People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

N° d'ordre :
N° de série :

KASDI MERBAH UNIVERSITY OUARGLA

Mathematics Department

MASTER Memoir in Mathematics

Speciality : Probability and Statistics

Prepared by : Ghedamsi Oumessaad

Theme:

SUPPORT VECTOR DENSITY ESTIMATION

Represented in : 29/06/2021

Jury committee:

Meflah Mebrok	Kasdi Merbah University-Ouargla	Chairman
Boussaad Abdelmalek	Kasdi Merbah University-ouargla	Supervisor
Mansoul Brahim	Kasdi Merbah University-Ouargla	Examiner
Kouidri Mohammed	Kasdi Merbah University-Ouargla	Examiner

بسم الله الرحمان الرحيم

الحمد لله على نعمة العلم و على الوصول الى ما أنا عليه. أود شكر كل من دعمني وخاصة من أهلي أمي ، أبي و جدتي. من أساتذتي كاليش كلثوم، بوسعد عبد المالك ومزابية عبد الهادي. كما أشكر أختان عزيزتان على قلبي ايمان قمولة و رشيدة بلعباس على كل ماقدماه لي.

List of Abbreviations

COP	Convex Optimization Problem
SPSD	Symmetric Positive Semi-Definite
LP	Linear Programming
QP	Quadratic Programming
KKT	Karush-Kuhn-Tucker
SVM	Support Vector Machines
SVC	Support Vector Classification
SVR	Support Vector Regression
SV_s	Support Vectors
SV	The set of support vectors
RBF	Radial Basis Function
iid	independently and identically distributed
PDF	Probability Density Function
CDF	Cumulative Distribution Function
EDF	Empirical Distribution Function

Contents

List of Figures

Introduction

There are three types of machine learning: supervised, unsupervised and reinforcement. In supervised learning [10], the training data set includes the pairs (x,y) such that: x is the input and y is the output. As shown in [10], the supervised learning is divided into two categories, regression in the case of outputs are continuous values, and classification in the case of outputs are discrete values. During the process of supervised learning the dataset divides into two parts, the first part is used for learning in the aim to build a mathematical model between the inputs and outputs, the second part for testing the model to predict the output of test data set.

Statistical learning Theory (SLT) is a basic theoretical tool of the problem of function estimation a given set of data and also for developing practical algorithms for estimating multi-dimensional functions. As shown in [12] the model of learning can be described as follows: Given a set of n training data $D = \{(x_1, y_1), ..., (x_n, y_n)/x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$, with independently and identically distributed (iid) unknown probability distribution $p(x, y) = p(x)p(y/x)$. The problem of learning is that of choosing from a given set of functions $\{f(x, \alpha)/\alpha \in H\}$ the one which predicts the supervisor's response in the best possible way. In order to select the best approximation, the following expected value:

$$E(L(y, f(x, \alpha)) = \int L(y, f(x, \alpha))dp(x, y)$$

is chosen as a criterion to be minimized over the class of functions $\{f(x, \alpha)/\alpha \in H\}$, with $L(y, f(x, \alpha))$ is a loss function that measure the error between y_i and $f(x_i, \alpha)$.

Support Vector Machines (SVM) is a supervised learning algorithm developed in the framework of (SLT) by V.Vapnik in 1995. It is widely used for solving classification problems. The idea of SVM is to find an optimal hyperplane which is equivalently to solve a learning problem, where $f(x) = w^T\phi(x) + b$ with w, b are the parameters of the machine and ϕ is a mapping from $\mathbb{R}^m$ to $\mathbb{R}^{m'}$ $(m' > m)$. This kind of algorithm is used in different applications such as image segmentation, regression problem ...etc. In our dissertation, we focus on the support vector regression (SVR) for estimating the cumulative distribution function (CDF) which is proposed by [4]. Concerning this manuscript is organized as follows.

- **Chapter 01: Preliminaries**

  This chapter is divided into two parts:

  Part01:

  In this part, the Hilbert spaces notion and the optimization problems with some theories are presented.

  Part02:

  Probability theory and the estimation of the cumulative distribution function are detailed in this part.

- **Chapter 02: Support Vector Machines**

  In this chapter, we define the support vector machines (SVM) for classification and regression as a princpal tool in the estimation of the probability density function (PDF).

- **Chapter 03: Support Vector Density Estimation**

  In this final chapter, we try to detail a technique [4] for estimating the cumulative distribution function which is based on **the empirical distribution function and the support vector machines (SVM)**.

  Finally, a conclusion is written to summarize this work.

Chapter 1

Preliminaries

Part I

Optimization

1.1 Reminder

1.1.1 Hilbert Space

Definition 1.1.1 *[10] A space $(X, <, >)$ is called an inner product where $<, >$ the application from $X \times X$ in $\mathbb{R}_+$ verifies for all $x, y, z \in X$ and $\alpha \in \mathbb{R}$:*

- $<x, y> = <y, x>$.
- $<x, x> = 0 \Leftrightarrow x = 0$.
- $<\alpha x, y> = \alpha <x, y>$.
- $<x + y, z> = <x, z> + <y, z>$.

The quantity $<x, y>$ is called the inner product of x and y and the couple $(X, <, >)$ is called a semi-Hilbertian sapce.

Remark 1.1.1 *The application $||.||$ from X to $\mathbb{R}$:*

$$||x|| = \sqrt{<x, x>},$$

call norm and $(X, ||.||)$ call normed space.

