Introducción a los SVMs con R

Innova-TSN *Octubre 2016*

Agenda

1. **Innova-TSN:**
   i. **¿Quiénes somos?**
      - ✓ **Misión, Visión, Valores**
      - ✓ **Equipo**
   ii. **Ámbitos de Negocio**

2. Introducción a los SVMs
   1. Introducción
   2. Prolegómenos: Primeros Modelos de Clasificación
   3. Concepto de SVM

3. Índice ampliado para las VIII Jornadas de usuarios de R en Albacete

4. Información de contacto

INNOVA-TSN

- ¿QUIÉNES SOMOS?

Innova-TSN: ¿Quiénes Somos?

Empresa de capital español con 12 años de vida cuya actividad está enfocada a la consultoría de definición e implantación de **soluciones de Business Intelligence, CRM, Business Analytics, Big Data y Business Discovery aplicadas a la toma de decisiones.**

Innova-TSN cuenta con una amplia experiencia en diferentes **sectores de actividad**: **Banca y Finanzas, Seguros, Telecomunicaciones, Utilities, Sector Público, Retail, Servicios, etc**.

La compañía cubre los tres pilares básicos y necesarios para la implantación, gestión y soporte de este **tipo de soluciones**: **Tecnología/Infraestructura, Proyecto/Servicio, Consultoría.**

DANDO FORMA A LAS IDEAS

«QUE NUESTROS CLIENTES TOMEN MEJORES DECISIONES DE NEGOCIO BASADAS EN LA INFORMACIÓN QUE LE PROPORCIONAN NUESTRAS SOLUCIONES».

INNOVA-TSN

- EQUIPO

Innova-TSN: Equipo

Innova-TSN es una compañía que **apuesta por el talento.** Nuesta ventaja competitiva se asienta en un equipo especializado, motivado, comprometido y joven, que comparte una filosofía de confianza y respeto mutuo.

Nuestro equipo está en **constante evolución**. La composición **multidisciplinar** de nuestros profesionales asegura el cumplimiento conjunto de los objetivos de negocio, al interiorizar las necesidades de nuestros clientes.

Con un equipo de más de **180 profesionales**, **la pasión, la creatividad y la dedicación** de cada miembro por su trabajo nos convierte en una empresa de confianza, líder en el mercado.

INNOVA-TSN

- ÁMBITOS DE NEGOCIO

ÁMBITOS DE NEGOCIO

Innova-TSN ofrece un amplio abanico de soluciones integrales, personalizadas e innovadoras dentro de los ámbitos indicados a continuación

Business Intelligence

Business Analytics

Big Data

CRM

Business Discovery

Agenda

1. Innova-TSN:
   i. ¿Quiénes somos?
      ✓ Misión, Visión, Valores
      ✓ Equipo
   ii. Ámbitos de Negocio

2. **Introducción a los SVMs**
   1. **Introducción**
   2. Prolegómenos: Primeros Modelos de Clasificación
   3. Concepto de SVM

3. Índice ampliado para las VIII Jornadas de usuarios de R en Albacete

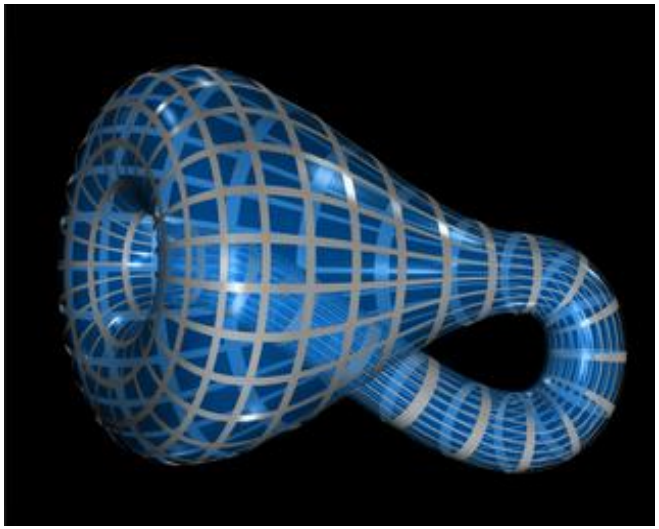
4. Información de contacto

Introducción

¿Qué sucede más allá de la tercera dimensión?

