

¿Pueden los androides aprender qué es una oveja mecánica?

Un paseo por los modelos generativos

Alfredo Cuesta Infante

E.T.S. Ing. Informática, URJC (Móstoles)

Master UCM en Ingeniería Informática

20 de Junio de 2016

Contenidos

- 1 Motivación
- 2 Aprendizaje automático
 - Clasificación según tareas
 - Aproximación probabilística
- 3 Modelos generativos
 - Ideas clave
 - Modelo de Mezclas de Gaussianas
 - Modelado mediante funciones cópula
 - Modelos gráficos probabilísticos
 - Otros
- 4 Conclusiones

¿Cómo definimos...

..aprender?

- Adquirir el conocimiento de algo por medio del estudio o de la experiencia.
- Fijar algo en la memoria

..inteligencia?

- Capacidad de entender
- Capacidad de resolver problemas

..entender?

- Discurrir, inferir, deducir

Dos ejemplos

A. Conan Doyle: “Estudio en escarlata”

(1887)

Quien se guiase por la lógica podría inferir de una gota de agua la posibilidad de la existencia de un océano Atlántico o de un Niágara, sin necesidad de haberlos visto u oído hablar de ellos.

J.L. Borges : “Funes el memorioso”

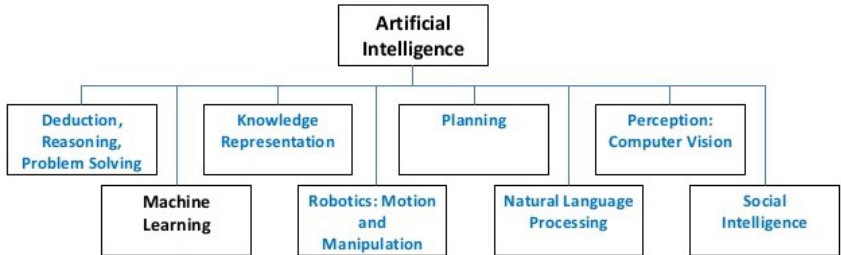
(1944)

Funes no sólo recordaba cada hoja de cada árbol de cada monte, sino cada una de las veces que la había percibido o imaginado... No sólo le costaba comprender que el símbolo genérico perro abarcara tantos individuos dispares de diversos tamaños y diversa forma; le molestaba que el perro de las tres y catorce (visto de perfil) tuviera el mismo nombre que el perro de las tres y cuarto (visto de frente)... Había aprendido sin esfuerzo el inglés, el francés, el portugués, el latín. Sospecho, sin embargo, que no era muy capaz de pensar. Pensar es olvidar diferencias, es generalizar, abstraer.

¿Qué tipo de inteligencia puede tener a una máquina?

La regla de oro

**Ante un caso nuevo, nunca visto antes,
la máquina debe ser capaz de responder correctamente.**



Memorizar cantidades ingentes de datos NO es I.A.

¿En qué consiste el *Machine Learning*?

- *Extraer patrones o modelos a partir de los datos proporcionados a modo de ejemplos o instancias del sistema que se quiere aprender.*
- Según la tarea a realizar se puede clasificar en:
 - ✓ **Aprendizaje supervisado**
Cada ejemplo proporcionado lleva una etiqueta asociada.
Se pretende asignar la etiqueta correcta ante un nuevo ejemplo.
 - Etiqueta discreta = *Clasificación*
 - Etiqueta continua = *Regresión*
 - ✓ **Aprendizaje no supervisado**
Solo tenemos ejemplos, no etiquetas.
Se pretende agrupar los ejemplos en *clusters*, o también decidir si un nuevo caso es del mismo grupo que los anteriores.
 - ✓ **Sistemas recomendadores**
Tenemos un conjunto de emparejamientos, dado un nuevo elemento se pretende determinar la/s mejores parejas para él.
 - ✓ **Aprendizaje por refuerzo**
Hay información parcial que me dice como de bien estoy aprendiendo.

