Universidad Popular del Cesar

Ciencia de Datos y BigData

Aprendizaje No Supervisado

Dr. José Ramón Iglesias
DSP-ASIC BUILDER GROUP
Director Semillero TRIAC
Ingenieria Electronica
Universidad Popular del Cesar

Qué es aprendizaje no supervisado?

Análisis exploratorio de datos (*Exploratory Data Analysis*)

- Detección de anomalías
- Prevención de fallas

Data mining

- Detección de patrones
- Reglas de asociación
- Segmentación de perfiles

Acercarse a las **causas latentes** de los **fenómenos observables**

Para qué?

- Cuando no sabemos lo que queremos
- Cuando sospechamos de los datos
- Para refinar las clases que queremos

La promesa de NO supervisado

A partir de datos crudos obtenemos patrones accionables (reglas, clases)

La promesa de NO supervisado

A partir de datos crudos obtenemos patrones accionables (reglas, clases)

- los datos nunca pueden ser crudos
- los resultados nunca están listos para ser usados
  - no es claro qué significan, hay que interpretarlos

    ➜ lo que ahorramos en anotación de datos lo gastamos en análisis de resultados

La promesa de NO supervisado

A partir de datos crudos obtenemos patrones accionables (reglas, clases)

- los datos nunca pueden ser crudos
- los resultados nunca están listos para ser usados
  - no es claro qué significan, hay que interpretarlos

    ➜ lo que ahorramos en anotación de datos lo ~~gastamos~~ invertimos en análisis de resultados

La promesa de NO supervisado

A partir de datos crudos obtenemos patrones accionables (reglas, clases)

- los datos nunca pueden ser crudos
- los resultados nunca están listos para ser usados
  - no es claro qué significan, hay que interpretarlos

    ➜ lo que ahorramos en anotación de datos lo ~~gastamos~~ invertimos en análisis de resultados
  - hay que consultar con el experto de dominio

La promesa de NO supervisado

A partir de datos crudos obtenemos patrones accionables (reglas, clases)

- los datos nunca pueden ser crudos
- los resultados nunca están listos para ser usados
  - no es claro qué significan, hay que interpretarlos
    ➜ lo que ahorramos en anotación de datos lo ~~gastamos~~ invertimos en análisis de resultados
  - hay que consultar con el experto de dominio
  - iterar varias veces

La promesa de NO supervisado

A partir de datos crudos obtenemos patrones accionables (reglas, clases)

- los datos nunca pueden ser crudos
- los resultados nunca están listos para ser usados
  - no es claro qué significan, hay que interpretarlos
    ➜ lo que ahorramos en anotación de datos lo ~~gastamos~~ invertimos en análisis de resultados
  - hay que consultar con el experto de dominio
  - iterar varias veces

nosotras prometemos darles herramientas y enseñarles cómo no usarlas

Tecnologías relacionadas

- ANOVA, testeo de hipótesis
- Proyecciones (*embeddings*)
- Reglas de asociación
- Vecinos más cercanos (recomendación)
- Clustering
- Detección de Anomalías
- Propiedades de Grafos
- Modelos de lenguaje

Problemas metodológicos

- No hay evaluación intrínseca
  - Evaluación indirecta, por impacto en otras aplicaciones
  - Evaluación interpretativa subjetiva
- Evaluación anecdótica, nunca exhaustiva
- Medidas de calidad de utilidad cuestionable

La clave está en

- hacer **buenas preguntas**,
- exprimir los datos buscando respuestas,
- cuestionar todas las respuestas

- Un espacio de búsqueda muy grande ➜ mínimos locales

Aplicaciones clásicas

- Análisis del carrito de la compra
- Segmentación de mercado (clientes)
- Caracterización epidemiológica de población (enfermos)
- Caracterización de comportamiento de usuarios (web, celular, redes sociales, electricidad)
- Detección de fallos en líneas de producción
- Detección de fraude (tarjetas de crédito, impuestos)
- Detección de temas en documentos
- Detección de tipos de objetos en imágenes
- Detección de comunidades