Remark 1.1.2 *If $X = \mathbb{R}^m$ and for $x \in \mathbb{R}^m$:*

$$||x|| = \sqrt{\sum_{i=1}^{m} x_i^2}.$$

Definition 1.1.2 *([10],[8]) A Hilbert space is complete (every element from this space is the limit of suite of cauchy converges) separable inner product space.*

Definition 1.1.3 *[8] We call $(X_n)_{n \in N}$ is the suite of cauchy if:*

$$\forall \epsilon > 0, \exists n_\epsilon \in N / \forall n, m \geq n_\epsilon \longrightarrow ||X_n - X_m|| < \epsilon.$$

Definition 1.1.4 *[8] We call the function f is integrable noted $f \in L^1(\mathbb{R}^n)$, if:*

$$\int_{\mathbb{R}^n} |f(x)|dx < \infty, \quad \forall x \in \mathbb{R}^n.$$

1.1.2 Convexity and Matrixes

Definition 1.1.5 *[10] A set $M \subseteq \mathbb{R}^n$ is convex if:*

$$tx + (1-t)y \in M, \quad \forall x, y \in \mathbb{R}^n, \ \forall t \in (0;1).$$

Definition 1.1.6 *[10] A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if:*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y), \quad \forall x, y \in \mathbb{R}^n, \ \forall \lambda \in [0;1].$$

Remark 1.1.3 *If this inequality be equal then the function call linear.*

Remark 1.1.4 *A linear function from a Hilbert space to anthon call linear operator.*

Definition 1.1.7 *[10] A matrix H is symmetric if and only if:*

$$H^T = H.$$

Definition 1.1.8 *[10] A symmetric matrix H is positive semi-definite if and only if:*

$$x^T Hx \geq 0, \quad \forall \ x \in \mathbb{R}^n, \ x \neq 0.$$

1.1.3 The convex optimization problem

Definition 1.1.9 *[10] Optimization problem is a technique to find the optimal (best) points (max or min) of the objective function subject to some constraints.*

Definition 1.1.10 *[10] The objective function is the value you are training to optimize (op-*

timal decision, minimal error,...).

Definition 1.1.11 *[10] Constraints set boundaries for where the optimizer cannot go, there are two types (equality and inequality) constraints.*

Definition 1.1.12 *[7] The variables of the problem can be of various nature (real, integer, boolean,...) and are expressed from qualitative or quantitative data.*

Definition 1.1.13 *[10] The feasible region is the domain where the objective function is defined, we will denote it by a convex subset:*

$$M = \{x \in \mathbb{R}^n / \ the \ constraints \ are \ satisfied \ \}. \tag{1.1}$$

Definition 1.1.14 *[10] The optimization problem is called a linear programme (LP) if the objective function and the constraints are linear functions, if the objective is quadratic function and the constraints are linear, then the optimization problem call quadratic programme (QP).*

Problem 1.1.1 *[10] Finding x the solution of the convex optimization problem:*

$$\begin{cases} & \min\limits_{x \in M} f(x) \\ subject\ to & \\ & g_i(x) \leq 0, \ i = 1, ...n \\ & h_i(x) = 0, \ i = 1, ...m \end{cases} \tag{1.2}$$

where the functions are convex.

Remark 1.1.5 *To go from maximum to minimum, we need to change the sign of the objective function.*

Remark 1.1.6 *As the objective function and all the constraints are convex, together with convexity of the feasible region so the optimization problem is convex.*

1.2 Optimization Theory

1.2.1 Lagrange Formulation

Definition 1.2.1 *[10] To solve an optimization problem* (1.2)*, it suffices to search for a stationary point of the lagrangian function given by:*

$$L(x, \alpha, \beta) = f(x) + \sum_{i=1}^{n} \alpha_i g_i(x) + \sum_{i=1}^{m} \beta_i h_i(x), \tag{1.3}$$

where the coefficient α and β are called the Lagrange multipliers.

Theorem 1.2.1 *Let (x^*, α^*, β^*) be a feasible solution of the lagrangien function, then:*

$$L(x^*, \alpha^*, \beta^*) \leq f(x^*), \tag{1.4}$$

where $x^* \in M$ is a feasible solution of the primal problem (1.2)*.*

Proof. [10] From the definition of $L(x^*, \alpha^*, \beta^*)$, we have:

$$\begin{aligned} L(x^*, \alpha^*, \beta^*) &= f(x^*) + \sum_{i=1}^{n} \alpha_i^* g_i(x^*) + \sum_{i=1}^{n} \beta_i^* h_i(x^*), \\ &\leq f(x^*). \end{aligned}$$

Because the only point verifie for all i, $g_i(x) \leq 0$ and $h_i(x) = 0$ are the solution of the primal optimisation problem. ■

1.2.2 Fermat's Theorem

Theorem 1.2.2 *For a point to be a minimum or maximum of function continuously differentiable its derivative is null, this sufficient condition with convexity of the function.*

Proof. Suppose f defined on the interval $[a; b]$, and let $c \in [a; b]$, we proof that c is the extremi point (max).

Since f differentiable on c (the derivative exists) we have that:

$$f^{'}(c) = \lim_{h \longrightarrow 0} \frac{f(c+h) - f(c)}{h},$$

we have: $\displaystyle\lim_{h \longrightarrow 0^{-}} \frac{f(c+h) - f(c)}{h} \geq 0$ and $\displaystyle\lim_{h \longrightarrow 0^{+}} \frac{f(c+h) - f(c)}{h} \leq 0$,

then

$$0 \leq f^{'}(c) \leq 0 \Rightarrow f^{'}(c) = 0$$

■

1.2.3 Karush-Kuhn-Tucker Conditions

Theorem 1.2.3 *If x^* minimizes the problem* (1.2)*, then there exists lagrange multipliers α^* and β^* (vectors) such that the following:*

$$\begin{aligned}
\min_{x} L(x, \alpha^*, \beta^*) &= \max_{\alpha,\beta} L(x^*, \alpha, \beta) = L(x^*, \alpha^*, \beta^*). && (1.5)\\
\alpha &\geq 0, \quad \beta \geq 0. && (1.6)\\
\alpha_i^* g_i(x^*) &= 0, \quad i = 1, ...n. && (1.7)\\
\beta_i^* h_i(x^*) &= 0, \quad i = 1, ...m. && (1.8)
\end{aligned}$$

The last two equations call KKT complementary conditions.