¿Qué significado cabe dar a una "cuarta dimensión?

La botella de Klein es uno de los primeros ejemplos de superficie matemática que "parcialmente" nos ofrece una idea de qué es lo que es en 3 dimensiones, pero para su completa comprensión, se necesita una cuarta dimensión



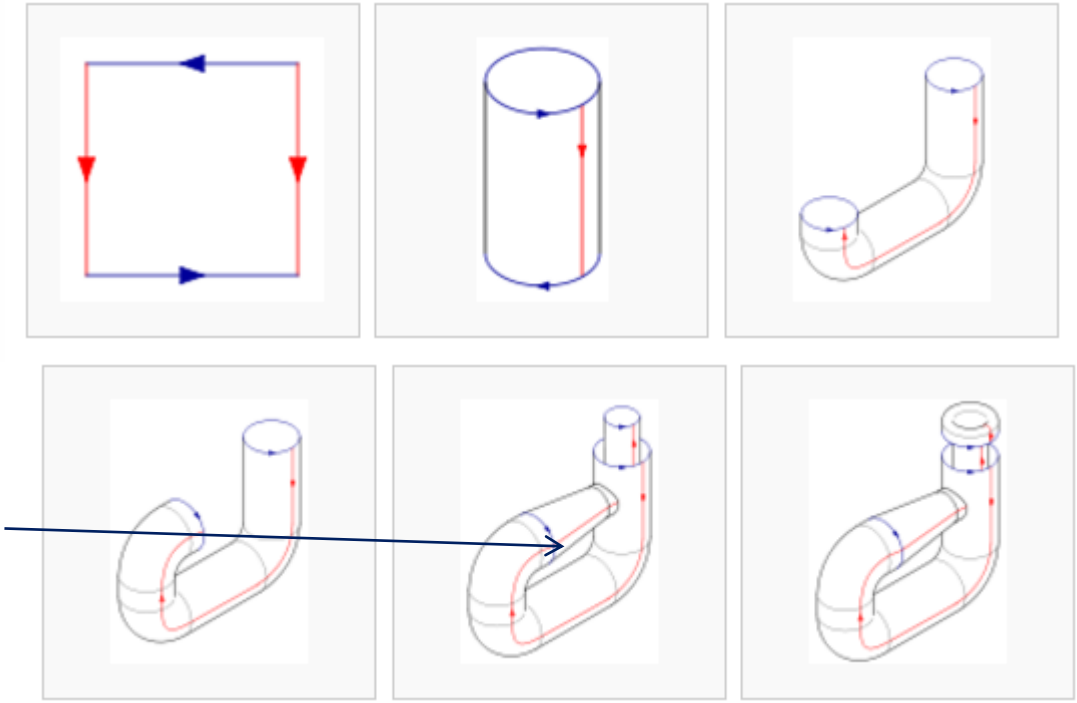
Fue descrita por primera vez en 1882 por el Alemán Felix Klein (1849 – 1925)

$$\left(x^2+y^2+z^2+2y-1\right)\left[\left(x^2+y^2+z^2-2y-1\right)^2-8z^2\right]$$
$$+16xz\left(x^2+y^2+z^2-2y-1\right)=0$$

Introducción

Actualmente hay empresas que venden botellas de klein de cristal, no hay que olvidar que la razón del nombre de "botella de klein" viene de una incorrecta traducción del alemán y que debería denominarse "superficie de klein"

Es una superficie abierta, no orientable, de característica de euler 0 y por tanto no se corta así misma