Perspectiva general

Clustering

Agrupar mis datos, viendo qué elementos son semejantes entre sí

Perspectiva general

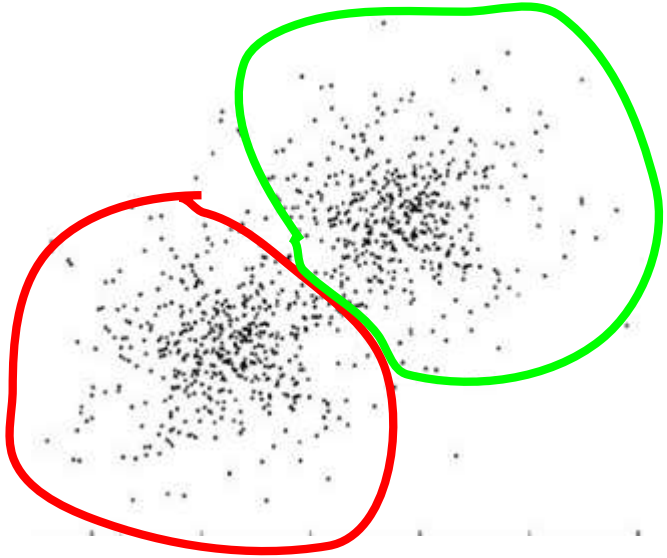
Clustering

Agrupar mis datos, viendo qué elementos son semejantes entre sí

Perspectiva general

Clustering

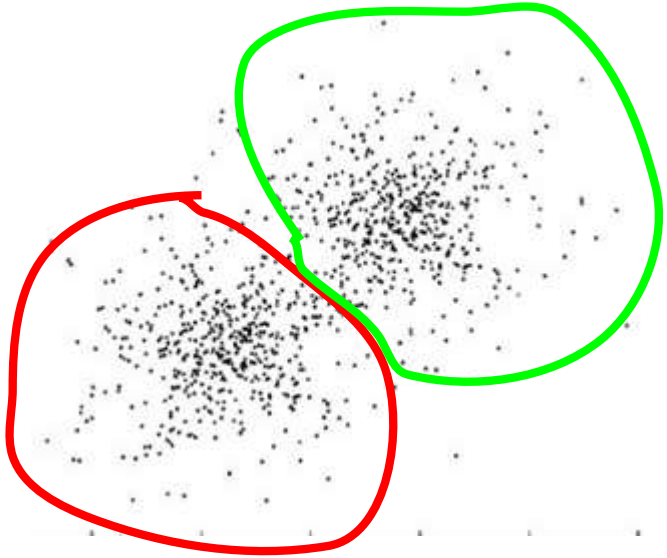
Agrupar mis datos, viendo qué elementos son semejantes entre sí



Perspectiva general

Clustering

Agrupar mis datos, viendo qué elementos son semejantes entre sí



Tengo clases!

gratis!

Perspectiva general

Clustering

Agrupar mis datos, viendo qué elementos son semejantes entre sí

- ¿Con qué caracterizo a mis datos?
- Medida de distancia, criterio de aglomeración o separación
- ¿Cuántos grupos?
- ¿Cómo evalúo?
    - Medidas geométricas
    - Inspeccionar el contenido
    - Medir la correspondencia del contenido por correspondencia con algún conocimiento del dominio (pares de elementos bandera, clases)

Perspectiva general

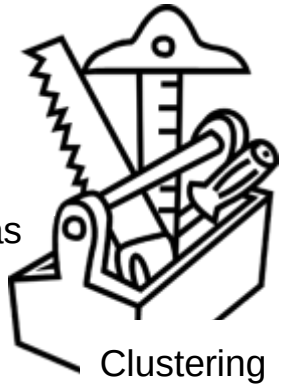

Clustering

Selección de Características

- Eliminar características que introducen ruido
- Eliminar características redundantes
- Quedarse con las características más determinantes
  ➜ ¿cómo determinamos cuán determinantes son?
  en aprendizaje supervisado, por co-varianza con la clase o por cómo las usa un clasificador, pero en no-supervisado?
    ➜ nos inventamos una *tarea de pretexto*!
- Agrupar características
  ➜ usando un clasificador