Part II

Statistic

1.3 Reminder

1.3.1 Probability Space

Definition 1.3.1 *[8] We call to the family $\mathcal{A}$ σ-algebra (tribe) on the non empty set Ω if $\mathcal{A}$ containing the empty set, stable by complementary and countable union. The couple $(\Omega, \mathcal{A})$ is called measurable space.*

Exemple 1.3.1 *Borelian σ-algebra of $\mathbb{R}$ is the smallest tribe containing all open intervals, noted $\mathcal{B}_R$.*

Definition 1.3.2 *[8] Let $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$ be two measurable spaces. The function $f : \Omega_1 \longrightarrow \Omega_2$ is said to be measurable if:*

$$f^{-1}(B) \in \mathcal{A}_1, \quad \forall B \in \mathcal{A}_2.$$

Definition 1.3.3 *[8] Let $(\Omega, \mathcal{A})$ be measurable space. The random variable is a measurable function from Ω to $\mathbb{R}$, noted X.*

Definition 1.3.4 *[8] A probability on $(\Omega, \mathcal{A})$ is an application $\mathbb{P}$ from $\mathcal{A}$ in [0;1] such as:*

- $\mathbb{P}(\varnothing) = 0$ *and* $\mathbb{P}(\Omega) = 1$.
- $\mathbb{P}(\cup_{i \in N} B_i) = \sum_{i \in N} \mathbb{P}(B_i)$, *for any sequence of disjoints sets* $B_n \in \mathcal{A}$.

The triplet $(\Omega, \mathcal{A}, \mathbb{P})$ is called probability space.

1.3.2 Probability Law

Definition 1.3.5 *[8] Let X be a random variable definied on $(\Omega, \mathcal{A}, \mathbb{P})$, the law of X is the probability $\mathbb{P}_X$ on $(\mathbb{R}, \mathcal{B}_R)$ defined by $\mathbb{P}_X(B) = \mathbb{P}(X \in B)$, for any event $B \in \mathcal{B}_R$.*

Remark 1.3.1 *If X is a discrete variable then:*

$$\mathbb{P}(X \in B) = \sum_{x \in B} \mathbb{P}(X = x).$$

Definition 1.3.6 *[6] We defined the function f call density as the solution of the equations:*

- $\int_{\mathbb{R}} f(x)dx = 1.$
- $f(x) \geq 0, \quad \forall x \in \mathbb{R}.$

Remark 1.3.2 *If X is a continues variable then:*

$$\mathbb{P}(X \in B) = \int\limits_B f(x)dx.$$

Definition 1.3.7 *[6] A monotonic function of the real random variable X is called cumulative distribution function, noted $F_X(x)$ verifie the conditions:*

- *Continuous on the right limit on the left.*
- $\lim\limits_{x \longrightarrow -\infty} F_X(x) = 0$ *and* $\lim\limits_{x \longrightarrow +\infty} F_X(x) = 1.$

Remark 1.3.3 *The CDF for the real random variable X is:*

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} \sum\limits_{x_i \leq x} \mathbb{P}(X = x_i) & , \quad \textit{in the discrete case,} \\ \int_{-\infty}^{x} f(t)dt & , \quad \textit{in the continues case.} \end{cases}$$

Definition 1.3.8 *[6] We call esperance the linear functin of the real random variable X, noted $\mathbb{E}(X)$ such as:*

$$\mathbb{E}(X) = \begin{cases} \sum\limits_{i=1}^{n} x_i \mathbb{P}(X = x_i) & , \quad \textit{in the discrete case,} \\ \int\limits_{\mathbb{R}} x f(x)dx & , \quad \textit{in the continues case.} \end{cases}$$

Definition 1.3.9 *[6] We call variance the quadratic funtion of the real random variable X, noted $Var(X)$ such as:*

$$\begin{aligned} Var(X) &= \mathbb{E}[(X - \mathbb{E}(X))^2], \\ &= \mathbb{E}(X^2) - \mathbb{E}^2(X), \quad if\ X \in L^2(\Omega). \end{aligned}$$

Where

$$\mathbb{E}(X^2) = \begin{cases} \sum_{i=1}^{n} x_i^2 \mathbb{P}(X = x_i) & , \quad in\ the\ discrete\ case, \\ \int_{\mathbb{R}} x^2 f(x) dx & , \quad in\ the\ continues\ case. \end{cases}$$

Remark 1.3.4 *X is square integrable ($X \in L^2(\Omega)$) if $|X|^2$ has a finite esperance.*

Definition 1.3.10 *[6] Let $(X_n)_{n\in N}$ and X be random variables in $\mathbb{R}$, the sequence $(X_n)_{n\in N}$ converges to X in probability if:*

$$\forall \epsilon > 0 \quad \lim_{n\longrightarrow\infty} P(|X_n - X| < \epsilon) = 1. \tag{1.9}$$

Noted $X_n \longrightarrow^P X$.

1.3.3 Basic notion

Definition 1.3.11 *[2] The population is a set of objects on which a study is carried out. These objects are called individuals. And any property studied in individuals called character or variable noted X.*

Definition 1.3.12 *[2] A sample is a part from the population that has a small size that achieves a chracteristic.*

Definition 1.3.13 *[6] Consider the sample $X_1, X_2, ..., X_n$, we say that is independent if:*

$$\begin{aligned}\mathbb{P}(X_1 \leqslant x_1, X_2 \leqslant x_2, ..., X_n \leqslant x_n) &= \mathbb{P}(X_1 \leqslant x_1)\mathbb{P}(X_2 \leqslant x_2)\cdots\mathbb{P}(X_n \leqslant x_n), \\ &= [\mathbb{P}(X_1 \leqslant x_1)]^n, \quad if\ they\ are\ identical.\end{aligned}$$

Definition 1.3.14 *[2] We call estimator of θ any statistic (variable defined as being a function of the sample $X_1, \ldots, X_n$) noted $\hat{\theta}$, whose values are plausibly close to θ.*

Definition 1.3.15 *[6] We call $\hat{\theta}$ is an unbiased estimator if and only if:*

$$\mathbb{E}(\hat{\theta}) = \theta \Leftrightarrow \mathbb{E}(\hat{\theta} - \theta) = 0. \tag{1.10}$$

If $\mathbb{E}(\hat{\theta}) \neq \theta$, then $\hat{\theta}$ call bias estimator.

