's define $M=1$ and $F=-1$. The mean vector is:

$$\mathbf{m} = \frac{[1 \ 87 \ 75 \ 65] + [-1 \ 74 \ 63 \ 96] + [1 \ 89 \ 90 \ 88] + [-1 \ 72 \ 60 \ 92] + [1 \ 78 \ 74 \ 85]}{5} = [0.2 \ 80 \ 72.4 \ 85.2]$$

- The elements of the covariance matrix is computed as

$$S_{11} = \frac{\sum_{k=1}^5 (x_1^k - 0.2)^2}{4} = \frac{(1-0.2)^2 + (-1-0.2)^2 + (1-0.2)^2 + (-1-0.2)^2 + (1-0.2)^2}{4} = 1.2$$

$$S_{12} = \frac{\sum_{k=1}^5 (x_1^k - 0.2)(x_2^k - 80)}{4} = \frac{(1-0.2)(87-80) + (-1-0.2)(74-80) + (1-0.2)(89-80) + (-1-0.2)(72-80) + (1-0.2)(78-80)}{4} = 7$$

- We continue the computations in the same way

$$\mathbf{S} = \begin{bmatrix} 1.2 & 7 & 10.9 & -8.8 \\ 7 & 58.5 & 82.25 & -58.75 \\ 10.9 & 82.25 & 140.3 & -47.35 \\ -8.8 & -58.75 & -47.35 & 144.7 \end{bmatrix}$$

111

Example


- We may compute the correlation matrix as

$$\mathbf{R} = \begin{bmatrix} 1.00000 & 0.83547 & 0.84005 & -0.66782 \\ 0.83547 & 1.00000 & 0.90788 & -0.63855 \\ 0.84005 & 0.90788 & 1.00000 & -0.33232 \\ -0.66782 & -0.63855 & -0.33232 & 1.00000 \end{bmatrix}$$

$$r_{ij} = \frac{S_{ij}}{S_i S_j}$$


- From the correlation matrix, $r_{12}=0.835$ indicates a positive correlation between gender and mathematics. Since Male has been defined as +1, this indicates that male is doing better in math. However, $r_{14}=-0.667$, which indicates females are doing better in biology. Furthermore, $r_{23}=0.908$ indicates a high positive correlation between math and physics.

112



Estimation of Missing Values

- Frequently, values of certain variables may be missing in observations. The best strategy is to discard those observations altogether, but if we do not have large enough samples to be able to afford this and we do not want to lose data as the non-missing entries do contain information. We try to fill in the missing entries by estimating them. This is called *imputation*.
- In *mean imputation*, for a numeric variable, we substitute the mean (average) of the available data for that variable in the sample. For a discrete variable, we fill in with the most likely value, that is, the value most often seen in the data.
- In *imputation by regression*, we try to predict the value of a missing variable from other variables whose values are known for that case. Depending on the type of the missing variable, we define a separate regression or classification problem that we train by the data points for which such values are known. If many different variables are missing, we take the means as the initial estimates, and the procedure is iterated until predicted values stabilize. If the variables are not highly correlated, the regression approach is equivalent to mean imputation.




University of Vaasa

Prof. Mohammed Elmusrati


113

113



Estimation of Missing Values

- Using Kalman filters is one very well-known method to estimate missing samples.
- Depending on the context, however, sometimes the fact that a particular attribute value is missing may be essential and cannot be just estimated. For example, if the applicant does not declare their telephone number (full or partial), that may be a critical piece of information. In such cases, this is represented as a separate value to indicate that the value is missing and is used as such.
- Or maybe in some applications, we need the credit card number. If some of the numbers are missing (e.g., validity date), it is not possible to estimate them.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

114

114

Multivariate Classification



University
of Vaasa

- When $\mathbf{x} \in \mathbb{R}^d$, if the class-conditional densities, $f(\mathbf{x}|C_i)$, are taken as normal density, $N_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, we have (Revise Part 3)

$$f(\mathbf{x}|C_i) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_i|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right]$$

- The main reason for this assumption is its analytical simplicity. Besides, the Normal density is a model for many naturally occurring phenomena in that examples of most classes can be seen as mildly changed versions of a single prototype, $\boldsymbol{\mu}_i$, and the covariance matrix, $\boldsymbol{\Sigma}_i$, denotes the amount of noise in each variable and the correlations of these noise sources.
- While actual data may not often be precisely multivariate Normal, it is a valid approximation in many situations.