Ejemplos



Browse Through: 350 Data Sets

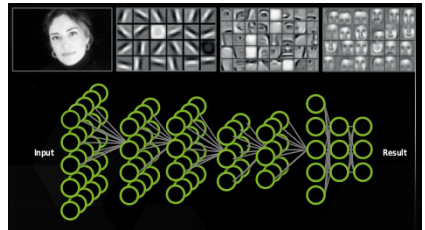
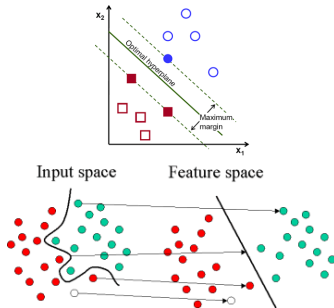
Default Task
Classification (256)
Regression (61)
Clustering (52)
Other (51)

Attribute type
Categorical (37)
Numerical (205)
Mixed (56)
Data Type
Multivariate (273)
Univariate (16)
Sequential (35)
Time-Series (62)
Text (30)
Domain-Theory (22)
Other (21)

Name	Data Types	Default Task	Att
 Abalone	Multivariate	Classification	Cat Inte
 Adult	Multivariate	Classification	Cat Inte
 Annealing	Multivariate	Classification	Cat Inte
 Anonymous Microsoft Web Data		Recommender-Systems	Cat
 Arthritis	Multivariate	Classification	Cat

Algunos métodos

- Árboles de clasificación
 - Modelos lineales generalizados
 - K-medias
- Máquinas de vectores soporte ► Aprendizaje profundo

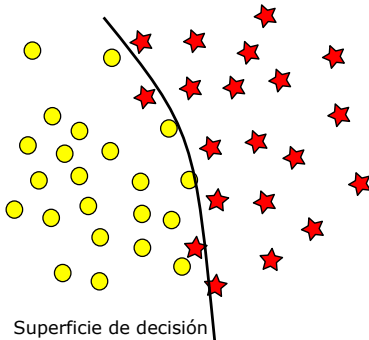


10 capas $\Rightarrow$ 1.000 millones de parámetros
¡¡ El cerebro tiene “sólo” 1.000 veces más!!
(Larry Brown, 2015)

Un ejemplo sencillo de clasificación

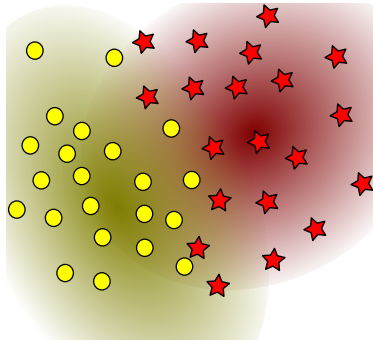
Queremos aprender:

La superficie de decisión



⇒ **Modelos Discriminativos**

La distribución de cada clase



⇒ **Modelos Generativos**

Formalizando el ejemplo

Notación

x son los ejemplos , y son las etiquetas

Discriminativos

¿Cuál es la probabilidad de que un ejemplo dado sea de una clase?

$$p(y|x) \quad ??$$

Generativos

¿Cómo están distribuidos los ejemplos de una cierta clase?

$$p(x|y) \quad ??$$

Relación

Teorema de Bayes:

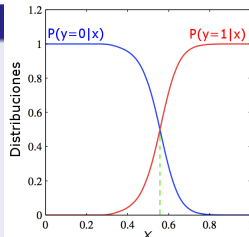
$$p(y|x) = \frac{p(y) \cdot p(x|y)}{p(x)} = \frac{p(x, y)}{p(x)}$$

Modelos discriminativos vs. generativos

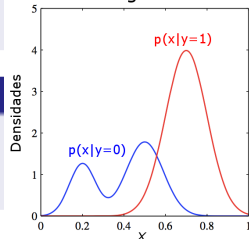
Modelos discriminativos

- ✓ Se construye directamente la *distribución a posteriori*, $P(y|x)$.
- ✓ Son más sencillos de construir.
Se utiliza una función sigmoide como salida del método de clasificación.
- ✗ ¿Qué ocurre cuando no tengo etiquetas?
- ✗ No utiliza conocimiento *a priori*, $p(y)$

Modelo discriminativo

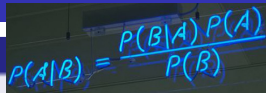


Modelo generativo



Modelos generativos

- ✓ Válido para aprendizaje no supervisado
- ✓ Se puede incorporar $p(y)$


$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Ideas clave de los modelos generativos

- ✓ **Clasificación:** La clase $\hat{y}$ estimada para un ejemplo x :

$$\hat{y} = \arg \max p(y)p(x|y)$$

- ✓ **Clustering:** Encontrar el modelo que mejor explica los datos x .