La auto-intersección no se produciría en una cuarta dimensión

Agenda

1. Innova-TSN:
   i. ¿Quiénes somos?
      ✓ Misión, Visión, Valores
      ✓ Equipo
   ii. Ámbitos de Negocio

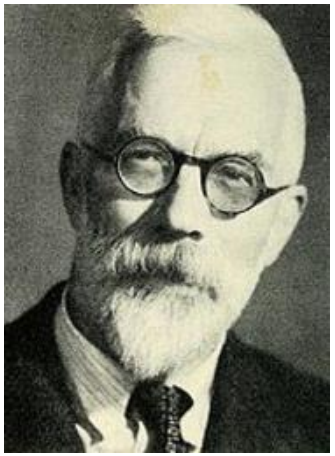
2. **Introducción a los SVMs**
   1. Introducción
   2. **Prolegómenos: Primeros Modelos de Clasificación**
   3. Concepto de SVM

3. Índice ampliado para las VIII Jornadas de usuarios de R en Albacete

4. Información de contacto

Prolegómenos: Primeros Modelos de Clasificación

El primer algoritmo de reconocimiento de patrones fue ideado por: Sir Ronal Aylmer Fisher (1860 – 1962)



Este algoritmo es el que se conoce en la actualidad como:

ANÁLISIS DISCRIMINANTE

Y fue creado durante los años 30 del siglo XX

Objetivo del Algoritmo:

¿Cómo separar dos poblaciones normales de vectores de n-dimensiones de modo óptimo?

$N(\vec{\mu}_1; \underline{\Sigma}_1)$ $N(\vec{\mu}_2; \underline{\Sigma}_2)$

Prolegómenos: Primeros Modelos de Clasificación

El problema se resuelve de modo general con la siguiente regla de asignación:

$$F_{sq}(\vec{x}) = sign\left[-\frac{1}{2}(\vec{x}-\vec{\mu}_1)^T \underline{\Sigma_1}^{-1}(\vec{x}-\vec{\mu}_1) + \frac{1}{2}(\vec{x}-\vec{\mu}_2)^T \underline{\Sigma_2}^{-1}(\vec{x}-\vec{\mu}_2) + \frac{1}{2} ln\left(\frac{|\underline{\Sigma_2}|}{|\underline{\Sigma_1}|}\right) - Ln\left(\frac{|\pi_2|}{|\pi_1|}\right)\right]$$

Es decir, dado un vector, se quiere saber si está en la clase 1 o en la clase 2. Para ello:

-Se requiere conocer las proporciones de las poblaciones de vectores a priori, junto con las medias y las varianzas de cada una de las poblaciones

-Se hace las operaciones “matriciales” anteriores y al final se obtiene un número

-En función de si ese número es positivo o negativo, se dirá que el vector a clasificar pertenecerá a una población o a otra

Prolegómenos: Primeros Modelos de Clasificación

Simplificación recomendada por Fisher:

$$F_{lin}(\vec{x}) = sign\left[(-\vec{\mu}_1 + \vec{\mu}_2)^T \underline{\Sigma}^{-1}\vec{x} - \frac{1}{2}\left(\vec{\mu}_1^T \underline{\Sigma}^{-1}\vec{\mu}_1 - \vec{\mu}_2^T \underline{\Sigma}^{-1}\vec{\mu}_2\right) - Ln\left(\frac{|\pi_2|}{|\pi_1|}\right)\right]$$

Hipótesis del supuesto: **Las Varianzas son Iguales**

Fisher recomendaba esta formulación para clasificación, incluso en casos donde el supuesto de normalidad no se sostuviese por el ahorro en cálculo que supone.