University of Vaasa

Prof. Mohammed Elmusrati

115

115

Multivariate Classification



University
of Vaasa

- Let us say that we want to predict the type of a car that a customer would be interested in. Different cars are the classes, and \mathbf{x} are observable data of customers, for example, age and income. $\boldsymbol{\mu}_i$ is the vector of mean age and income of customers who buy car type i and $\boldsymbol{\Sigma}_i$ is their covariance matrix: σ_{i1}^2 and σ_{i2}^2 are the age and income variances, and σ_{i12}^2 is the covariance of age and income in the group of customers who buy car type i .
- When we define the discriminant function as $g_i(\mathbf{x}) = \log f(\mathbf{x}|C_i) + \log P(C_i)$, and assuming $f(\mathbf{x}|C_i) \sim N_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, we have

$$g_i(\mathbf{x}) = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Sigma}_i| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \log P(C_i)$$



University of Vaasa

Prof. Mohammed Elmusrati

116

116

Multivariate Classification

- We should find the parameters of the previous discrimination function.
- Given a training sample for $K \geq 2$ classes, $X = \{\mathbf{x}^t, \mathbf{r}^t\}$, where $r_i^t = 1$ if $\mathbf{x}^t \in C_i$ and 0 otherwise, estimates for the means and covariances are found using maximum likelihood separately for each class and then substituted in $g_i(\mathbf{x})$ function:

$$\begin{aligned}\hat{P}(C_i) &= \frac{\sum_t r_i^t}{N} \\ \mathbf{m}_i &= \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t} \\ \mathbf{S}_i &= \frac{\sum_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^T}{\sum_t r_i^t}\end{aligned}$$

117

Multivariate Classification

- We obtain: $g_i(\mathbf{x}) = -\frac{1}{2} \log |\mathbf{S}_i| - \frac{1}{2} (\mathbf{x} - \mathbf{m}_i)^T \mathbf{S}_i^{-1} (\mathbf{x} - \mathbf{m}_i) + \log \hat{P}(C_i)$

This could be expanded to:

$$g_i(\mathbf{x}) = -\frac{1}{2} \log |\mathbf{S}_i| - \frac{1}{2} (\mathbf{x}^T \mathbf{S}_i^{-1} \mathbf{x} - 2 \mathbf{x}^T \mathbf{S}_i^{-1} \mathbf{m}_i + \mathbf{m}_i^T \mathbf{S}_i^{-1} \mathbf{m}_i) + \log \hat{P}(C_i)$$

- which defines a *quadratic discriminant*, it could be rewritten as

$$g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

where

$$\mathbf{W}_i = -\frac{1}{2} \mathbf{S}_i^{-1}$$

$$\mathbf{w}_i = \mathbf{S}_i^{-1} \mathbf{m}_i$$

$$w_{i0} = -\frac{1}{2} \mathbf{m}_i^T \mathbf{S}_i^{-1} \mathbf{m}_i - \frac{1}{2} \log |\mathbf{S}_i| + \log \hat{P}(C_i)$$

118

Example

- In the previous example in Slide 110, find \mathbf{m}_i and \mathbf{S}_i for Male and Female.
- Based on that, find the discriminant function for both males and females.
- We may look to these functions as we teach the machine how will be the performance of males and females in courses.
- You can test the machine by applying only the grades of students, and the machine should guess if the student is male or female.
- The machine's performance would depend on the data size and if the assumption of Normally distribution is accurate.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