Modelo $\equiv$ Parámetros de la distribución $p(x)$; e.g.: $\Theta = \{\mu, \sigma\}$

$$\Theta^* = \arg \max p(x|\Theta)$$

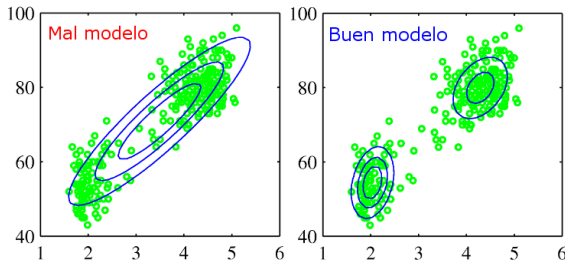
Tipos

- Modelos de Mezclas de Gaussianas (GMM)
- Modelado mediante funciones Cópula
- Modelos gráficos probabilísticos (PGM)
- Otros...

A. Conan Doyle: “El sabueso de los Baskerville”

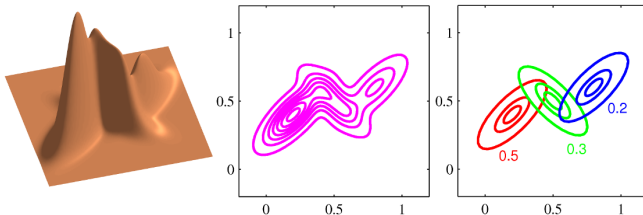
(1902)

...uno comienza a deformar los hechos para hacerlos encajar en las teorías en lugar de encajar las teorías en los hechos.



- Los (GMM) son un método potente de construir distribuciones complejas a partir de otras más sencillas.

- Un GMM es una suma ponderada de distribuciones normales.



- Formalmente:

$$P(x) = w_1 \cdot \mathcal{N}(x|\mu_1, \sigma_1) + w_2 \cdot \mathcal{N}(x|\mu_2, \sigma_2) + \dots$$

cumpliendo que $w_1 + w_2 + \dots = 1$

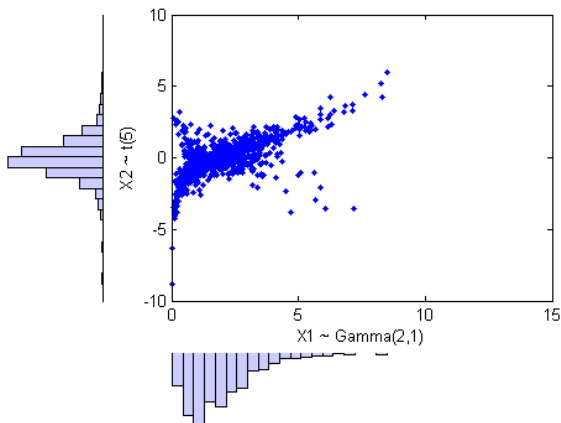
- Conjunto de parámetros $\Theta = \{\mu_1, \mu_2, \dots, \sigma_1, \sigma_2, \dots, w_1, w_2, \dots\}$
- **Estimación** de Θ mediante EM.

Ventajas y aplicaciones

- Es un método rápido y sencillo .
- Permite manejar distribuciones multimodales.
- Evidentemente en cuanto se estiman los parámetros se obtienen clusters de los datos.
- Si los datos están etiquetados simplemente hay que modelar cada clase por separado.
- Es un método frecuente para *Sustracción de fondo* en visión artificial
 - Se guarda el valor de cada pixel en los N fotogramas pasados.
 - Es esperable que el valor del pixel de fondo sea aquel que más aparece porque está fijo.
 - Pero la distribución puede tener más de una moda (máximo) si algún objeto en movimiento se queda quieto.