1.4 Non parametric Estimation

1.4.1 Cumulative Distribution Function

Definition 1.4.1 *[6] Suppose the sample $X_1, X_2, \ldots, X_n$ iid of F (unknown).*
A good estimator for F is the empirical distribution function, noted F_n defined by:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{]-\infty;x]}(x_i),$$

where

$$I_{]-\infty;x]}(x_i) = \begin{cases} 1 & if\ x_i \leq x, \\ 0 & otherwise. \end{cases}$$

1.4.2 Elementary Properties

- **Bias of EDF**

$$\mathbb{E}(F_n(x)) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(I_{]-\infty;x]}(x_i)) = \mathbb{P}(X \leq x) = F(x).$$

Therefore, for any point x, $F_n(x)$ is an unbiased estimator of $F(x)$ [4].

- **Variance of EDF**

For all x the variance of the estimator $F_n(x)$ is given by:

$$Var(F_n(x)) = \frac{1}{n}F(x)(1 - F(x)).$$

Noted $\sigma^2 = Var(F_n(x))$ [4].

Proof. We use the definition of variance with the iid of the sample.

$$\begin{aligned}
\sigma^2 &= var(F_n(x)) = \frac{1}{n^2}\sum_{i=1}^{n} var(I_{]-\infty;x]}(x_i)) = \frac{1}{n}var(I_{]-\infty;x]}(x_i)),\\
&= \frac{1}{n}\mathbb{E}[(I_{]-\infty;x]}(x_i))^2] - \frac{1}{n}\mathbb{E}^2(I_{]-\infty;x]}(x_i)),\\
&= \frac{1}{n}\mathbb{E}(I_{]-\infty;x]}(x_i)) - \frac{1}{n}\mathbb{E}^2(I_{]-\infty;x]}(x_i)),\\
&= \frac{1}{n}F(x) - \frac{1}{n}F^2(x) = \frac{1}{n}F(x)(1 - F(x)).
\end{aligned}$$

■

- **Convergence of EDF**

For all x, $F_n(x)$ converge to F in probability, we take $\epsilon = \sigma$, then [9]:

$$|F_n(x) - F(x)| < \sigma \quad for\ n\ very\ large\ \ p.s$$

Chapter 2

Support Vector Machines

The Support Vector Machines (SVM) is a supervised learning technique. The idea of SVM is to find a hyperplane in high-dimensional space (number of features) has a maximum margin to increase the likelihood of separating data with confidence. One of its main advantages is that it uses part of the training data (support vectors) to search for hyperplane and after the separation, the system can easily predict new data labels. In this chapter we will look at the case of the binary classification and how to find the term of the classifier in the linear separable data and generalize it to non linear seprable data in hard and soft margin. The kernel trick that can be used to the separation without transform the data. In the last part from the chapter we take SVR for SVM example and we only care about the error outside the tube not like the linear regression before (calculate the error for all the data).

2.1 Hard Margin

Suppose the training data set, $D = \{(x_1, y_1), ..., (x_n, y_n),\ x_i \in \mathbb{R}^m, y_i \in \{-1, +1\}\}$, where x_i is the i^{th} feature and y_i called the class (label).

2.1.1 Linear Classification

Definition 2.1.1 *[7] Hyperplane is a binary classifier is represented by the equation:*

$$w^T x + b = 0, \quad w \in \mathbb{R}^m, \quad b \in \mathbb{R}, \tag{2.1}$$

where w is the weight and b is the bias.

Definition 2.1.2 *[7] The decision rule for the classification is given by the sign of (2.1), if the sign positive then x in class 1 otherwise x in the class -1.*

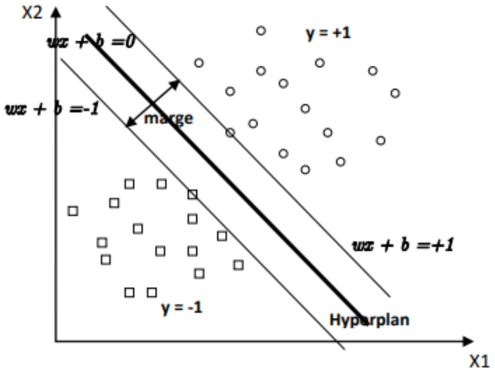

Figure 2.1: Hard Margin in 2-D space.

Remark 2.1.1 *Because the data are linear separable (figure 2.1 [3]).*

Equivalently:

$$y_i = \begin{cases} +1 & if \quad w^T x_i + b \geq 1, \\ -1 & if \quad w^T x_i + b \leq -1. \end{cases} \tag{2.2}$$

The combination of (2.2) is:

$$y_i(w^T x_i + b) \geq 1, \ i = 1, ..., n. \tag{2.3}$$

Definition 2.1.3 *[1] **(Margin)** given that w is perpendicular to the hyperplane, the distance between the hyperplane and any point x_i in terms the size of w is:*

$$\frac{\left|w^T x_i + b\right|}{||w||},$$

if x_s is the closest point to the hyperplane, we obtain: $\dfrac{\left|w^T x_s + b\right|}{||w||} = \dfrac{1}{||w||}$, then the margin is equal to $\dfrac{2}{||w||}$.

Maximizing Margin:

In order to maximize the margin, for mathematical convenience, $\frac{1}{2}w^T w$ is minimized subject to the constraint (2.3) ([7],[14]), then:

1. The COP:

$$\begin{cases} \quad \min\limits_{w,b} \dfrac{1}{2}w^T w \\ \textit{subject to} \\ \quad y_i(w^T x_i + b) \geq 1 \ , \ i = 1, \dots n. \end{cases} \tag{2.4}$$

2. We use the lagrange multipliers for transforming (2.4) to the lagrangian function:

$$L(w, b, \alpha) = \frac{1}{2}w^T w - \sum_{i=1}^{n} \alpha_i [y_i(w^T x_i + b) - 1], \qquad \alpha_i \geq 0. \tag{2.5}$$

3. The lagrange L should be minimized w and b and maximized the vector α. We apply the Fermat's theorem, this function is solved by calculating the partial derivatives ([7],[13]):

$$\nabla_w L = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0. \tag{2.6}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0. \tag{2.7}$$

4. Set off the value of (2.6) and (2.7) into equation (2.5), we get:

$$\begin{cases} \quad \max_{\alpha \geq 0} L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ subject\ to \\ \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \quad \alpha_i \geq 0 \quad \forall i = 1, \ldots, n. \end{cases} \tag{2.8}$$

5. The KKT conditions are satisfied. Consequently,

$$\alpha_i [y_i(w^T x_i + b) - 1] = 0, \qquad \forall i = 1, ...n. \tag{2.9}$$

From the condition (2.9), most of $\alpha_i = 0$ and the others verifie $y_i(w^T x_i + b) = 1$ called the **support vectors** [10].