Cabe pensar que los primeros ordenadores no verían la luz hasta los años 40 y Fisher jamás pudo disfrutar de un ordenador de sobremesa como mi antiguo Spectrum 128K

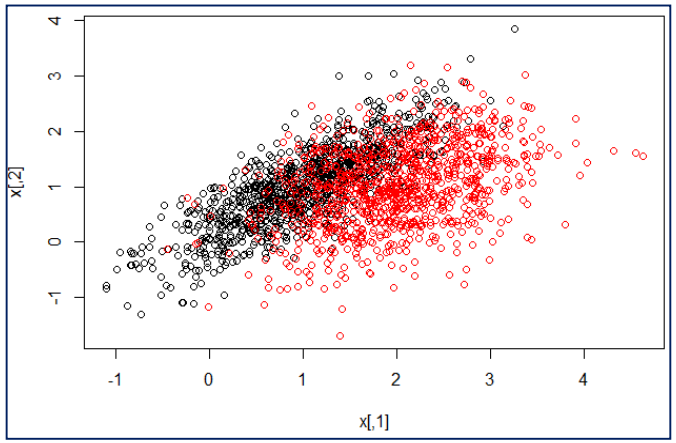
Prolegómenos: Primeros Modelos de Clasificación

Ejemplo I: Análisis sencillo de un LDA en R (Proyecto_01.R)

$$\vec{\mu}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \underline{\Sigma}_1 = \begin{pmatrix} 0.5 & 0.4 \\ 0.4 & 0.5 \end{pmatrix}$$

$$\vec{\mu}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \qquad \underline{\Sigma}_2 = \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.5 \end{pmatrix}$$

Para cada una de las poblaciones se van a simular 1000 vectores de dimensión n = 2



Los puntos de color negro se clasifican como +1
Los puntos de color rojo se clasifican como -1

Se toma como conjunto training una muestra aleatoria estratificada del 70% y se predice sobre el otro 30%

Prolegómenos: Primeros Modelos de Clasificación

Ejemplo I: Análisis sencillo de un LDA en R (Proyecto_01.R)

Resultados obtenidos:

(1) Preparación de datos: Construcción de conjuntos training y validación

(2) Implementación directa de la Fórmula de Fisher

(3) Nivel de clasificación a través de la Fórmula de Fisher

(4) Nivel de clasificación a través de la técnica QDA en R

(5) Nivel de clasificación a través de la técnica LDA en R

(6) Nivel de clasificación a través de un SVM básico en R

```
#2.- Fórmula Fischer modificada

Fischer <- function(x, mu1, mu2, sigma1, sigma2, pr1, pr2){

  Num <- sign(-0.5* (x - mu1) %*% solve(sigma1) %*% (x - mu1) +
              0.5* (x - mu2) %*% solve(sigma2) %*% (x - mu2) +
              0.5* log(det(sigma2)/det(sigma1), base = exp(1)) -
              log(pr2/pr1, base = exp(1)))

  return(Num)

}
```

```
                Predicted Class
Actual Class   -1   1
          -1  229  79
           1   39 253
```

Prolegómenos: Primeros Modelos de Clasificación

Ejemplo I: Análisis sencillo de un LDA en R (Proyecto_01.R)

QDA

```
                Predicted Class
Actual Class   -1    1
           -1 230   78
           1   41  251
```

LDA

```
                Predicted Class
Actual Class   -1    1
           -1 230   78
           1   42  250
```

SVM

```
                Predicted Class
Actual Class   -1    1
           -1 222   86
           1   31  261
```

¿Qué sucedería si en el QDA se pusiesen las mismas probabilidades a priori que se utilizaron en la fórmula de Fisher anterior?