119

119

Multivariate Classification: Few Samples

- The number of parameters to be estimated are $K \cdot d$ for the means and $K \cdot d(d+1)/2$ for the covariance matrices. When d is large, and samples are small, \mathbf{S}_i can be singular, and its inverse does not exist. Or, $|\mathbf{S}_i|$ may be nonzero but too small or even negative!, in which case it will be unstable; small changes in \mathbf{S} will cause large changes in \mathbf{S}^{-1} (ill-conditioned matrix). For the estimates to be reliable on small samples, one may want to decrease dimensionality, d , by redesigning the feature extractor and selecting a subset of the features or somehow combining existing features (as we will explain in the next part).
- Another possibility is to pool the data and estimate a common covariance matrix for all classes:

$$\mathbf{S} = \sum_i \hat{P}(C_i) \mathbf{S}_i$$

- In this case of equal covariance matrices, the equation reduces to

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \mathbf{S}^{-1}(\mathbf{x} - \mathbf{m}_i) + \log \hat{P}(C_i)$$



University of Vaasa

120

120

Multivariate Classification: Few Samples



University
of Vaasa

- The number of parameters is $K \cdot d$ for the means and $d(d+1)/2$ for the shared covariance matrix. If the priors are equal, the optimal decision rule is to assign input to the class whose mean's Mahalanobis distance to the input is the smallest.
- As before, unequal priors shift the boundary toward the less likely class. Note that in this case, the quadratic term $\mathbf{x}^T \mathbf{S}^{-1} \mathbf{x}$ cancels since it is common in all discriminants, and the decision boundaries are linear, leading to a *linear discriminant* that can be written as

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

where

$$\mathbf{w}_i = \mathbf{S}^{-1} \mathbf{m}_i$$

$$w_{i0} = -\frac{1}{2} \mathbf{m}_i^T \mathbf{S}^{-1} \mathbf{m}_i + \log \hat{P}(C_i)$$



University of Vaasa

Prof. Mohammed Elmusrati

121

121

Classification Implementation Example



University
of Vaasa

- We have designed a randomly correlated database for explanation purposes. It has been formulated as an Excel file.
- The data consists of 7 attributes as Income [yearly], Age [years], Weight [kg], Gender [2 categories], Education [as 6 categories], Health [10 categories], Height [cm].
- Users have been classified in the database. The classification could be, for example, honesty, success in personal life, number of crimes, car accidents, etc..
- However, in this example, we assumed 13 different classes.
- Screenshot of the database is given in the next slide.
- We aim to build a Machine learning code that extracts the knowledge presented by the database and uses it to classify new users with new attributes.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

122

122

Parametric Classification Example



University
of Vaasa

File name is Database2.xlsx

We assumed 13 different classes from -2 to 10.

There are 10,000 entry samples

123

Classification Implementation Example



University
of Vaasa

- In this simple implementation, we have used Octave. However, any other programming language could be used as well.
- The main function to read the excel file is **xlsread** and the main function to write and save in excel format is **xlswrite**
- Next slide shows how we could upload the excel file.
- The number of samples in the file is 10,000. We use the first 5000 samples to teach our system about the means, covariance matrices, and the prior probability of each class.
- Then, we tested our system with 101 elements that were not shown before and computed the error percentage.
- Once we are satisfied with the functionality of our algorithm, we should repeat the computation of the parameters (means, covariance matrices, and priori probabilities for each class) using all available data.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

124

124

Classification Implementation Example



University
of Vaasa

```

1 - clear all
2 - % let's read the first 5000 inputs
3 - N=5000;
4 - AR=xlsread('Database2.xlsx','Database1','A2:H5001');
5 - Class=AR(:,8);
6 - s=0;
7 - classNum=-2:10; %Available classes
8 - m=length(classNum);
9 - for i = classNum,
10 -     s=s+1;
11 -     D(s)=AR(find(Class==i),1:7);
12 -     Smean(s)=mean(D{s}); % Mean of each class
13 -     Scov(s)=cov(D{s}); % Covariance of each class
14 -     n(s)=length(find(Class==i));
15 -     Ps(s)=n(s)/N; % A priori probability of each class
16 - end
17 - % validation set consists of 101 elements
18 - x=xlsread('Database2.xlsx','Database1','A8400:G8500');
19 -
20 - % the results of the validation set from the database
21 - xx=xlsread('Database2.xlsx','Database1','H8400:H8500');
22 -
23 - K=length(xx);