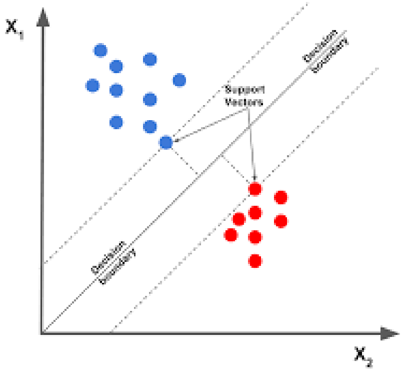

Figure 2.2: Support Vectors in 2-D space.

On definied the set:

$$SV = \{x_i / \ \alpha_i > 0\}, \quad \forall i = 1, ...n. \tag{2.10}$$

6. On obtain ([14],[3]):

$$\begin{aligned} w &= \sum_{x_i \ \in \ SV} \alpha_i y_i x_i. \qquad (2.11) \\ b &= y_i - w^T x_i. \qquad (2.12) \end{aligned}$$

7. This results given after we move to QP ([7],[10]):

$$\begin{cases} \min\limits_{\alpha \geq 0} \ \dfrac{1}{2}\alpha^T H\alpha - 1^T\alpha \\ subject.to \\ Y^T\alpha = 0 \\ \alpha \geq 0 \end{cases}$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)^T$, 1 is an (n,1) unit vector and H denotes a matrix SPSD.

Solution 2.1.1 *([7],[14],[1]) Using the equation of hyperplane* (2.1) *we obtain the maximum margin hyperplane:*

$$f(x) = \sum_{x_i \ in \ SV} \alpha_i y_i x^T x_i + b. \tag{2.13}$$

Note that the sign of this function in order to classify new data.

Exemple 2.1.1 *In (1-D) space the hyperplane has became a point and the number of support vectors is two.*

2.1.2 Non linear Classification(kernel method)

In this case the dataset is non linear separable in X (input space). Cosider the mapping function [10]: $\varPhi : X \longrightarrow Z$, this function can transform the dataset to Z space (high dimensional feature) which the data are linearly separable.

Remark 2.1.2 *We need from Z space [7]:*

1. $L(\alpha) = \sum\limits_{i=1}^{n} \alpha_i - \dfrac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} \alpha_i\alpha_j y_i y_j \varPhi(x_i)^T\varPhi(x_j)$.
2. $\sum\limits_{i=1}^{n} \alpha_i y_i = 0, \quad \alpha_i \geq 0 \quad \forall i = 1, \ldots n.$

And the decision function is given by:

1. $g(x) = sign(w^T\varPhi(x) + b)$.

2. $w = \sum_{\phi(x_i) are SV_s} \alpha_i y_i \phi(x_i), \quad b = y_i - w^T \Phi(x_i).$

Generalized Inner Product:

Definition 2.1.4 *Given two point x_1 and x_2 from X space [10].*

Let:

$$\Phi(x_1)^T \Phi(x_2) = k(x_1, x_2) \qquad (the \ \ kernel). \tag{2.14}$$

Exemple 2.1.2 *[7] let, $x_1 = (x_{11}, x_{12})^T$,we choose, $\quad \Phi(x_1) = (x_{11}^2, \sqrt{2}x_{11}x_{12}, x_{12}^2)^T$,*

then:

$$\begin{aligned} k(x_1, x_2) &= (x_{11}^2, \sqrt{2}x_{11}x_{12}, x_{12}^2)(x_{21}^2, \sqrt{2}x_{21}x_{22}, x_{22}^2)^T, \\ &= (x_{11}^2 x_{21}^2 + 2x_{11}x_{12}x_{21}x_{22} + x_{12}^2 x_{22}^2). \end{aligned}$$

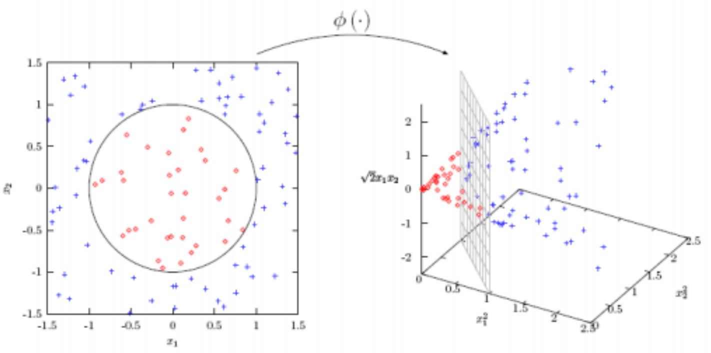

Figure 2.3: The hyperplane in X and Z space.

<u>**The trick:**</u> We can compute $k(x_1, x_2)$ without transforming x_1 and x_2, consider:

$$\begin{aligned} k(x_1, x_2) &= (x_1^T x_2)^2, \\ &= (x_{11}x_{21} + x_{12}x_{22})^2, \\ &= \Phi(x_1)^T \Phi(x_2). \end{aligned}$$

Remark 2.1.3 *Note that in order to calculate $\Phi(x_1)^T \Phi(x_2)$ in Z space, we calculate this* *product directly in X space by the kernel trick. The type of example calls the polynomial kernel.*

Generality [14]

If $X \equiv \mathbb{R}^m$ and the scalar product are a polynomial of order p. Equivalent:

$$\begin{aligned} k(x_1, x_2) &= (c + x_1^T x_2)^p, \\ &= (c + x_{11}x_{21} + x_{12}x_{22} + \ldots + x_{1m}x_{2m})^p. \end{aligned}$$

This calculation can be very difficult if the number of features (Z space) is very large.

