

Aprendizaje Semi-supervisado



Diplomatura en Ciencia de Datos,
Aprendizaje Automático y sus Aplicaciones
FaMAF-UNC
agosto 2018

Para saber más

Un buen tutorial de Jerry Zhu

<http://pages.cs.wisc.edu/~jerryzhu/pub/sslchicago09.pdf>

Xiaojin Zhu and Andrew B. Goldberg. Introduction to Semi-Supervised Learning. Morgan & Claypool, 2009.

Contexto

Los datos etiquetados son escasos y caros

Los datos no etiquetados son abundantes y gratis

Objetivo: aprender de datos etiquetados y no etiquetados, para obtener:

- Menos overfitting, mejor generalización
- Más capacidad para tratar ejemplos no vistos

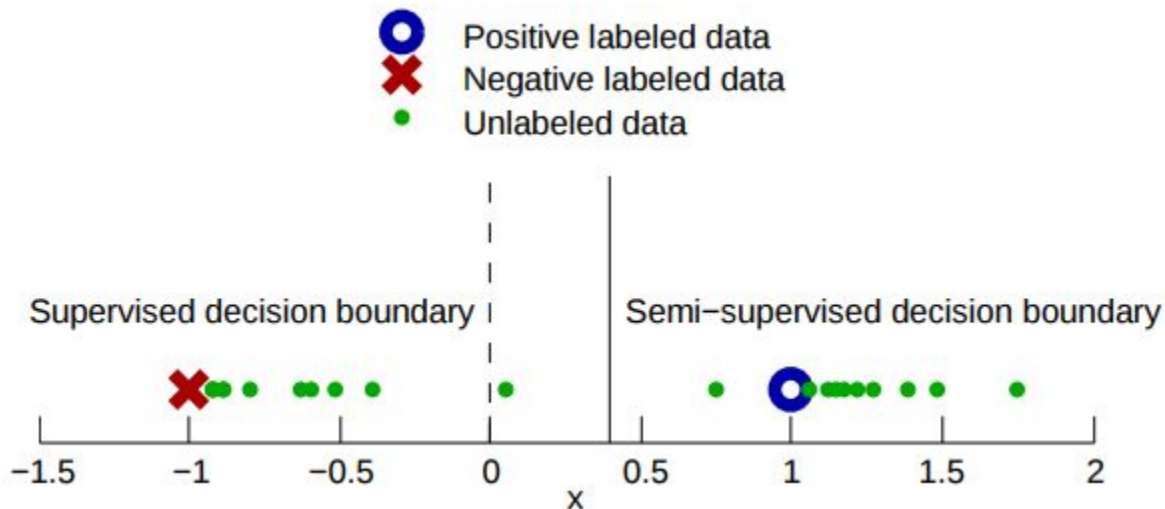
También: usar datos etiquetados para mejorar algoritmos no supervisados

- Clustering with rules
- Reglas de asociación con clase

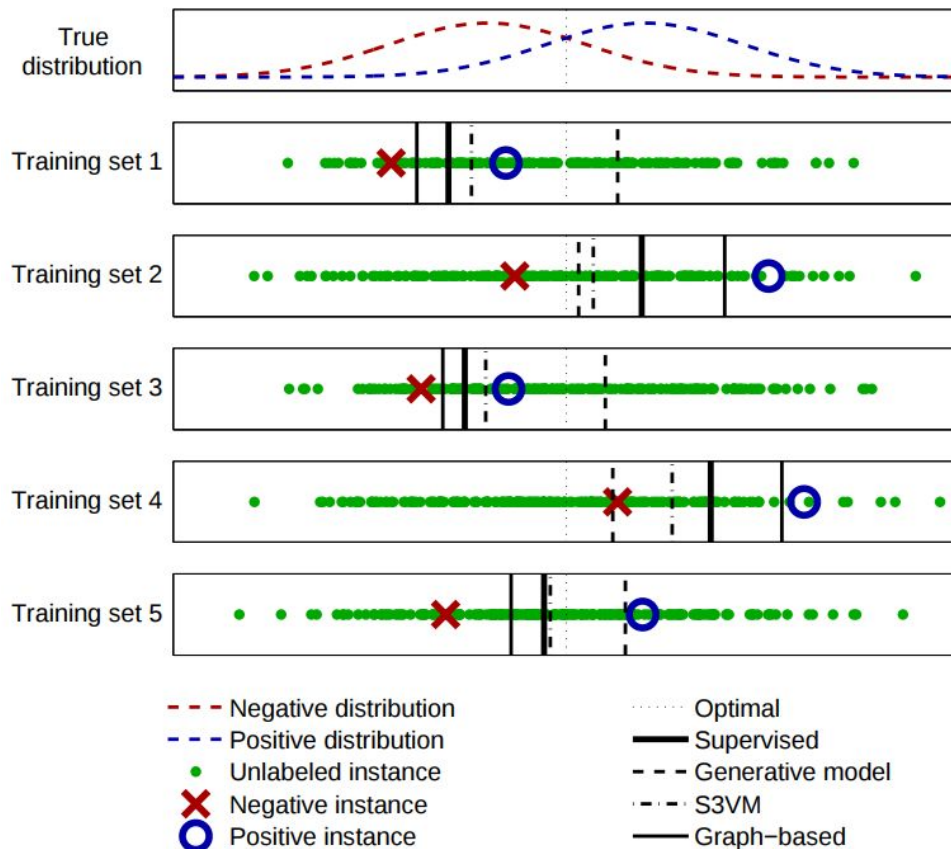
Fundamento cognitivo

Promesa: mejorar la
performance gratis!

Cómo ayudan los datos no etiquetados?



Asunciones equivocadas... empeoran



Notación

ejemplo de entrada x , etiqueta y

aprendedor

$$f: X \rightarrow Y$$

datos etiquetados

$$(X_p, X_y) = \{(x_{1:p}, y_{1:p})\}$$

datos no etiquetados

$$X_u = \{x_{l+1:n}\} \quad \text{disponibles durante entrenamiento}$$

normalmente: $l \ll n$

datos de test

$$X_{\text{test}} = \{x_{n+1:n}\} \quad \text{NO disponibles durante entrenamiento}$$

Modelos disjuntos vs. conjuntos

Aprender conjuntamente vs. concatenar módulos

Autoaprendizaje (self-learning) (bootstrapping)

Algoritmo de autoaprendizaje

1. Obtener un conjunto pequeño de datos etiquetados
 2. Aprender un clasificador de los datos etiquetados
 3. Aplicar el clasificador sobre datos no etiquetados
 4. Incorporar datos etiquetados automáticamente al conjunto de entrenamiento
 5. Volver a 2.
-
- ¿Qué ejemplos etiquetados automáticamente incorporamos?
 - Mayor confianza
 - Los n mejores
 - Todos

Un ejemplo: Yarowsky (1995)