Sin haber definido aún los SVM, se observa que con una sencilla línea de código, se tiene una clasificación superior a un LDA clásico

```
Modelo_svm <- svm(xtrain, ytrain, kernel = "linear")
```

Prolegómenos: Primeros Modelos de Clasificación

En el ejemplo anterior se conoce el sistema de generación de la información, esto en la realidad no se conocerá en general

Los elementos anteriores, como la media, la desviación típica y probabilidades a priori tendrán que ser estimados a priori

Hay que estimar un número de parámetros igual a: $\frac{v(v+3)}{2}g$

Agenda

1. Innova-TSN:
   i. ¿Quiénes somos?
      ✓ Misión, Visión, Valores
      ✓ Equipo
   ii. Ámbitos de Negocio

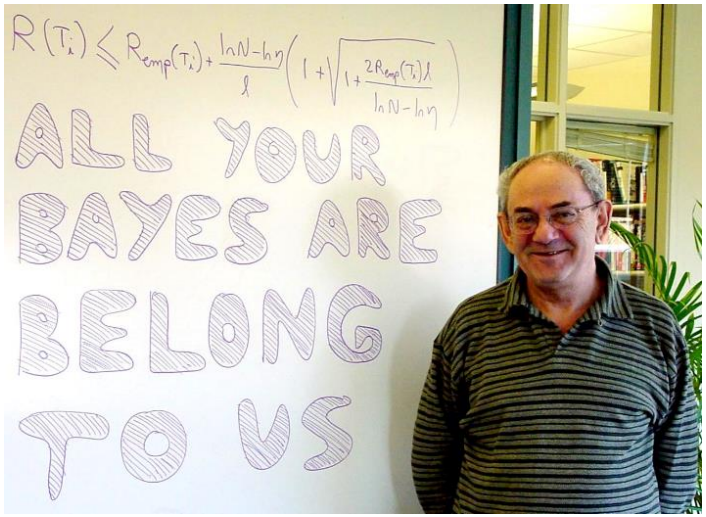
2. **Introducción a los SVMs**
   1. Introducción
   2. Prolegómenos: Primeros Modelos de Clasificación
   3. **Concepto de SVM**

3. Índice ampliado para las VIII Jornadas de usuarios de R en Albacete

4. Información de contacto

Introducción a los SVM. Concepto de SVM

Vladimir Vapnik (1936-?) de los laboratorios AT&T crea en la primera mitad de los 90 esta técnica



Originariamente la variable dependiente era discreta, de tipo binario o nominal

Objetivo del Algoritmo:

Reconocimiento de Texto a partir de patrones de escritura

Se trata de separar un conjunto de vectores en un espacio, en regiones disjuntas acorde a las características de dichos vectores

Introducción a los SVM. Concepto de SVM

Los Support Vector Machine están dentro de los conocidos como Clasificadores Lineales (grupo al que pertenecerían los anteriores algoritmos LDA y QDA)

La diferencia de los SVM con las anteriores técnicas es que geométricamente la separación no tiene porqué realizarse en el espacio de partida, sino esta técnica permite recurrir al uso de un espacio de dimensión superior donde se realizaría dicha separación

$$\Phi: X = \{\vec{x}_1; \dots; \vec{x}_n\} \subseteq \mathbb{R}^p \rightarrow \mathbb{R}^m$$

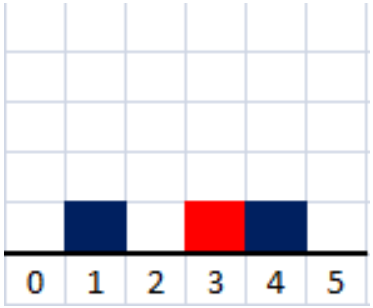
$$\vec{x}_i \in X: \rightarrow \Phi(\vec{x}_i) \in \mathbb{R}^m$$

¿Cómo será el tamaño de m respecto a p? ¿Qué se entiende por la expresión "curse of dimensionality" de los SVMs?