Gaussian (RBF) kernel [10]

We use the kernel function RBF (Radial Basis Function) given by:

$$k(x_1, x_2) = exp(-\frac{||x_1 - x_2||^2}{2\sigma^2}),$$

which σ represents the bandwidth went the dimensionality of Z space are infinite.

The two types of kernel (polynomial and Gaussian RBF) are the most popular function.

linear separable [7]

$$k(x_1, x_2) = x_1^T x_2.$$

Proposition 2.1.1 *[10]**(Mercer's condition)** The matrix H in quadratic programming is a SPSD, this condition (kernel validation) can confirm a Z space exists.*

Solution 2.1.2 *[7]**(The final hypothesis)***

$$\begin{aligned} g(x) &= sign(\sum_{\alpha_i > 0} \alpha_i y_i k(x_i, x) + b). && (2.15) \\ b &= y_j - \sum_{\alpha_i > 0} \alpha_i y_i k(x_i, x_j), \quad for\ any\ \alpha_j > 0. && (2.16) \end{aligned}$$

2.2 Soft Margin

In the general, we don't have the optimal hyperplane, that can be classified all the data. Sometime we must leave some data on the wrong side of a decision boundary, this misclassification represented by a non-negative slack variable $\xi \in \mathbb{R}^n$ given by [10]:

$$\xi_i = \max(0, 1 - y_i(w^T x_i + b)), \quad for\ all\ couple\ (x_i, y_i).$$

We introdute constraint ([13][7]):

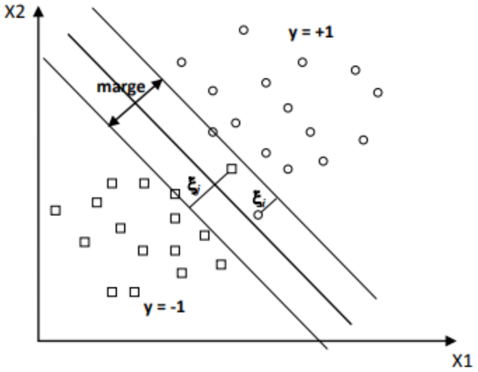

Figure 2.4: Margin violation.

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad i = 1, ...n. \tag{2.17}$$

And we try to maximize the margin and minimizing the sum of the errors [3], we obtaint the COP:

$$\begin{cases} \quad \min\limits_{w,b,\xi} \dfrac{1}{2} w^T w + C \sum\limits_{i=1}^{n} \xi_i \\ subject.to \\ \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad i = 1, \dots n. \\ \quad \xi_i \geq 0, \quad i = 1, \dots n. \end{cases} \tag{2.18}$$

Where $C > 0$ is a penalty parameter represent the balance between the two terms of the objective function, this problem has a meaning for some finite values.

Lagrange formulation:

$$L(w,b,\xi,\alpha,\beta)=\frac{1}{2}w^Tw+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i[y_i(w^Tx_i+b)-1+\xi_i]-\sum_{i=1}^{n}\beta_i\xi_i. \tag{2.19}$$

The same steps:

$$\nabla_w L = w-\sum_{i=1}^{n}\alpha_i y_i x_i=0. \tag{2.20}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n}\alpha_i y_i=0. \tag{2.21}$$

$$\frac{\partial L}{\partial \xi_i} = C-\alpha_i-\beta_i=0. \tag{2.22}$$

We get [14]:

$$\begin{cases} \qquad \max\limits_{\alpha\geq 0}\ \sum\limits_{i=1}^{n}\alpha_i-\frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j \\ subject.to \\ \qquad \sum\limits_{i=1}^{n}\alpha_i y_i=0 \quad ,i=1,...n. \\ \qquad 0\leq\alpha_i\leq C \quad ,i=1,\dots n. \end{cases} \tag{2.23}$$

The KKT comlementarity condition below:

$$\alpha_i[y_i(w^Tx_i+b)-1+\xi_i]=0, \quad i=1,...n. \tag{2.24}$$

Solution 2.2.1 *We get the same solution in the linear separable case (hard margin) for (w,b) and for the decision rule is given by:*

1. x_i is correctly classified if $\alpha_i=0$, $\xi_i=0$ or $\alpha_i=C$, $\xi_i<1$.

2. x_i is misclassified if $\alpha_i=C$ and $\xi_i>1$.

Types of support vectors [7]

1.x_i call margin support vectors: if $0<\alpha_i<C$ and $\xi_i=0$.

2.x_i call non-margin support vectors: if $\alpha_i=C$ and $\xi_i>0$.

Remark 2.2.1 *We used the soft margin in slightly non separable type, in the seriously non separable type we use the kernel.*
In the case of multiple classifications, we look for the hyperplanes of each two classes.

2.3 Support Vector Regression

SVR is a part from SVM for classification, it used for approximate a function (hyperplane) that has at most ϵ deviation from the actually obtained the labels for all the data ($y \simeq f(x)$). This new parameter call the ϵ-insensitivity loss function measure the error of the approximation [7]:

$$\xi = \begin{cases} 0 & , \; if \; |y - f(x)| \leq \epsilon, \\ |y - f(x)| - \epsilon & , \; otherwise. \end{cases} \tag{2.25}$$

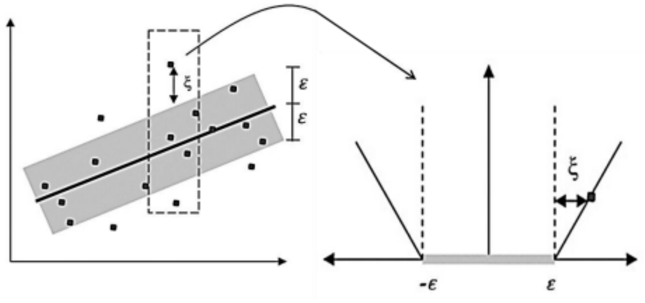

Figure 2.5: The deviation ϵ in the linear regression hyperplane.