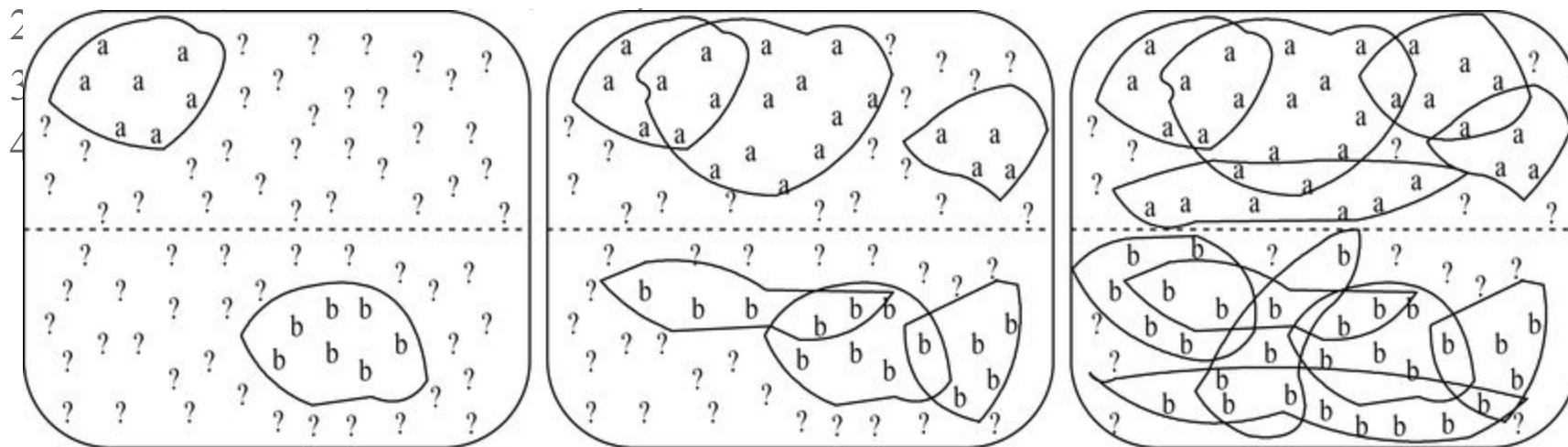
Desambiguación de palabras

1. Ejemplos iniciales
2. Aprender una lista de decisión
3. Buscar más ejemplos con la lista
4. Iterar a 2.

Un ejemplo: Yarowsky (1995)

Desambiguación de palabras

1. Ejemplos iniciales



Un ejemplo: Yarowsky (1995)

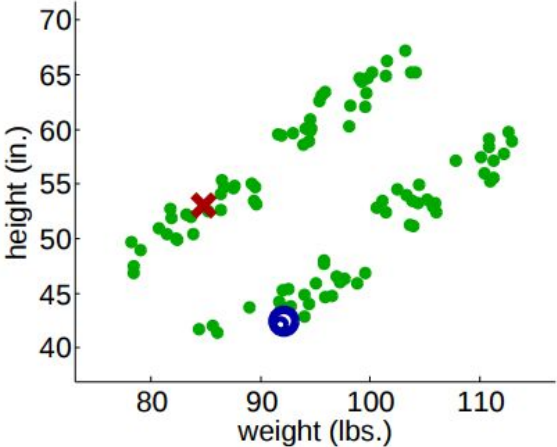
Desambiguación de palabras

One sense per collocation

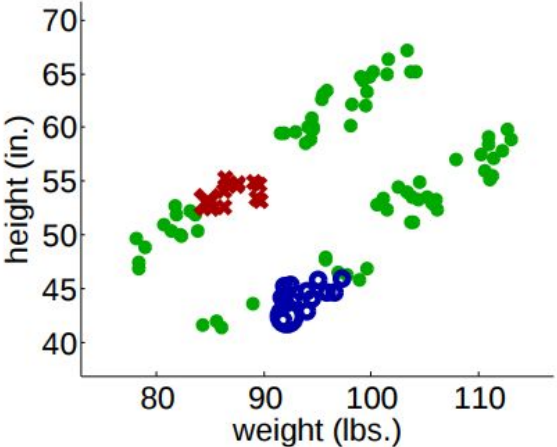
1. Ejemplos iniciales
2. Aprender una lista de decisión
3. Buscar más ejemplos con la lista
4. Iterar a 2.

One sense per discourse

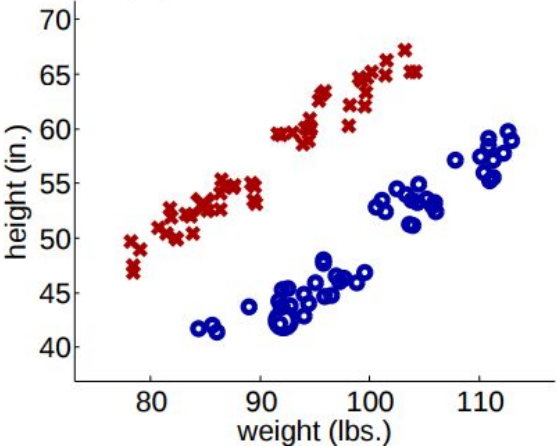
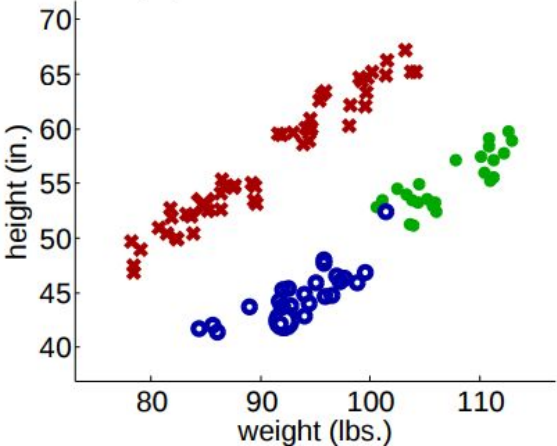
En cada documento, la misma palabra tiene siempre el mismo sentido

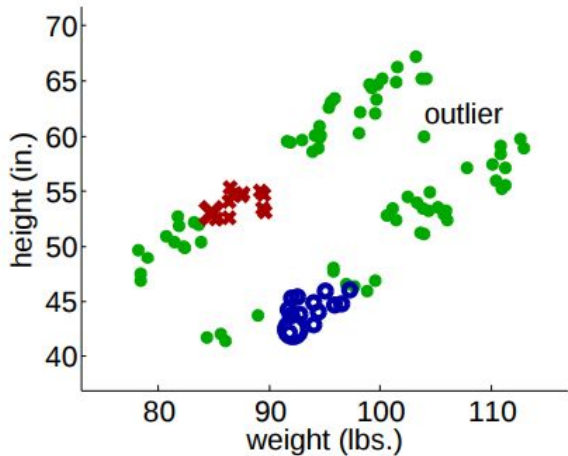


(a) Iteration 1

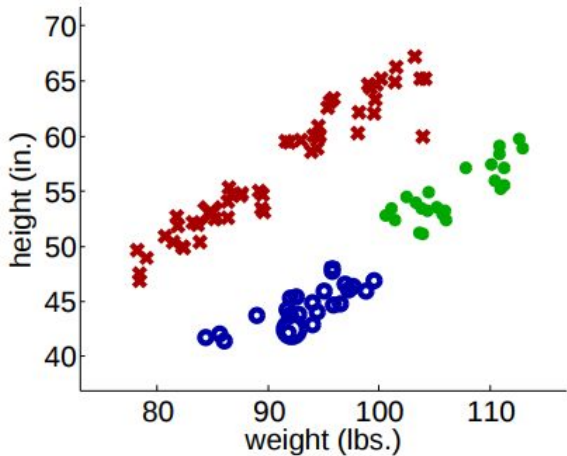


(b) Iteration 25

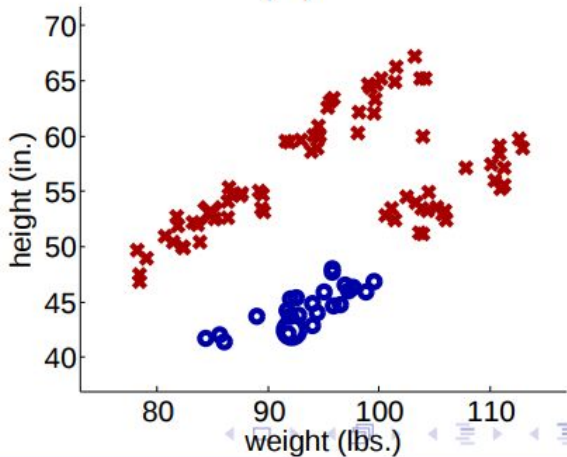
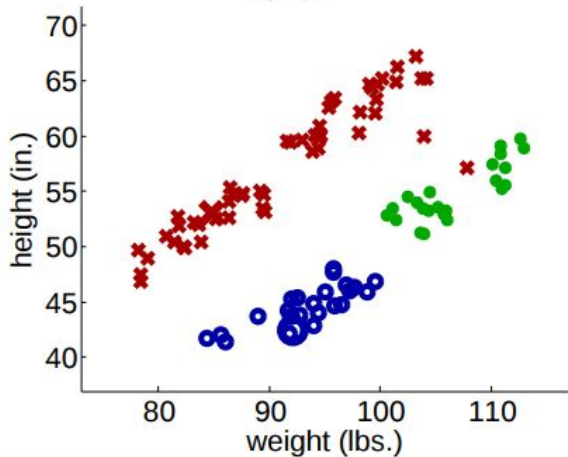