Perspectiva general

Selección de Características



Clustering

Reglas de asociación

Detección de patrones en el sentido más intuitivo:

La probabilidad condicional hecha regla

- muchísimos patrones, y la mayoría triviales
  - ➜ cómo encontrar las reglas
  - ➜ cómo filtrarlas y ordenarlas

Perspectiva general

Aprendizaje semi-supervisado

Combinar unos poquitos datos supervisados con no supervisados

Podemos usar los mismos ejemplos bandera, o reglas

Problemas:

- deriva semántica (propagación del error)
- regiones del espacio que no se cubren
- evaluación

El aprendizaje activo suele ser una buena forma de atacarlos

Perspectiva general

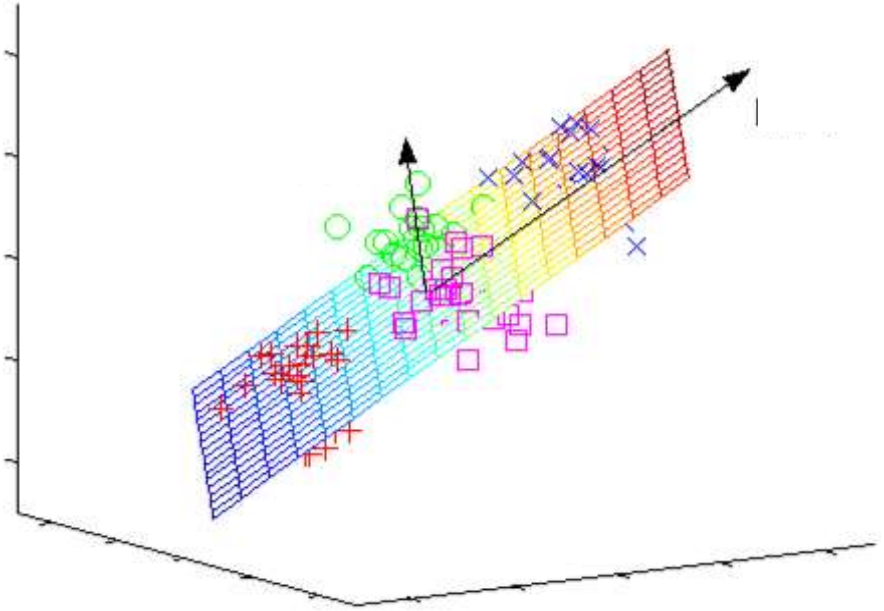
Embeddings

Un embedding es una proyección a otro espacio

- Selección de características
- Principal Component Analysis

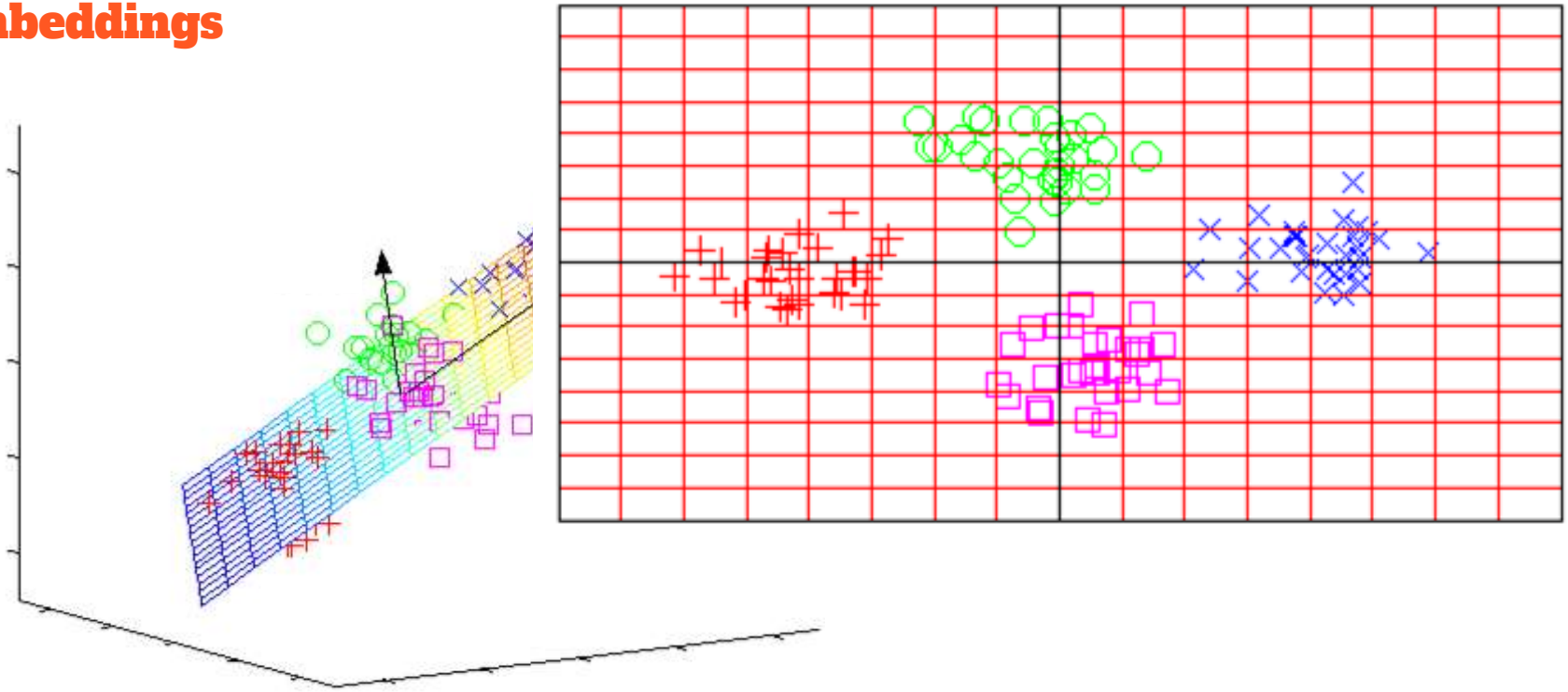
Perspectiva general

Embeddings



Perspectiva general

Embeddings

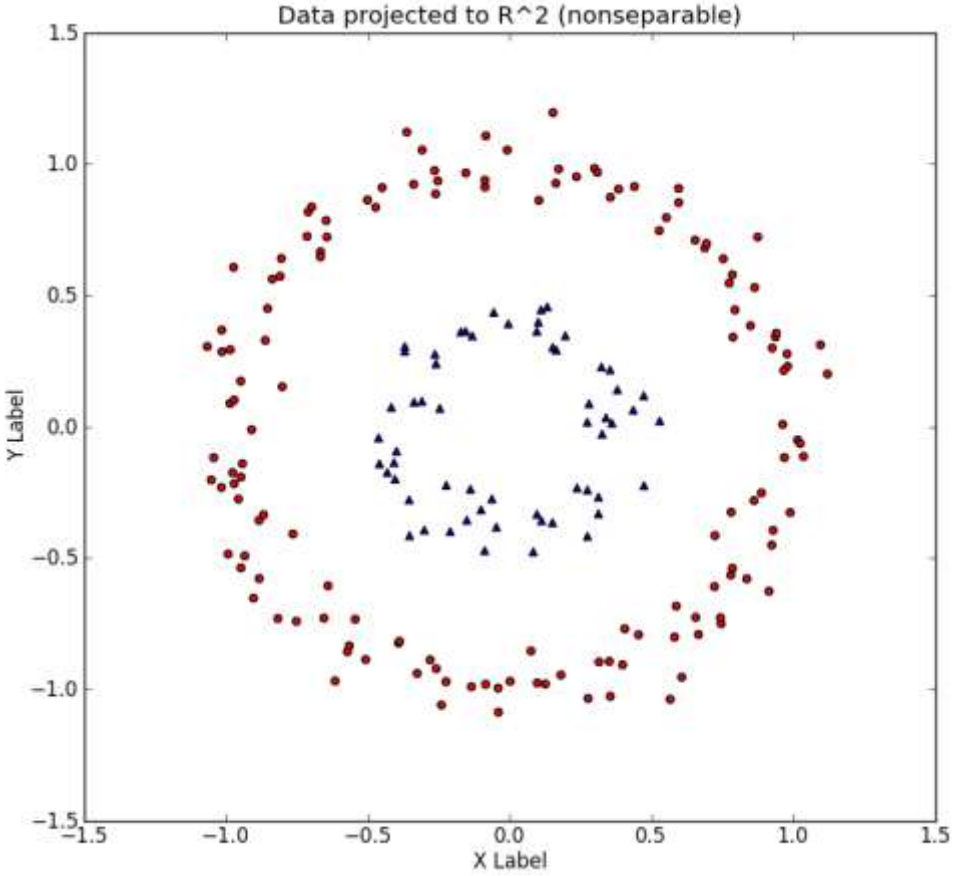