Suppose the training data set [5]: $D = \{(x_1, y_1), ..., (x_n, y_n), x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$, for all point outside the ϵ-tube (margin) we introdute the slack variables:

$$\xi_i^+ = y_i - f(x_i) - \epsilon, \quad for\ the\ points\ above\ an\ \epsilon - tube, \tag{2.26}$$

$$\xi_i^- = f(x_i) - y_i - \epsilon, \quad for\ the\ points\ below\ an\ \epsilon - tube. \tag{2.27}$$

The non linear regression hyperplane represented by:

$$f(x) = w^T \phi(x) + b \ , \qquad w \in R^m, b \in \mathbb{R}.$$

And we use the combination constraintes (2.25) with (2.26) and (2.27), the COP has the soft margin form [10]:

$$\begin{cases} \qquad \min\limits_{w,b,\xi} \ \dfrac{1}{2} w^T w + C \sum\limits_{i=1}^{n} (\xi_i^+ + \xi_i^-) \\ subject.to \\ \qquad y_i - w^T \phi(x_i) - b \leq \epsilon + \xi_i^+, \qquad i = 1, \dots n \\ \qquad w^T \phi(x_i) + b - y_i \leq \epsilon + \xi_i^-, \qquad i = 1, \dots n \\ \qquad \xi^+, \xi^- \geq 0 \end{cases}$$

where C represents the balance between the flatness of the hyperplane and the losses. We use the same principles as the SVC to solve the problem:

lagrangian function [7]

$$\begin{aligned} L(w, b, \xi^+, \xi^-, \alpha^+, \alpha^-, \beta^+, \beta^-) = & \frac{1}{2} w^T w + C \sum_{i=1}^{n} (\xi_i^+ + \xi_i^-) \\ & - \sum_{i=1}^{n} \alpha_i^+ (\epsilon + \xi_i^+ + w^T \phi(x_i) + b - y_i) \\ & - \sum_{i=1}^{n} \alpha_i^- (\epsilon + \xi_i^- + y_i - w^T \phi(x_i) - b) \\ & - \sum_{i=1}^{n} \beta_i^+ \xi_i^+ - \sum_{i=1}^{n} \beta_i^- \xi_i^- . \end{aligned}$$

The same steps:

$$
\begin{aligned}
\nabla_w L &= w - \sum_{i=1}^{n}(\alpha_i^+ - \alpha_i^-)\phi(x_i) = 0. \\
\frac{\partial L}{\partial b} &= \sum_{i=1}^{n}(\alpha_i^+ - \alpha_i^-) = 0. \\
\frac{\partial L}{\partial \xi_i^+} &= C - \alpha_i^+ - \beta_i^+ = 0. \\
\frac{\partial L}{\partial \xi_i^-} &= C - \alpha_i^- - \beta_i^- = 0.
\end{aligned}
$$

We get:

$$
\begin{cases}
\max\limits_{\alpha^+,\alpha^- \geq 0} \quad \sum\limits_{i=1}^{n} y_i(\alpha_i^+ - \alpha_i^-) - \epsilon \sum\limits_{i=1}^{n}(\alpha_i^+ + \alpha_i^-) - \frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}(\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-)k(x_i, x_j) \\
subject.to \\
\sum\limits_{i=1}^{n}\alpha_i^+ = \sum\limits_{i=1}^{n}\alpha_i^- \\
0 \leq \alpha_i^+, \alpha_i^- \leq C \quad , i = 1, \ldots n
\end{cases}
$$

The KKT comlementarity condition below:

$$
\alpha_i^+(\epsilon + \xi_i^+ + w^T\phi(x_i) + b - y_i) = 0, \quad i = 1, ..., n.
$$

$$
\alpha_i^-(\epsilon + \xi_i^- + y_i - w^T\phi(x_i) - b) = 0, \quad i = 1, ..., n.
$$

And we get for all SV_s:

$$
\begin{aligned}
f(x) &= \sum_i (\alpha_i^+ - \alpha_i^-)k(x_i, x) + b. \\
w &= \sum_i (\alpha_i^+ - \alpha_i^-)\phi(x_i). \\
b &= y_j - \sum_i (\alpha_i^+ - \alpha_i^-)k(x_j, x_i) - \epsilon, \quad for\ 0 < \alpha_j^+ < C.
\end{aligned}
$$

In the linear regression hyperplane, $\phi(x_i) = x_i$ and $k(x_i, x_j) = x_i^T x_j$.

Chapter 3

Support Vector Density Estimation

SVM is a non parametric approach, developed to solve the density estimation problem too. The objective of SVM is to estimate the probability density function by support vector regression or in other words “estimate the cumulative distribution function by the equation of the hyperplane and the density function can be easily obtained ”. In this chapter, we see how to estimate the targets and the advantage of this estimation (converge in probability) to control the deviation. For give the expression of density, we definied an operator and some conditions to choise the kernel to make it a density of probability. To complete this work, we use non parametric estimation on cross kernel in order to confirm that is density of probability.

3.1 Basic idea to use SVM approach

Given the iid sample [4]:

$$D = \{x_1, \ldots, x_n / x_i \in \mathbb{R}^m\}, \tag{3.1}$$

but the SVM is a supervised method, we must have the targets. We definite the EDF where the vectors x and x_i of dimension m [11]:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{]-\infty;x]}(x_i), \tag{3.2}$$

$$I_{]-\infty;x]}(x_i) = \prod_{k=1}^{m} I_{]-\infty;x^k]}(x_i^k). \tag{3.3}$$

For all x_i, we estimate $y_i \simeq F_n(x_i)$, constructing the data :

$$D = \{(x_1, F_n(x_1)), \ldots, (x_n, F_n(x_n))\}, \tag{3.4}$$

we have from SVR, the definition of the ϵ-insitivity combinate with what we get, then [9]:

$$\left| F_n(x_i) - \hat{F}(x_i) \right| \leq \epsilon + \xi_i, \quad i = 1, ..., n,$$

from an ather side, we know F_n is unbiased and converge in probability to F (for n very large):

$$|F_n(x_i) - F(x_i)| < \sigma, \quad for\ all\ fixed\ x_i, \tag{3.5}$$

where σ^2 is the variance.