(a)



(b)



Valoración de autoaprendizaje

Ventajas:

- Muy fácil de implementar
- Se adapta a cualquier aprendedor (es un wrapper)
- Funciona muy bien para muchas tareas

Desventajas:

- Deriva semántica (Amplificación del error)
- Puede haber regiones del espacio a las que no llega

Valoración de autoaprendizaje

Ventajas:

- Muy fácil de implementar
- Se adapta a cualquier aprendedor (es un wrapper)
- Funciona muy bien para muchas tareas

Desventajas:

- Deriva semántica (Amplificación del error) ← estrategias correctivas
- Puede haber regiones del espacio a las que no llega ← estrategias complementarias

Co-aprendizaje (co-training)

Combinar estrategias complementarias

Aprendedores complementarios sobre diferentes facetas de un mismo objeto

- Página web / producto: imagen y texto
- Entidades nombradas: palabra y contexto

Algoritmo de co-aprendizaje

1. Obtener un conjunto pequeño de datos etiquetados
 2. Aprender **dos** clasificadores **complementarios** de los datos etiquetados
 3. Aplicar los clasificadores sobre datos no etiquetados
 4. Incorporar datos etiquetados automáticamente al conjunto de entrenamiento
 5. ¿Eliminar datos etiquetados automáticamente del conjunto de entrenamiento?
 6. Volver a 2.
-
- ¿Qué ejemplos etiquetados automáticamente incorporamos?
 - Mayor confianza, uno solo, ambos?
 - Donde los dos clasificadores estén de acuerdo

Valoración de co-aprendizaje

Ventajas:

- Muy fácil de implementar
- Se adapta a cualquier aprendedor (es un wrapper)
- Funciona muy bien para muchas tareas

Desventajas:

- Muchos problemas no se dividen bien en facetas disjuntas
- Es posible que un solo clasificador usando ambas facetas tenga mejor desempeño