Introducción a los SVM. Concepto de SVM

Ejemplo II: Concepto de Embedimiento (Proyecto_03.R)

Se trata de separar, mediante un SVM "Heurístico" 3 puntos de una recta real, donde el 1 y el 4 están en un mismo grupo (azul) y el 3 estaría en otro




```
#Se definen 3 puntos en una recta a los que se asocia 2 colores mediante números
#azul con el número -1 y rojo con el número 1

p1 <- c(1,-1)
p2 <- c(3,1)
p3 <- c(4,-1)
```

Se propone un embedimiento de la recta en el plano real del siguiente modo:

$$\Phi: X = \{\vec{x}_1; \vec{x}_2; \vec{x}_3\} \subseteq \mathbb{R}^1 \rightarrow \mathbb{R}^2$$

$$\vec{x}_i = x_i \in X: \rightarrow \Phi(\vec{x}_i) = (x_i; x_i^3) \in \mathbb{R}^2$$

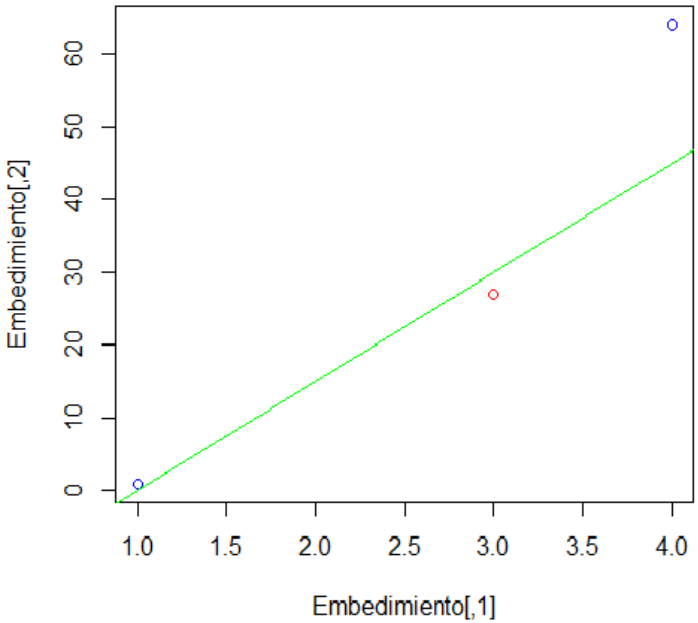
Introducción a los SVM. Concepto de SVM

Ejemplo II: Concepto de Embedimiento (Proyecto_03.R)

Embedimiento en el plano: se ha buscado "heurísticamente" una recta que en el plano separo a los puntos "embedidos"

$$Si \quad \begin{cases} y \geq -15 + 15x \; entonces \; Azul \\ y < -15 + 15x \; entonces \; Rojo \end{cases}$$

```
#Función de Embedimiento y = x^3

Emb1 = c(p1[1], p1[1]**3, -1)
Emb2 = c(p2[1], p2[1]**3, 1)
Emb3 = c(p3[1], p3[1]**3, -1)

#Representación del Embedimiento

Embedimiento <- rbind(c(Emb1[1], Emb1[2]),c(Emb2[1], Emb2[2]),
                      c(Emb3[1], Emb3[2]))
plot(Embedimiento, col = c("blue", "red", "blue"))

#Se busca una recta (hiperplano) que separe los azules del rojo, ahora que
#se puede

abline(a = -15, b = 15, col = "green")
```



Introducción a los SVM. Concepto de SVM

Ejemplo II: Concepto de Embedimiento (Proyecto_03.R)

Algoritmo de separación en la recta real (espacio original):

-La siguiente inecuación de grado 3, tiene 3 ceros que define 3 regiones disjuntas en la recta real

$$\begin{cases} x^3 \geq -15 + 15x \; entonces \; Azul \\ x^3 < -15 + 15x \; entonces \; Rojo \end{cases}$$

```
z <- matrix(c(15, -15, 0, 1), ncol = 1)

polyroot(z)
```

Obteniéndose como soluciones las siguientes:

$$x_1 = -4.299827; \; x_2 = 1.0852 \; ; \; x_3 = 3.214627$$

-Por tanto el algoritmo SVM de separación es:

$$SVM\ Heurístico = \begin{cases} (-\infty;\ -4.299827) & \rightarrow Rojo \\ [-4.299827;\ 1.0852) & \rightarrow Azul \\ [1.0852;\ 3.214627) & \rightarrow Rojo \\ [3.214627;\ +\infty) & \rightarrow Azul \end{cases}$$