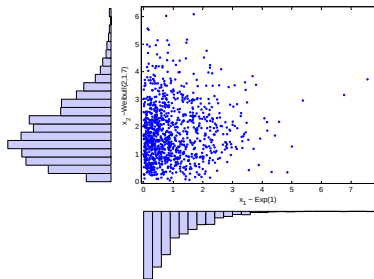
Desventajas

- ¿Y si los datos tienen una distribución NO Gaussiana pero conocida?



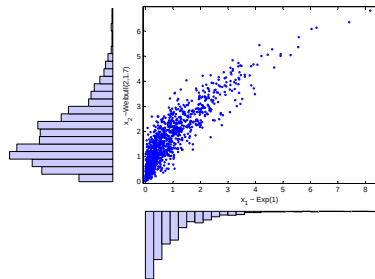
Independencia vs. Dependencia

Independencia



$$p(x_1, x_2) = p(x_1)p(x_2)$$

Dependencia



$$p(x_1, x_2) = p(x_1)p(x_2)c(x_1, x_2)$$

función Cópula

- $c(x_1, x_2)$ es el término que explica la dependencia entre las variables.
- Si las variables son continuas, C siempre existe y es único.

MIT News

Browse

or

Search



System predicts 85 percent of cyber-attacks using input from human experts

Virtual artificial intelligence analyst developed by the Computer Science and Artificial Intelligence Lab and PatternEx reduces false positives by factor of 5.

Adam Conner-Simons | CSAIL
April 18, 2016

▼ Press Inquiries

PRESS MENTIONS

Today's security systems usually fall into one of two categories: human or machine. So-called "analyst-driven solutions" rely on rules created by living experts and therefore miss any attacks that don't match the rules. Meanwhile, today's machine-learning approaches rely on "anomaly detection," which tends to trigger false positives that both create distrust of the system and end up having to be investigated by humans, anyway.

But what if there were a solution that could merge those two worlds? What would it look like?

In a new paper, researchers from MIT's [Computer Science and Artificial Intelligence Laboratory](#) (CSAIL) and the machine-learning startup [PatternEx](#) demonstrate an artificial intelligence platform called AI^2 that predicts cyber-attacks significantly better than existing systems by continuously incorporating input from human experts. (The name comes from merging artificial intelligence with what the researchers call "analyst intuition.")

bution are considered outliers. The outlier score is simply the probability density of a point in the multidimensional space.

To build a multivariate model from marginal distributions which are not all Gaussian, we exploit *copula* functions. A copula framework provides a means of inference after model-

MIT researchers have developed an artificial intelligence platform that uses input from human analysts to predict cyber-attacks, reports Brian Mastroianni for CBS News. "We realized, finding the actual attacks involved a mix of supervised and unsupervised machine-learning," explains research scientist Kalyan Veeramachaneni.

CBS NEWS

Wired reporter Brian Barrett writes that MIT researchers have developed a new system to help detect cyber-attacks. Barrett explains that the system, "reviews data...and pinpoints anything suspicious. A human takes it from there, checking for signs of a breach. The one-two punch identifies 86 percent of attacks while sparing analysts the tedium of chasing bogus leads."

WIRED

MIT News

Browse

or

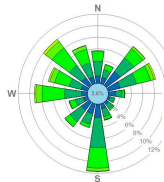
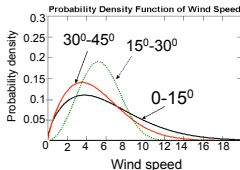
Search



Siting wind farms more quickly, cheaply

New model predicts wind speeds more accurately with three months of data than others do with 12.

Larry Hardesty | MIT News Office
July 17, 2015

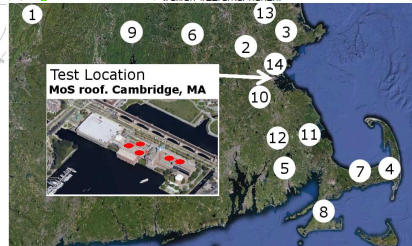


▼ Press Inquiries

RELATED

Paper: "Copula graphical models for wind resource estimation"

Kelven Vaaramananeni



¿Y si las variables son discretas?