Perspectiva general

Embeddings

Un embedding es una proyección a otro espacio

- Selección de características
- Principal Component Analysis
- Kernel trick



Semi-supervisado

Selección de Características

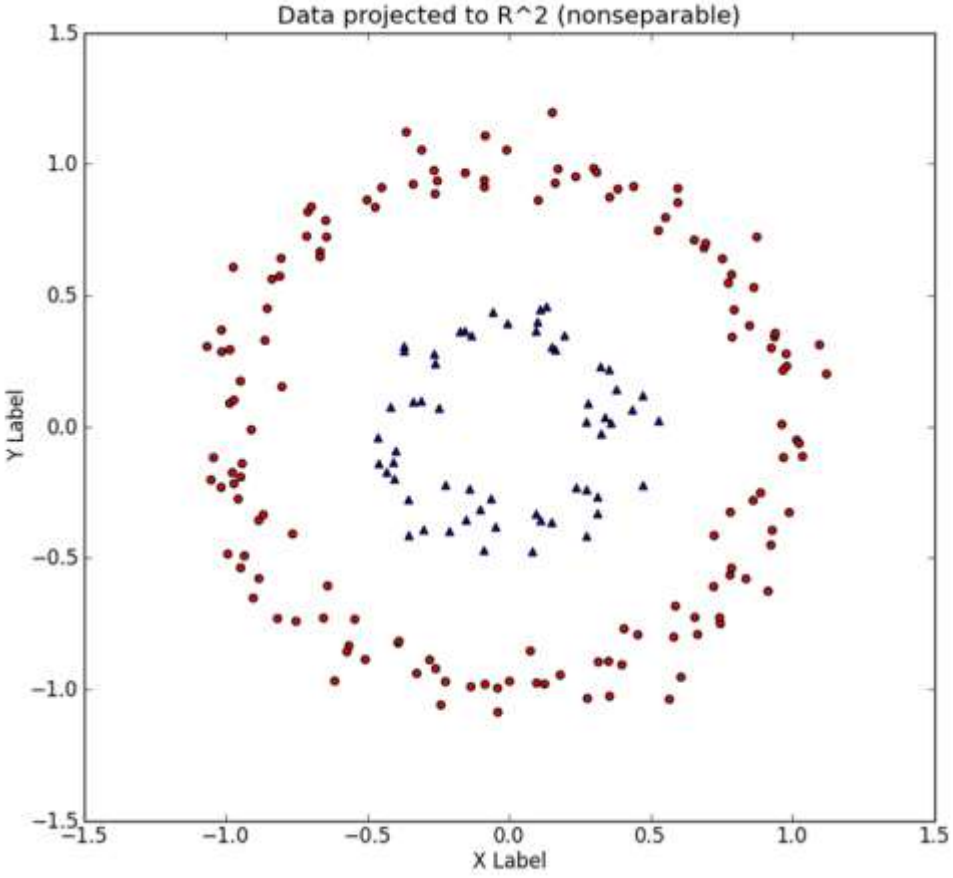
Reglas de Asociación

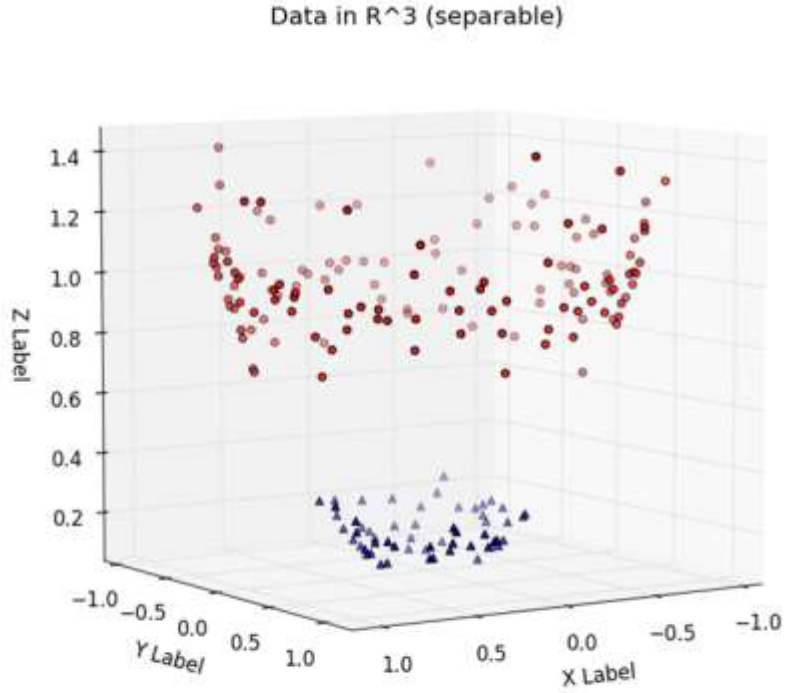
Clustering

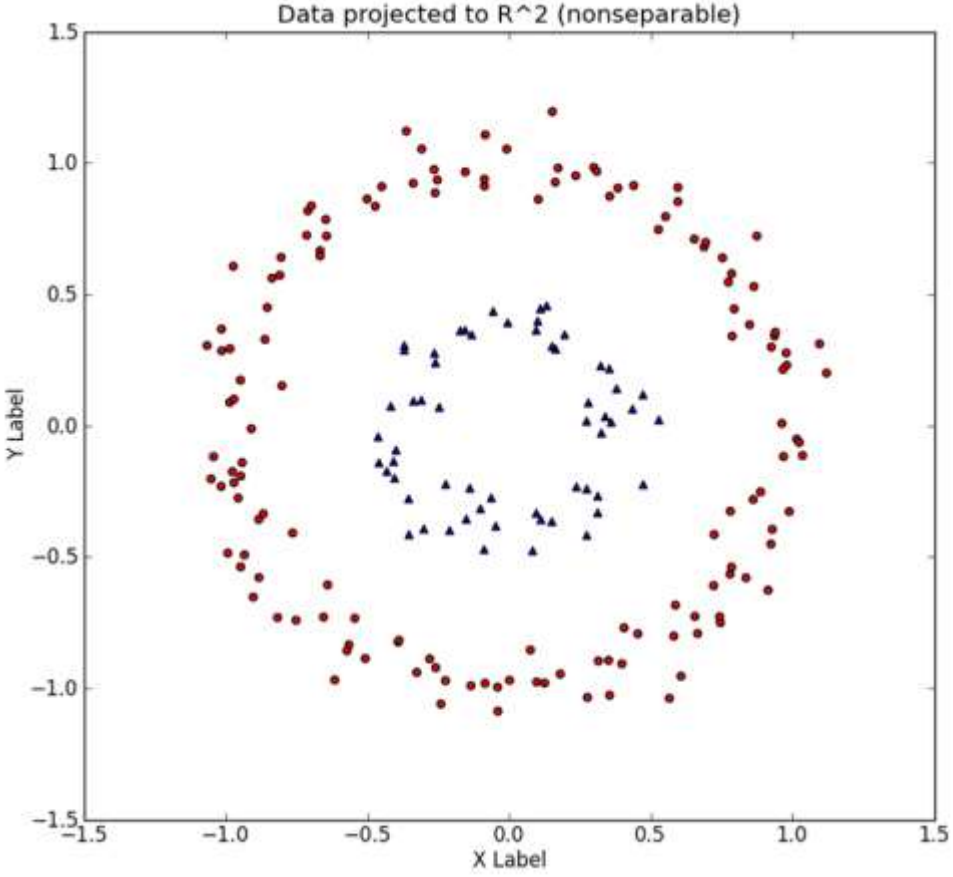
spacio

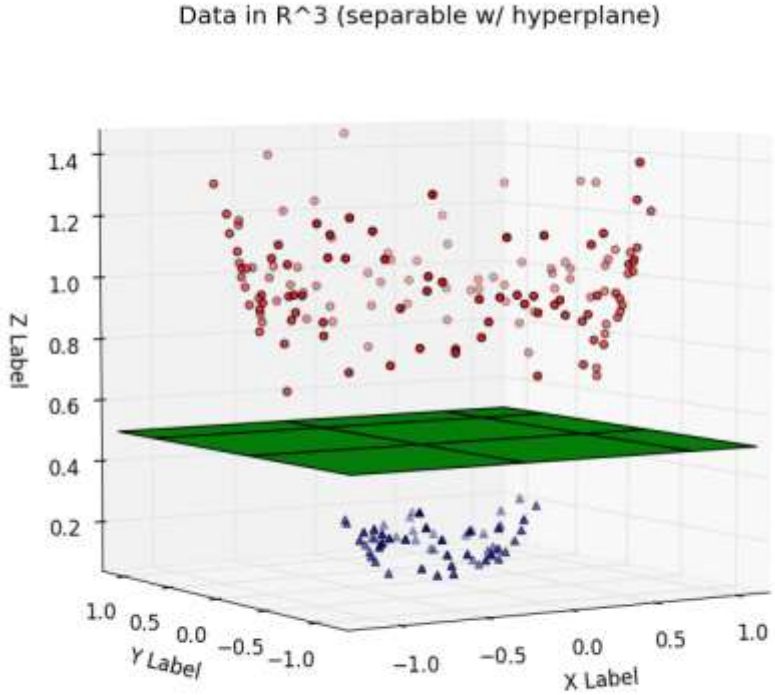
a de pretexto para aprender un

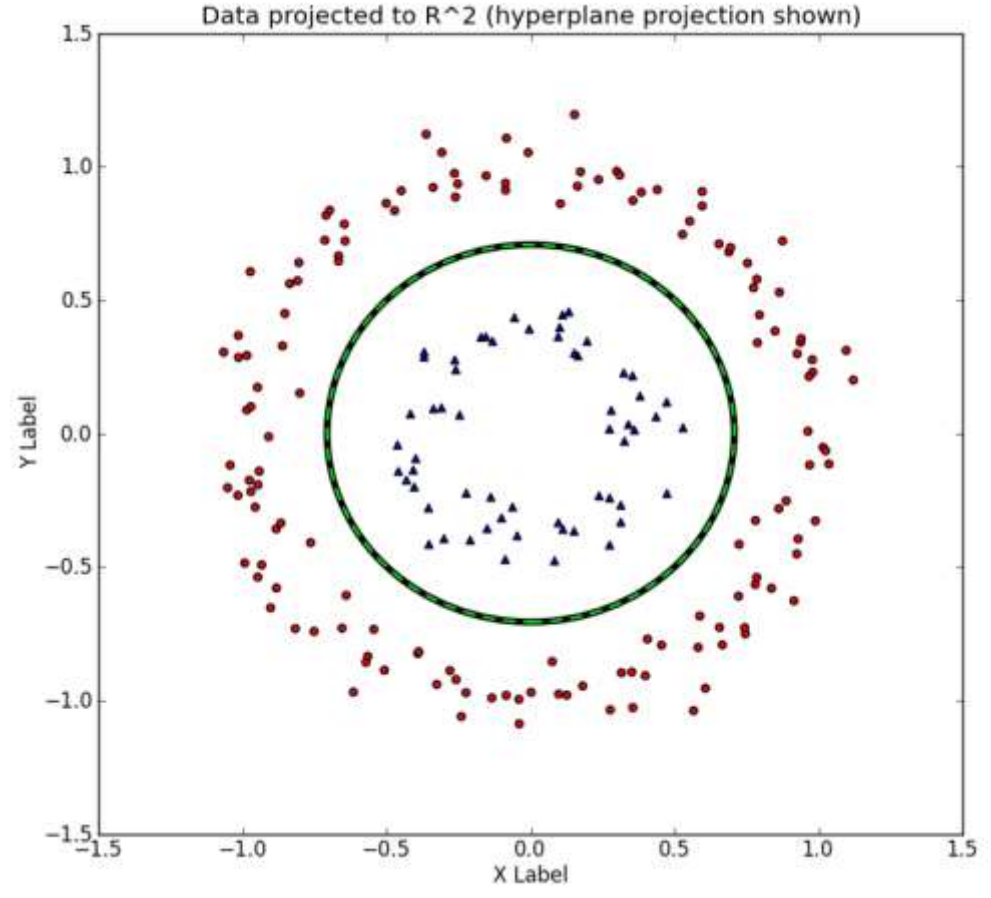