From (3.5) we can control the free parameter ϵ in the SVR by the standard deviation:

$$\epsilon_i = \hat{\sigma}_i = \sqrt{\frac{1}{n} F_n(x_i)(1 - F_n(x_i))}. \tag{3.6}$$

Using the data:

$$D = \{(x_1, F_n(x_1), \epsilon_1), \ldots, (x_n, F_n(x_n), \epsilon_n)\}. \tag{3.7}$$

3.1.1 Linear operator equations

Definition 3.1.1 *[15] We definied the linear operator A from a Hilbert space to anthor:*

$$Af(x) = F(x), \quad f(x) = w^T\psi(x).$$

Where f is a linear combination of functions and w is the coefficients of the hyperplane.

Then:

$$F(x) = Af(x) = w^T\phi(x), \tag{3.8}$$

and we have from SVR, $w = \sum_i(\alpha_i^+ - \alpha_i^-)\phi(x_i)$ $(\phi(x_i)$ *are* $SV_s)$.
Set off in (3.8):

$$\begin{aligned} \hat{F}(x) &= \sum_i(\alpha_i^+ - \alpha_i^-)\phi(x_i)^T\phi(x), \\ &= \sum_i(\alpha_i^+ - \alpha_i^-)k(x_i, x), \end{aligned} \tag{3.9}$$

and

$$\begin{aligned} \hat{f}(x) &= \sum_i(\alpha_i^+ - \alpha_i^-)\phi(x_i)\psi(x), \\ &= \sum_i(\alpha_i^+ - \alpha_i^-)\kappa(x_i, x). \end{aligned} \tag{3.10}$$

Where $\kappa(x_i, x)$ call the cross kernel.

3.1.2 Kernel Validation

1. To guarantee the positivity of PDF (CDF monotonic function), we choose a monotonic non-symmetrical kernel from L^1 with positive coefficients $(\alpha_i^- = 0, \quad \forall\ i = 1, \dots, n)$,

then:

$$\hat{F}(x) = \sum_i \alpha_i^+ k(x_i, x), \tag{3.11}$$

$$\hat{f}(x) = \sum_i \alpha_i^+ \kappa(x_i, x). \tag{3.12}$$

2. If $x \in [0;1]$ we have $F(0) = 1$ *and* $F(1) = 1$, we choose the kernel satisfys these conditions, then:

$$\sum_i \alpha_i^+ k(0, x_i) = 0, \tag{3.13}$$

$$\sum_i \alpha_i^+ k(x_i, 1) = 1. \tag{3.14}$$

For more explain see [4].

Conclusion

During the preparation of this modest work, we have tried in the first to present the support vector machines (SVM) with mathematical formulation, then we explained how to use it in different applications such as classification and regression problems. In the second part of this dessertation, we focused on the ability of (SVM) to approximate the probability density function. At the end of this work, we presented a technique for estimating the cumulative distribution function which is based on the empirical distribution function and the support vector machines (SVM).

Finally, we hope to have the ability to explore this vast field of artificial intelligence.

References

[1] **Ayat Nedjem-eddine**. Sélection de modèle automatique des machines à vecteurs de support. Doctorate's thesis, Le 20 January 2004.

[2] **Chesneau Christophe**. Sur L'Estimateur du Maximum de Vraisemblance (emv). Université de Caen. le 23 Octobre, 2018.

[3] **Djeffal Abdelhamid**. Utilisation des méthodes Support Vector Machine dans l'analyse des bases de donnes. Doctorate's thesis, 2011/2012.

[4] **J. Weston, A. Gammerman, M.Stitson, V. Vapnik, V. Vovk, C. Watkins**. Density Estimation using Support Vector Machines. Article, February 5, 1998.

[5] **Jie Liu**. Failure Prognostics by Support Vector Regression of Time Series Data under Stationary/Nonstationary Environmental and Operational Conditions. Doctorate's thesis, Feb.12, 2015.

[6] **Lejeune Michel**. Statistique La théorie et ses applications Deuxième édition. Springer-Verlag France, Paris, 2010.

[7] **Lipo Wang**. Support Vector Machines: Theory and Applications. Springer-Verlag, 2005.

[8] **Monique Jeanblanc**. Cours de calcul stochastique Master 2IF EVRY. Septembre, 2006.

[9] **Mukherjee Sayan**. Multivariate Density Estimation: An SVM Approach. Article, October 2004.

[10] **Nello Cristianini** and **John Shawe-Taylor**. An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000.

[11] **Refaat M Mohamed, Ayman El-Baz and Aly A. Farag, Senior Member IEEE**. Probability Density Estimation Using Advanced Support Vector Machines and the Expectation Maximization Algorithm. Article, 2007.

[12] **Vapnik Vladimir**. An Overview of Statistical Learning Theory. EEE transactions on neural networks, vol. 10, no. 5, september, 1999.

[13] **Vapnik Vladimir**. Statistical Learning Theory. John Wiley and Sons, Inc. New York, 1998.

[14] **Vapnik Vladimir**. The Nature of Statistical Learning Theory. Springer-Verlag New York, 1995.

[15] **ZHANG Zhao , ZHANG Su, ZHANG Chen-xi, CHEN Ya-zhu**. SVM for density estimation and application to medical image segmentation. Article, Feb. 27, 2006.

Résumé:

Notre objective dans ce mémoire est d'évoquer le domaine de l'apprentissage statistique, définir les machines à vecteurs de support ou séparateurs à vaste marge (SVM) comme un outil de base dans la théorie d'éstimation et de montrer leur capacité à estimer une densité de probabilité.

Les mots clés:
Apprentissage statistique, Séparateurs à vaste marge(SVM), Noyau, estimation d'une densité.

Abstract:

Our aim in this dissertation is to evoke the area of statistical learning, to define the support vector machines (SVM) as a basic tool in the estimation theory, and to show their ability to estimate a probability density function.

Key-words:
Statistical learning, Support vector machines (SVM), kernel, density estimation.

الملخص:

هدفنا في هذه المذكرة هو استحضار مجال التعلم الاحصائي لتعريف آلية المتجهات الداعمة كأداة أساسية في نظرية التقدير و اظهار قدرتها لتقدير الكثافة الاحتمالية.

الكلمات المفتاحية:
التعلم الاحصائي، آلية المتجهات الداعمة، النواة، تقدير الكثافة.