Introducción a los SVM. Concepto de SVM

Ejemplo III: Aplicación práctica de SVM

Este ejemplo se encuentra en Introduction to Statistical Learning

Se trata de ver una aplicación con datos reales donde tiene sentido la aplicación de un SVM

En el conjunto test, el acierto es en torno al 90%

Preguntas:

¿Es el modelo estable ante cambios en los datos?
¿Comparación con otras metodologías? ¿Vs Decission Trees, RRNN, …?

Introducción a los SVM. Concepto de SVM

Características básicos de los SVM:

-La variable dependiente es de tipo nominal finito (ideal si esta variable es de tipo binario)

-Se hace uso de supuestos de optimización convexa poco restrictivos, que obedecen a la formulación matemática de cómo se presenta el algoritmo, no a supuestos sobre los datos

-Gracias el Teorema de Mercer, el uso de las funciones de embedimiento no será necesario y se usará en su lugar, funciones kernel, mucho más manejables

-Como se verá a continuación, el algoritmo ofrece como resultado separaciones convexas óptimas (que se entienden muy bien en el caso de variables dependientes binarias y en consecuencia, mediante la generación de hiperplanos de separación óptima)

-Los SVM requieren que las variables explicativas sean numéricas, no tratándose el caso nominal

-Mal tratamiento de los "missing" o se estiman valores o se elimina el registro de la muestra

Agenda

1. Innova-TSN:
   i. ¿Quiénes somos?
      - ✓ Misión, Visión, Valores
      - ✓ Equipo
   ii. Ámbitos de Negocio

2. Introducción a los SVMs
   1. Introducción
   2. Prolegómenos: Primeros Modelos de Clasificación
   3. Concepto de SVM

3. **Índice ampliado para las VIII Jornadas de usuarios de R en Albacete**

4. Información de contacto

Índice Ampliado para las VIII Jornadas de Usuarios de R en Albacete

- INTRODUCCIÓN
- PROLEGÓMENOS: PRIMEROS MODELOS DE CLASIFICACIÓN
- CONCEPTO DE SVM
- PRINCIPALES TIPOS DE SVM DISCRETOS: Bases Teóricas y Aplicaciones sencillas

  - ✓ MAXIMAL MARGIN CLASSIFIER
  - ✓ SUPPORT VECTOR CLASSIFIERS
  - ✓ SUPPORT VECTOR MACHINES

- EJEMPLOS DE APLICACIONES DE SVM

  - ✓ RECONOCIMIENTO DE ESCRITURA
  - ✓ DETECCIÓN DE CORREO BASURA (SPAM)

- CONCLUSIONES Y GENERALIZACIONES DE LOS SVM

Agenda

1. Innova-TSN:
   i. ¿Quiénes somos?
      - ✓ Misión, Visión, Valores
      - ✓ Equipo
   ii. Ámbitos de Negocio

2. Introducción a los SVMs
   1. Introducción
   2. Prolegómenos: Primeros Modelos de Clasificación
   3. Concepto de SVM

3. Índice ampliado para las VIII Jornadas de usuarios de R en Albacete

4. **Información de contacto**

Información de contacto

Información de Contacto

Innova-TSN se diferencia por la atención directa y permanente a sus clientes y por el equipo de profesionales especializados en todas las ramas del negocio, garantizando el cumplimiento de objetivos y cierre exitoso de proyectos y servicios.

Para cualquier consulta, pueden dirigirse a:

Francisco J. Rodríguez Aragón
Analytic Consultant

Ph. D. in Statistics
Associate Professional Risk Manager Certified
SAS Advanced Programmer Certified

Teléfono:
91 513 00 73

E-mail:
francisco.rodriguez@innova-tsn.com

innova-tsn
GRACIAS