del clasificador como nuevo espacio

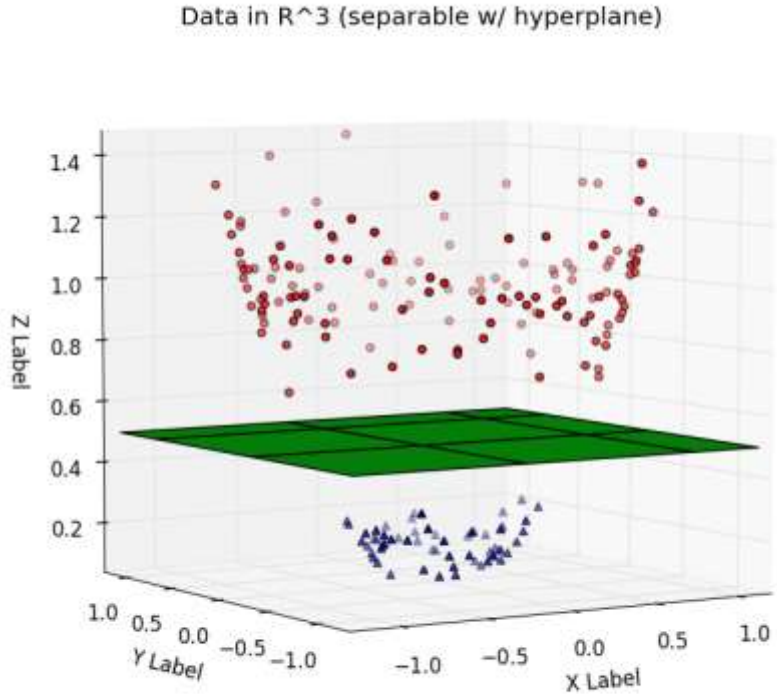

Data projected to R^2 (nonseparable)
Y Label
X Label
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


Data in R^3 (separable)
Z Label
Y Label
X Label
1.4
1.2
1.0
0.8
0.6
0.4
0.2
−1.0
−0.5
0.0
0.5
1.0


Data projected to R^2 (nonseparable)
Y Label
X Label
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Data in R^3 (separable w/ hyperplane)
Z Label
Y Label
X Label
1.4
1.2
1.0
0.8
0.6
0.4
0.2


Data projected to R^2 (hyperplane projection shown)
Y Label
X Label
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Data in R^3 (separable w/ hyperplane)
Z Label
Y Label
X Label
1.4
1.2
1.0
0.8
0.6
0.4
0.2

Perspectiva general

Embeddings

Un embedding es una proyección a otro espacio

- Selección de características
- Principal Component Analysis
- Kernel trick
- Filtros, preprocesos
- Embeddings neuronales: usar una tarea de pretexto para aprender un clasificador, y quedarse con el modelo del clasificador como nuevo espacio

Evaluación

Si no sabemos qué es lo bueno, cómo podemos calcular acierto o error?

- necesitamos un experto de dominio que analice resultados: adecuación, plausibilidad con respecto a la intuición o a la aplicación
- métricas de rendimiento en tiempo de desarrollo
- métricas de rendimiento en aplicación final
- métricas de consistencia del modelo: acumulación de probabilidad donde pensamos que tiene que estar, replicabilidad, cobertura...