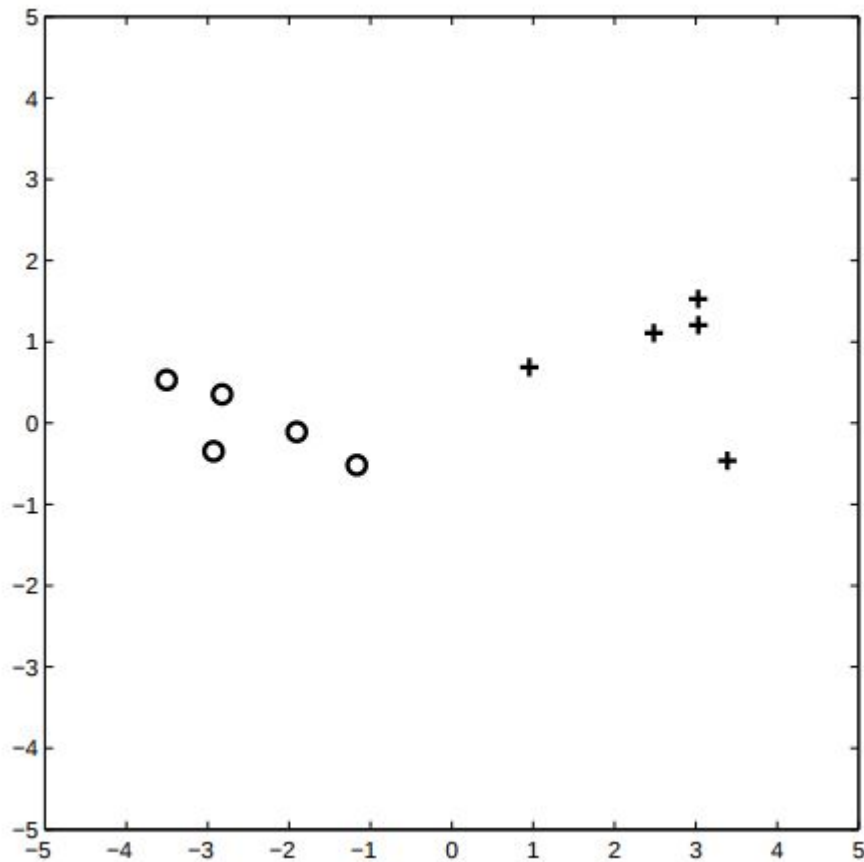
Modelos generativos

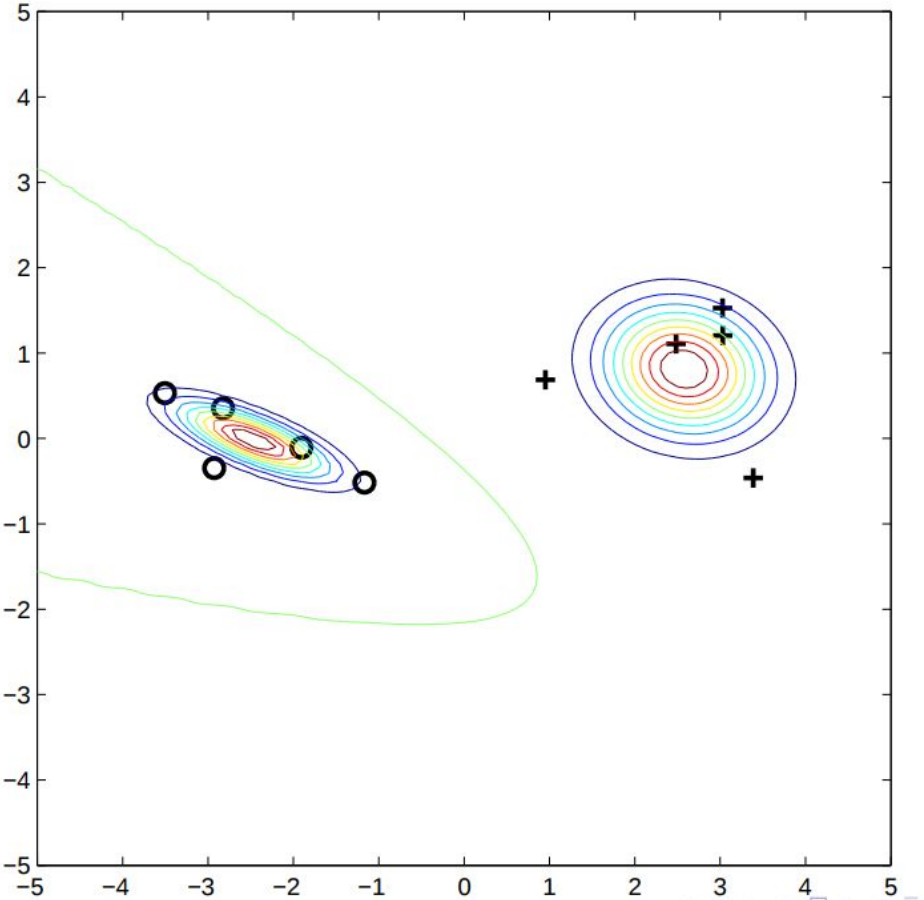
Modelos generativos

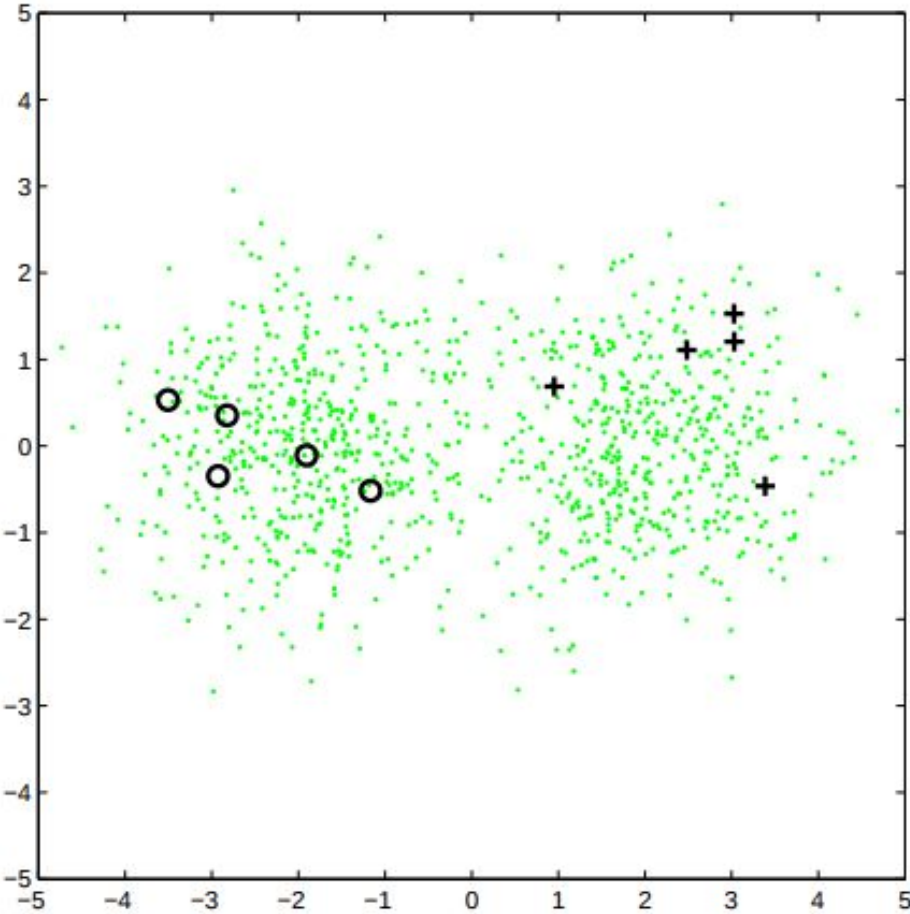
En el tutorial:

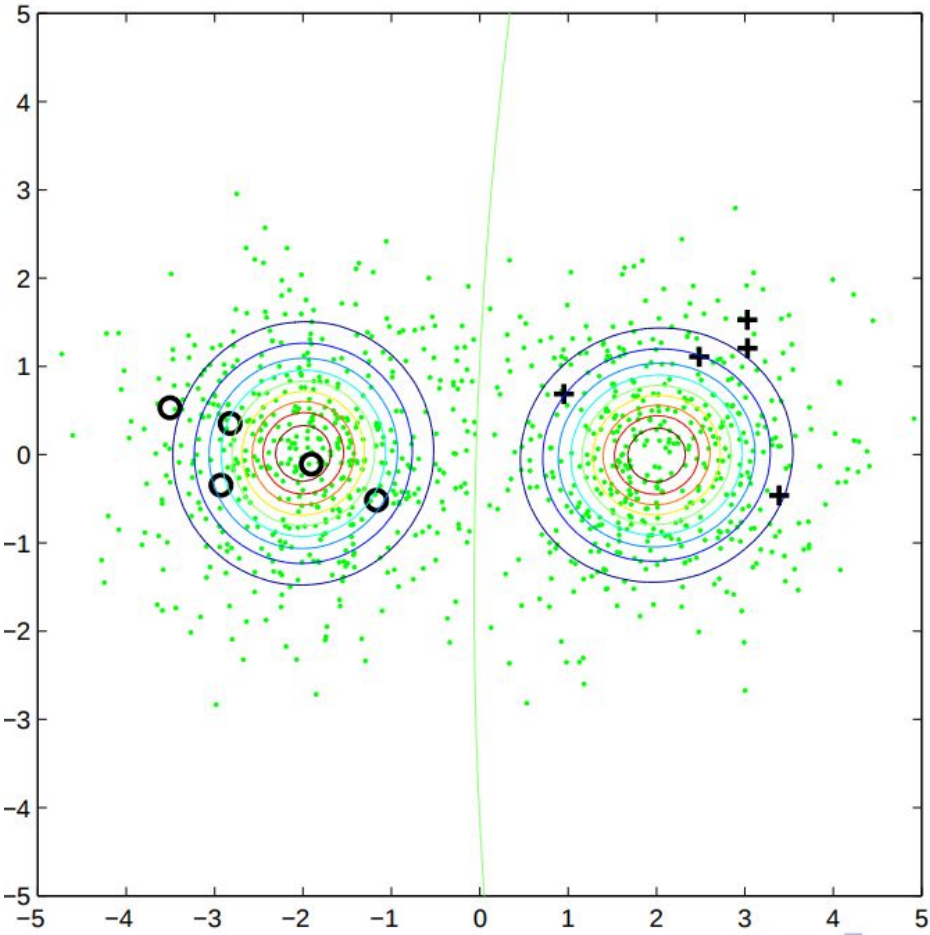
Modelos generativos con gaussianas

Usando Maximum Likelihood Estimation y Expectation Maximization

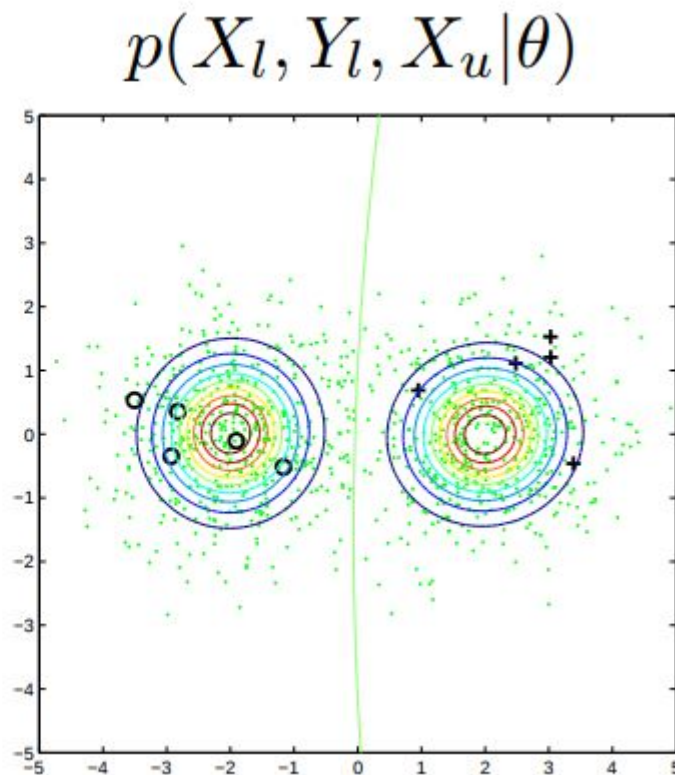
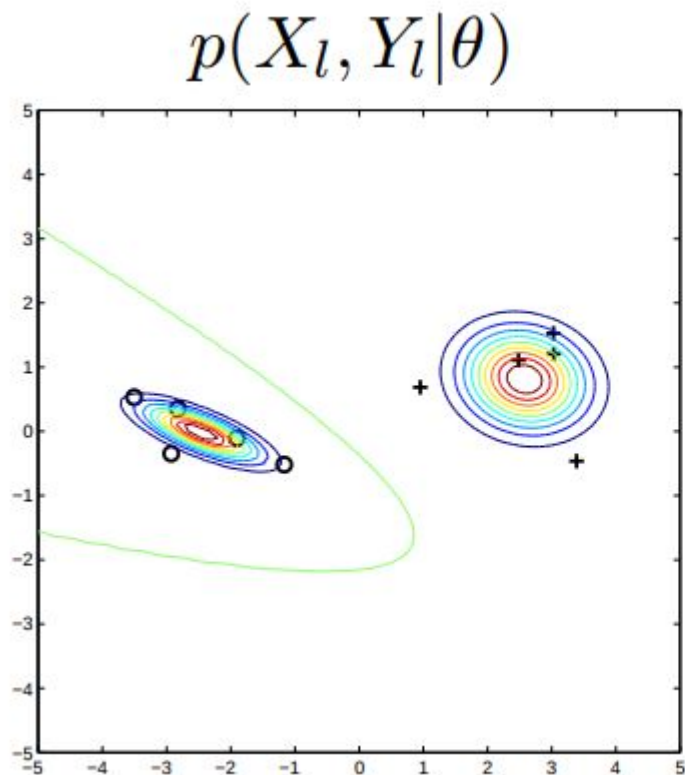








Maximizar diferentes parámetros



Cuánto podemos aprender?