Cóputas

Pero también hay otros modelos.

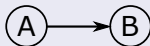
Work in Progress

$$p(y, x_1, x_2, x_3, \dots) = p(y) \cdot p(x_1|y) \cdot p(x_2|x_1, y) \cdot p(x_3|x_2, x_1, y) \cdots$$

- **Independencia condicional**

' x_i ' y ' x_j ' son independientes dado ' y ' si $p(x_1|x_2, y) = p(x_1|y)$

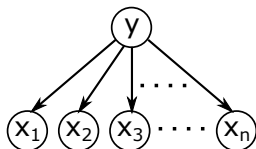
- Son una manera gráfica de representar la independencia condicional.



$$p(A, B) = p(A) \cdot p(B|A)$$

- Descubriendo o asumiendo independencia condicional podemos simplificar la distribución conjunta.

Naive Bayes

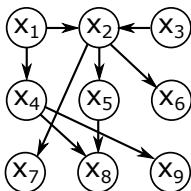


Distribución

$$p(y, x_1, x_2, \dots) = p(y) \cdot p(x_1|y) \cdot p(x_2|y) \cdots$$

- Las características se modelan independientemente para una clase.
- **Aplicación:** filtros anti-spam
 - ✓ y es la etiqueta *spam*
 - ✓ $\{x_1, \dots, x_n\}$ es una base de datos con las palabras que aparecen en los correos electrónicos con *spam* y el recuento de veces que aparecen.
 - ✓ N = número de mensajes spam vistos; M_i = recuento de la palabra i .

$$p(x_i|y) = M_i/N$$

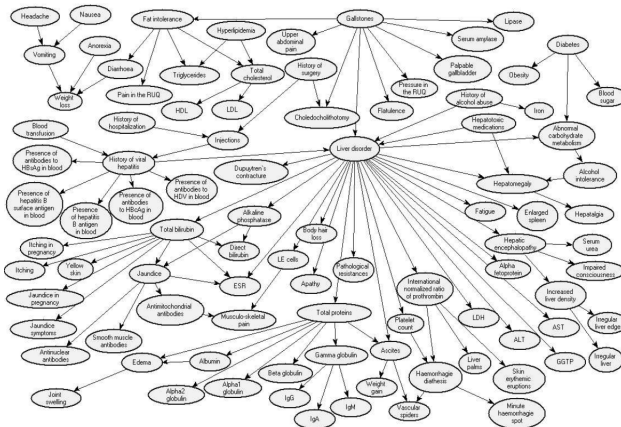


$$p(x_1, x_2, x_3, \dots, x_9) = p(x_1) \cdot p(x_3) \cdot p(x_4|x_1) \cdot p(x_2|x_1, x_3) \cdot p(x_5|x_2) \cdot p(x_6|x_2) \cdot p(x_7|x_2) \cdot p(x_8|x_4, x_5) \cdot p(x_9|x_4)$$

- Grafo dirigido acíclico (DAG)
- **Aplicación:** “Bayesian Networks” @ IEEE , Journals , [2006,2016]
1.180 coincidencias.
 - ✓ A B.N. approach to traffic flow forecasting (2006)
 - ✓ Wavelet B.N. Image Denoising (2013)
 - ✓ B.N. applied to Failure Diagnosis in Power Transformer (2013)
 - ✓ A Hierarchical B.N.-Based Approach to Keyword Auction (2015)

A Bayesian Network Model for Diagnosis of Liver Disorders

¹ Decision Systems Laboratory, School of Information Sciences, Intelligent Systems Program, and Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A.,



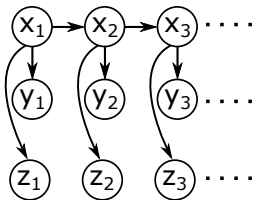
```

graph LR
    x1((x1)) --> x2((x2))
    x2 --> x3((x3))
    x1 --> y1((y1))
    x2 --> y2((y2))
    x3 --> y3((y3))
    x3 -.- dots1[...]
    y3 -.- dots2[...]
  