```



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

125

125

Classification Implementation Example



University
of Vaasa

```

24 - %building Discrimination Functions
25 - for k=1:K,
26 -     s=0;
27 -     for i=classNum,
28 -         s=s+1;
29 -
30 -         Mean=Smean{s};
31 -         Cov=Scov{s};
32 -         g(s)=-0.5*log(det(Cov))-0.5*(x(k,:)-Mean)*pinv(Cov)*(x(k,:)-Mean)'+log(Ps(s));
33 -     end
34 -     % desciding the class of each input k
35 -     c=find(g==max(g));
36 -     R(k)=classNum(c);
37 - end
38 - % error probability in miss-classification
39 - %Pmc=length(find(R!=xx))/101;
40 - Pmc=length(find(R~=xx))/101;

```




University of Vaasa

Mohammed S. Elmusrati – University of VAASA

126

126

Classification Implementation Example




University of Vaasa

- Many other interesting results could be obtained from the previous given code. For example, we can study the correlation between each attribute to any other attribute in each class.
- Based on the Database 2 file, what is the correlation factor between gender and income in Class number 0?
- We can easily compute the correlation element as shown on Slide 108 as:


$$\text{Scov}\{3\}(1,4)/\sqrt{\text{Scov}\{3\}(1,1)*\text{Scov}\{3\}(4,4)}$$
- ans = -0.14714
- Answer shows that the correlation is negative, it means that the salary goes slightly higher for Gender (-1). However, the relation is small

Classes= -2,-1,0,1,2,..., 10.
Therefore, the number of class 0 on the sequence is 3


University of Vaasa
Mohammed S. Elmusrati – University of VAASA
127


127

Classification Implementation Example



University of Vaasa

- Based on the Database 2 file, what is the correlation factor between age and health in Class number 10?
- $$\text{Scov}\{13\}(2,6)/\sqrt{\text{Scov}\{13\}(2,2)*\text{Scov}\{13\}(6,6)}$$
- ans = -0.32784
- Hence, besides the learning outcome, we can also have some statistical inferences about the data in each class.
- This data analysis could be very important in some applications.


University of Vaasa
Mohammed S. Elmusrati – University of VAASA
128

128

Exercise 3

- The same excel file used in the previous example has been uploaded on Moodle.
- Download the file and use the code to generate the same results.
- Study the code carefully and try to modify it in order to have applications.
- Use the same code with other excel files, maybe downloaded from some databases.



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

129

129

Exercise 4

- Download the Data_Part4_4D.txt file from Moodle. It contains certain four dimensional data with 1000 entry with three classes labeled (1,2,3).
- You may divide the data into two parts: training set and validation set.
- Estimate the mean vector and the covariance matrix of the training set of the data and construct the discrimination function.
- Try to use different sizes of the training part and check the performance based on the average error of validation set.
- Construct function based on the previous part that when you enter 4 elements, it will automatically classify it within one of the three classes.



University of Vaasa

Prof. Mohammed Elmusrati

130

130

Exercise 5

- Write a code consists of two parts:
- Part 1. Emulate the previous example of the exercise in Slide 32 of historical data about people that bought cars and their yearly income but add also their ages and gender.
- Part 2. Based on the available data of Part 1, compute the discrimination function in slide 56. Then generate 500 random customers with random annual income and then classify them according to the function.
- Draw figures of the classification results.



University
of Vaasa



University of Vaasa

Prof. Mohammed Elmusrati

131

131

THANK YOU



University
of Vaasa



University of Vaasa

Mohammed S. Elmusrati – University of VAASA

132

132