No free lunch!

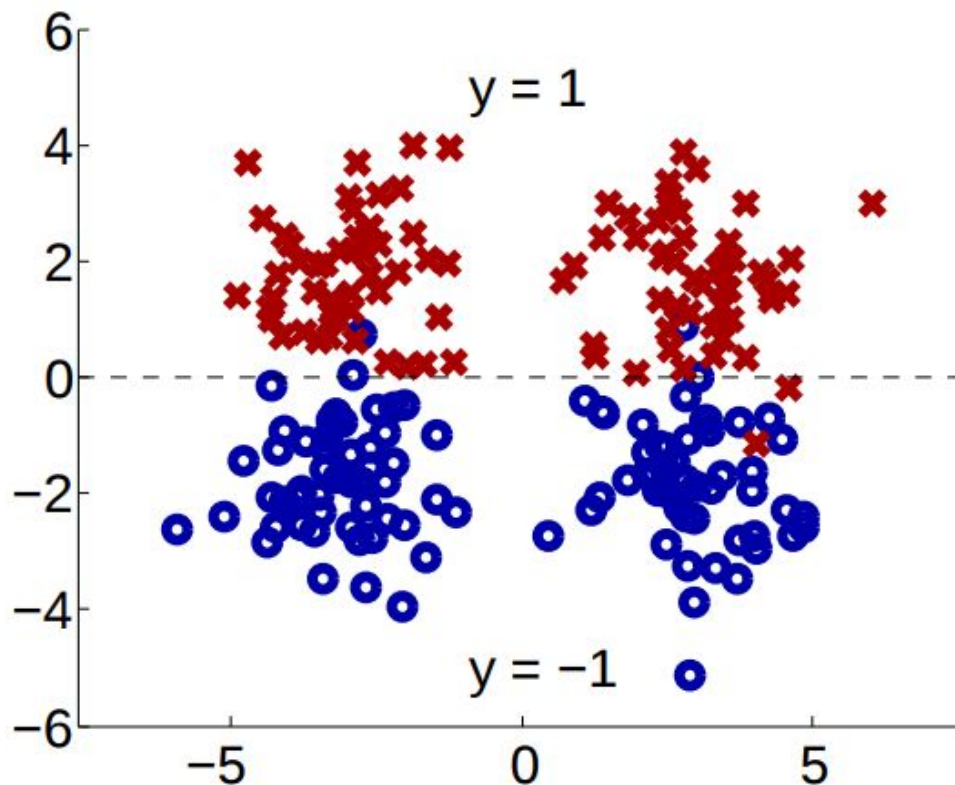
Si asumimos pocas cosas, ganamos poca información

Si asumimos muchas cosas, nos podemos equivocar

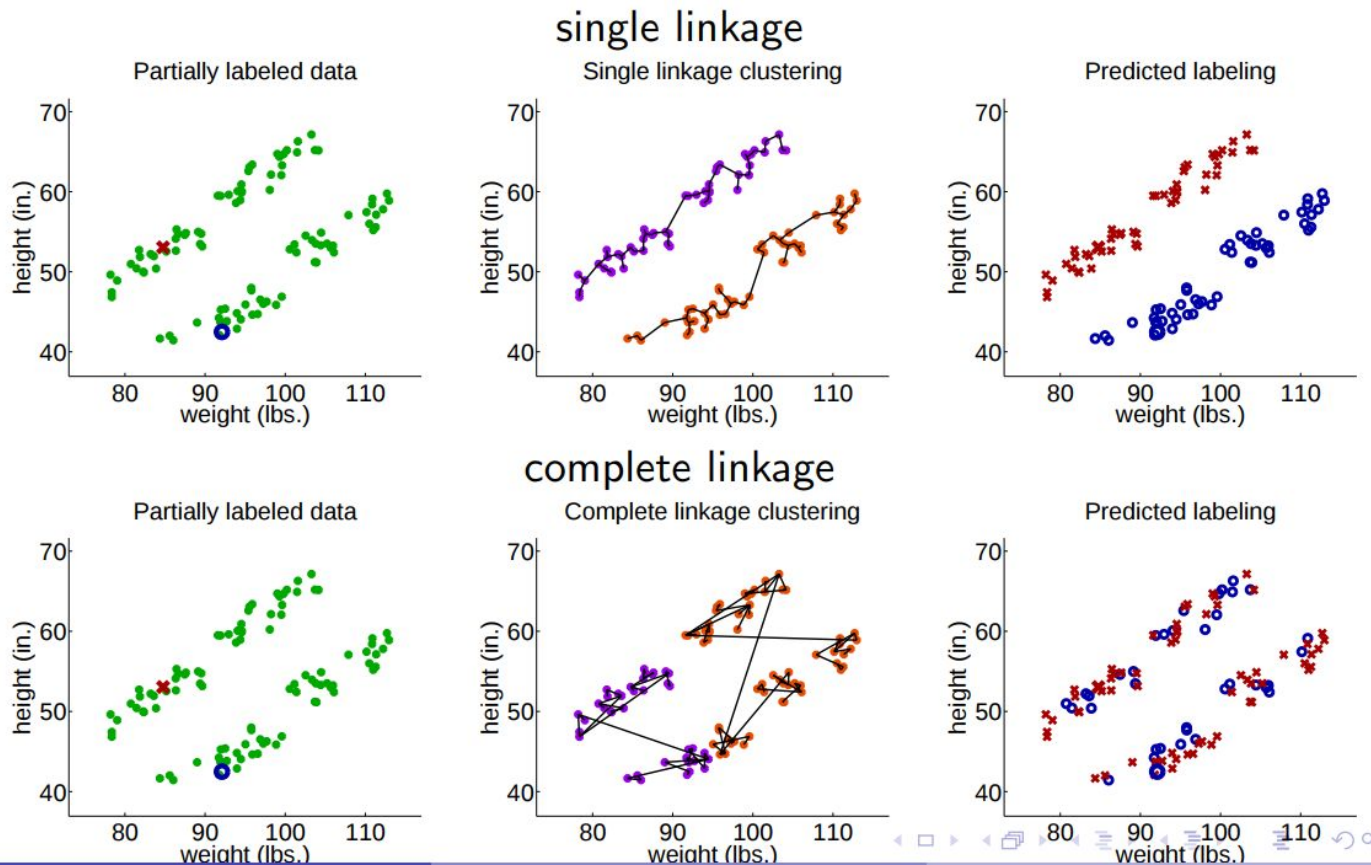
→ Mixtura de gaussianas

→ Modelos más complejos

Un modelo simple no lo captura bien



Relacionado: cluster-and-label



Valoración de modelos generativos

Ventajas:

- Buen fundamento matemático
- Se obtiene un modelo generativo

Desventajas:

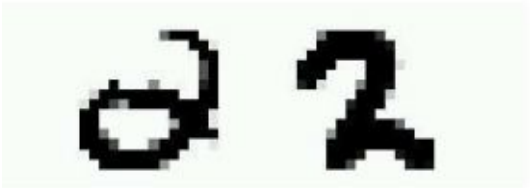
- Si la asunción está mal, el error es grande

Modelos basados en grafos

	<i>d₁</i>	<i>d₃</i>	<i>d₄</i>	<i>d₂</i>
asteroid	●	●		
bright	●	●		
comet		●		
year				
zodiac				
:				
:				
airport				
bike				
camp			●	
yellowstone			●	●
zion				●

	<i>d₁</i>	<i>d₃</i>	<i>d₄</i>	<i>d₂</i>
asteroid	•			
bright	•			
comet				
year				
zodiac		•		
.				
:				
.				
airport			•	
bike			•	
camp				
yellowstone				•
zion				•

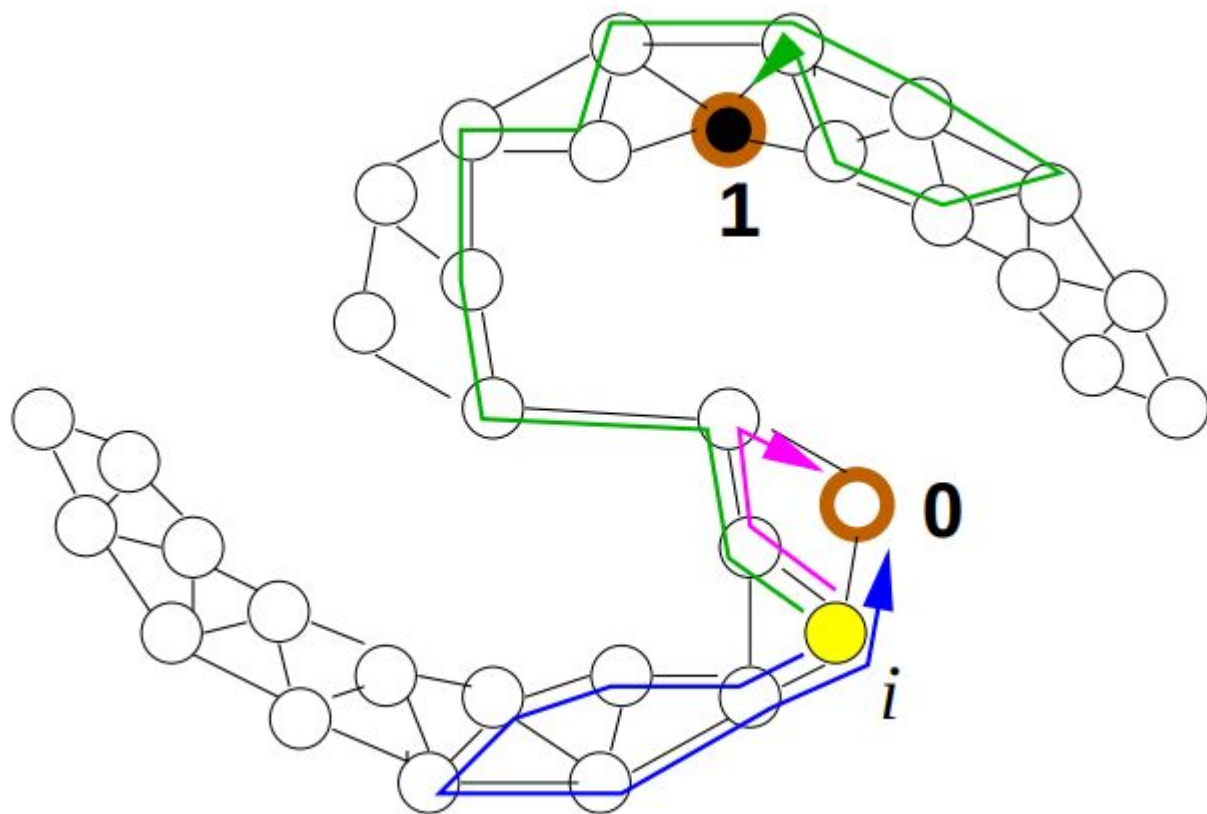
[illegible]



0 2



0 2 2 2 2 2



Otros algoritmos

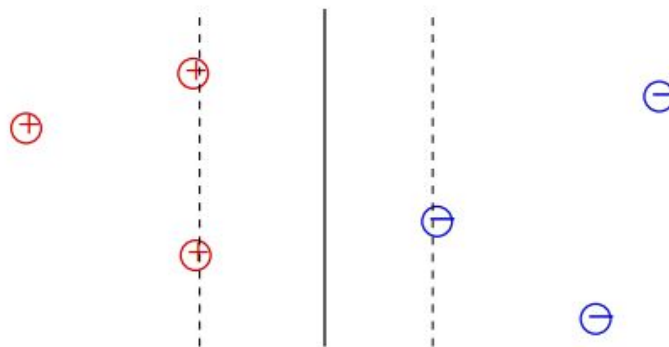
Otros algoritmos

- Multiview learning
- Manifold learning
- Semi-supervised Support Vector Machines
- Ladder Networks
- Positive Unlabelled