```

- Los ejemplos se suceden en el tiempo.
- Memorizar todos los ejemplos pasados no es práctico $\Rightarrow$ Modelar el estado

- Notación diferente.
 - x_t es el estado o variable oculta en t .
 - y_t es el ejemplo observado en t .
- Hay un modelo de observación $P(y_t|x_t)$ y un modelo de transición $P(x_{t+1}|x_t)$
- **Aplicación:** Filtrado y Smoothing

HMM doble

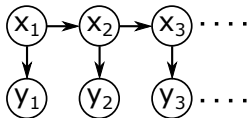


Puede haber varios observables

- Básicamente es el mismo modelo, pero se asume que los observables son condicionalmente independientes.

- Por ejemplo, dos micrófonos (Y, Z) grabando generan y_t, z_t ; queremos determinar en cada t de cual de los dos proviene el sonido.

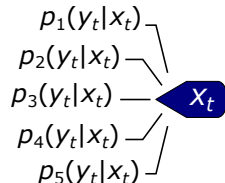
Markov Switching Model (MSM)



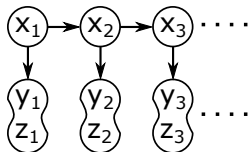
... y varios modelos

- El modelo observable depende del valor del estado $\Rightarrow$ el modelo global tiene varias dinámicas.

- Aplicación:** En mercados de compra/venta de productos financieros es muy frecuente que la dinámica alcista sea diferente de la dinámica bajista.



Markov Switching Copula Model (MSCM)

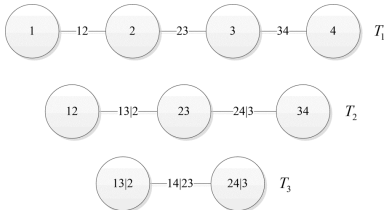


Uno para gobernarlos a todos

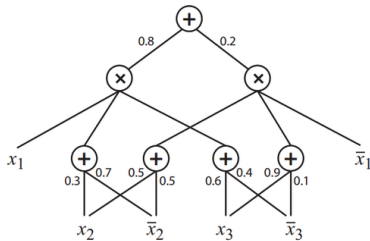
- Podemos NO suponer que los observables son condicionalmente independientes

- Utilizamos una cópula para unir los observables en un modelo.
- Si luego resulta que son independientes, $c(y_t, z_t) = 1$.
- **Aplicación:** “MSCM For Longitudinal Data”
 - A. Cuesta-Infante, K. Veeramachaneni; presentado al ICDM'16.
 - ✓ Clustering secuencial de alumnos en EdX según progresa el curso.
 - ✓ Predicción de AHE en pacientes UCI.

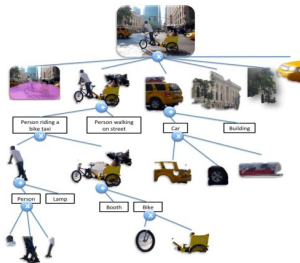
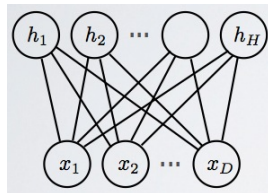
► Vines



► Sum-Product Networks



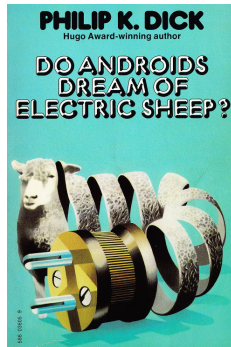
► Restricted Boltzman Machines



¿Por qué se llama así esta charla?

- Hemos recorrido varios métodos de obtener modelos generativos.
- Construir la distribución de las características.
 - ✓ Clasificación a través del T. de Bayes
 - ✓ Clustering
 - ✓ Datos secuenciales y no secuenciales
 - + Es posible simular NUEVAS muestras !!
- *¿Por qué es posible la inteligencia artificial?*
A.Ruiz de Elvira, El Mundo, 31/ene/2016.

La inteligencia humana solo es posible en unas mentes que tienen la libertad de generar conexiones neuronales nuevas. Y la libertad sólo es posible en esquemas aleatorios, no deterministas,



Créditos: Algunas figuras han sido tomadas del libro
Pattern Recognition and Machine Learning, de C.M. Bishop

Muchas gracias
¿Preguntas?