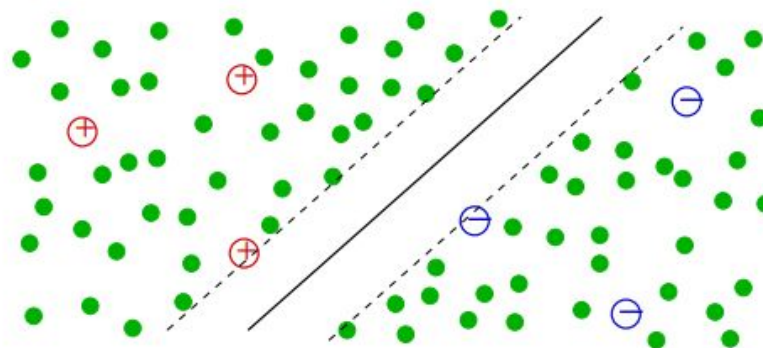
Aproximaciones disjuntas

- Usar embeddings como pre-proceso
- Usar clusters para generalizar

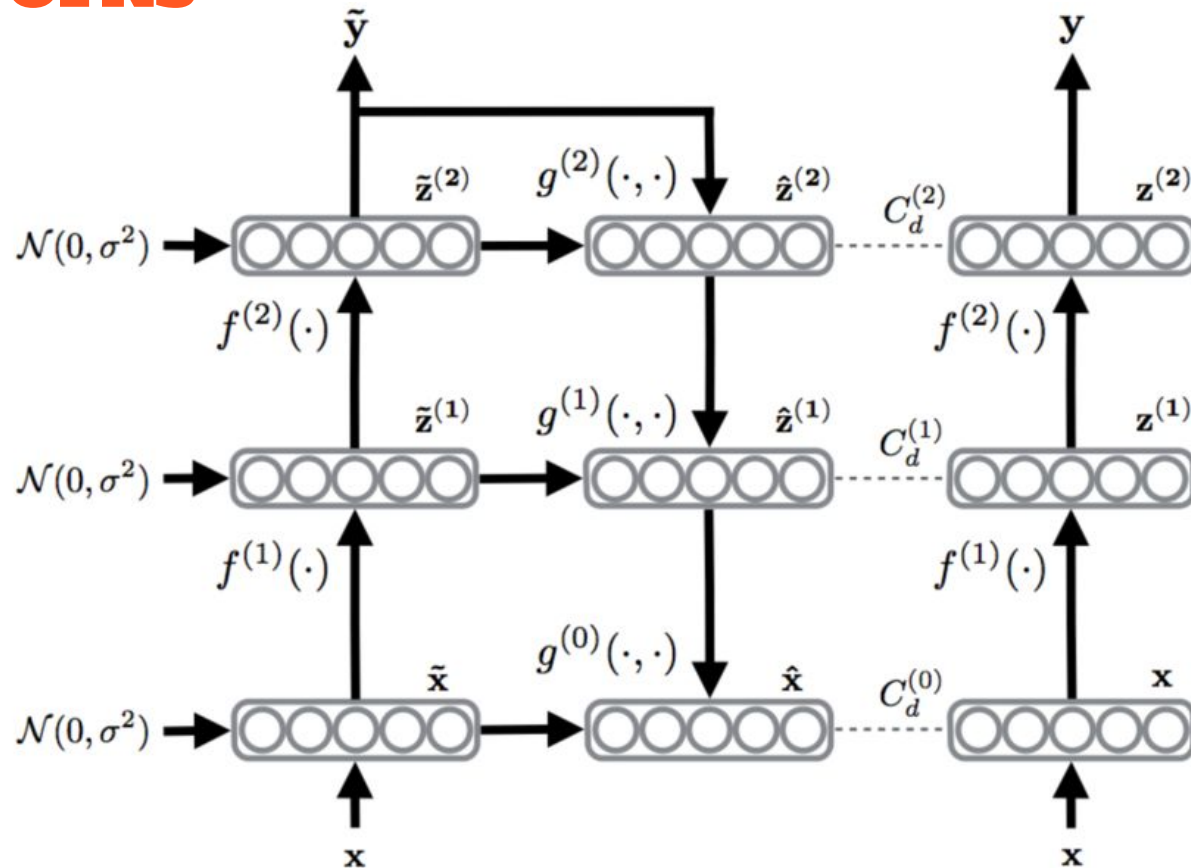
SVMs



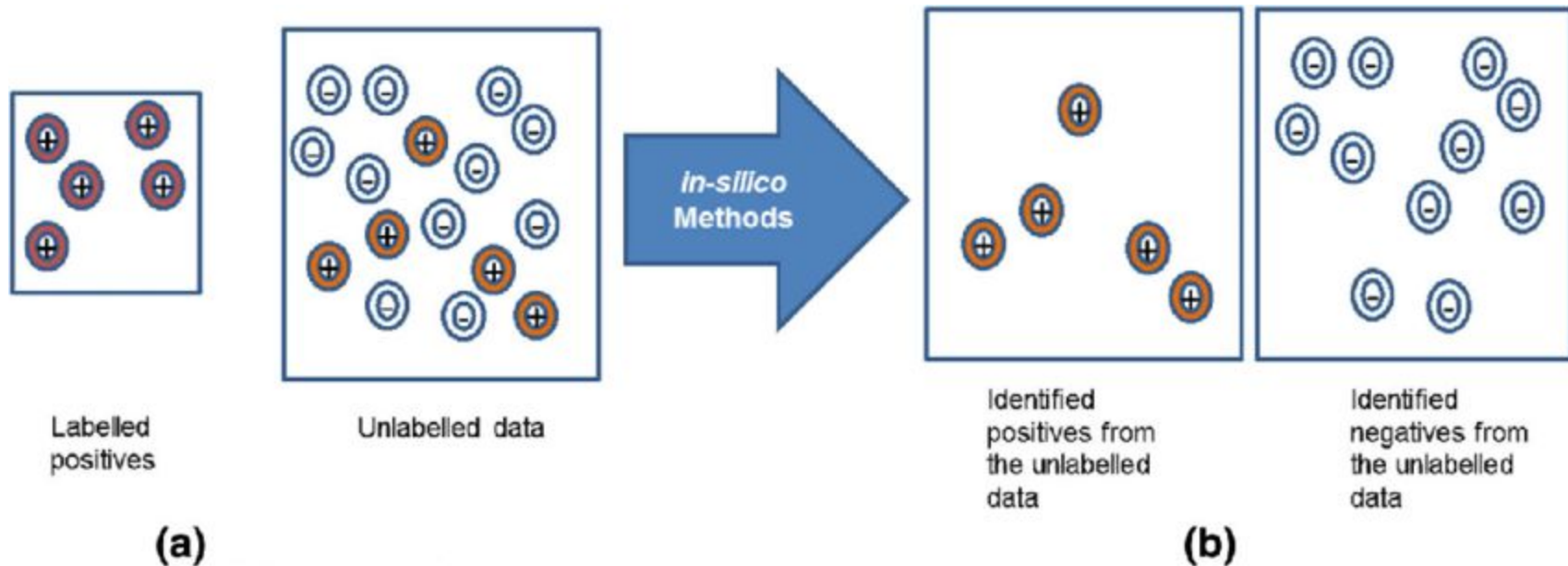
Semi-supervised SVMs (S3VMs) = Transductive SVMs (TSVMs)



Ladder Networks



Positive Unlabelled



¿Supervisado, semi-supervisado, no supervisado?

Aprendizaje activo

1. Obtener un conjunto pequeño de datos etiquetados
 2. Aprender un clasificador de los datos etiquetados
 3. Aplicar el clasificador sobre datos no etiquetados
 4. Seleccionar los ejemplos que, de tener etiqueta manual, maximizarían el rendimiento del clasificador
 5. Un oráculo (humano) etiqueta los ejemplos, y se incorporan a los datos etiquetados
 6. Volver a 2
- Qué ejemplos maximizan aprendizaje? Con mayor incertidumbre? Más representativos?
 - Combinar con self-learning

Aprendizaje por refuerzos

Alcanzar un objetivo lejano a través de pasos que no sabemos si son acertados

- ej.: videojuegos, armar un mueble, tratamiento de leucemia...

Cómo?

Aprendiendo una política que nos lleve hasta el objetivo a través de los pasos

→ aprender de los errores: asociar penalizaciones o recompensas a cada paso

- a diferencia de no supervisado, el objetivo está definido
- a diferencia de supervisado, no todos los eventos están asociados a una clase
- es una forma de semi-supervisado?

Recomendación

Es un problema supervisado, semi-supervisado, no supervisado?

los ejemplos iniciales influyen el comportamiento de los nuevos casos!

Tarea de pretexto

1. en datos no etiquetados, inventar una etiqueta presente en los datos
 2. entrenar un clasificador con estas etiquetas inventadas
- el clasificador que obtenemos nos provee un nuevo espacio
 - este espacio está configurado con otra perspectiva sobre los datos
 - la proyección a este espacio se puede integrar muy fácilmente en el preproceso de datos para aprendizaje supervisado o no supervisado
 - especialmente útil en redes neuronales

Tarea de pretexto para embeddings

1. entrenar un clasificador neuronal con una tarea de pretexto
2. quedarse con la penúltima capa del clasificador

Tarea de pretexto para embeddings

1. entrenar un clasificador neuronal con una tarea de pretexto
2. quedarse con la penúltima capa del clasificador

ej.: semántica de las palabras

tarea de pretexto

datos no etiquetados: *el gato come pescado*

datos etiquetados:

<i>_ come pescado</i>	-- etiqueta: el
<i>el _ come pescado</i>	-- etiqueta: gato
<i>el gato _ pescado</i>	-- etiqueta: come
<i>el gato come _</i>	-- etiqueta: pescado

Tarea de pretexto para embeddings

1. entrenar un clasificador neuronal con una tarea de pretexto
2. quedarse con la penúltima capa del clasificador

ej.: semántica de las palabras

tarea de pretexto

clasificador

dado un contexto, predecir la palabra

Tarea de pretexto para embeddings

1. entrenar un clasificador neuronal con una tarea de pretexto
2. quedarse con la penúltima capa del clasificador

ej.: semántica de las palabras

tarea de pretexto

clasificador

dada la palabra, predecir el contexto

Tarea de pretexto para embeddings

1. entrenar un clasificador neuronal con una tarea de pretexto
2. quedarse con la penúltima capa del clasificador

ej.: semántica de las palabras

tarea de pretexto

clasificador

embedding

la penúltima capa del clasificador

Tarea de pretexto para embeddings

1. entrenar un clasificador neuronal con una tarea de pretexto
2. quedarse con la penúltima capa del clasificador

ej.: semántica de las palabras

tarea de pretexto

clasificador

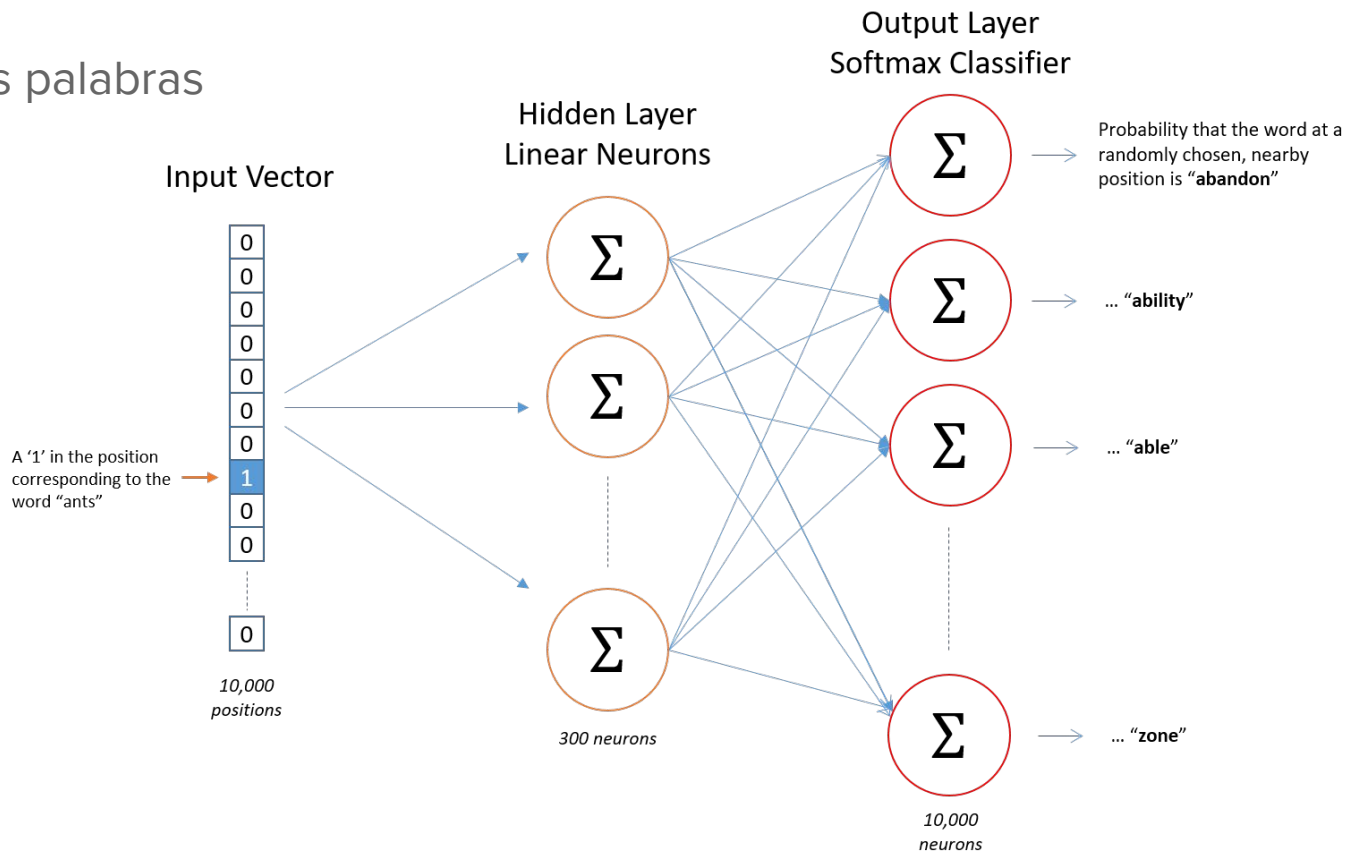
embedding

integración

para los nuevos ejemplos, los paso por el clasificador y los llevo hasta la penúltima capa. Su representación vectorial ya no es la del vector inicial (muy dimensional y ralo) sino la del vector de la penúltima capa (poco dimensional y denso)

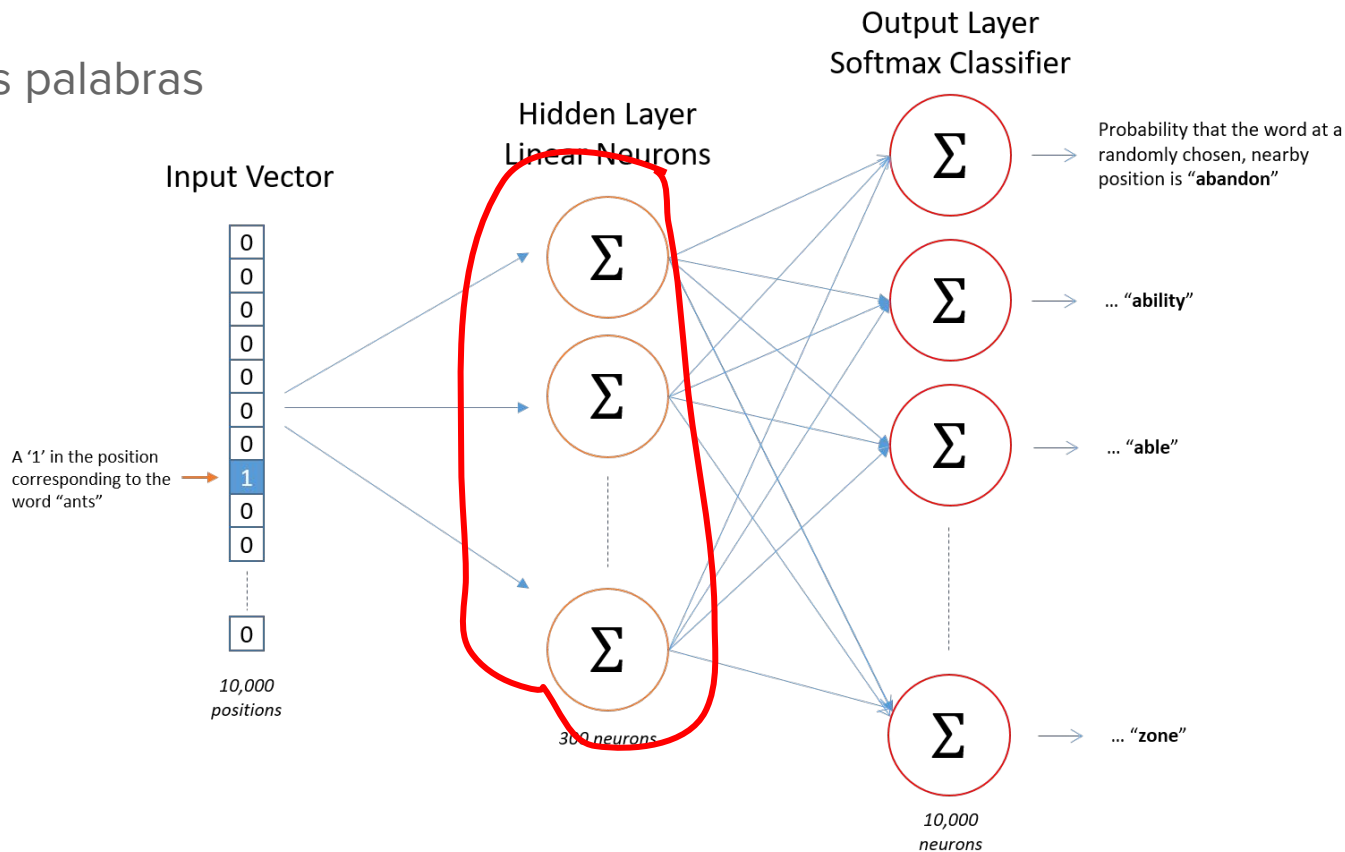
Tarea de pretexto para embeddings

semántica de las palabras



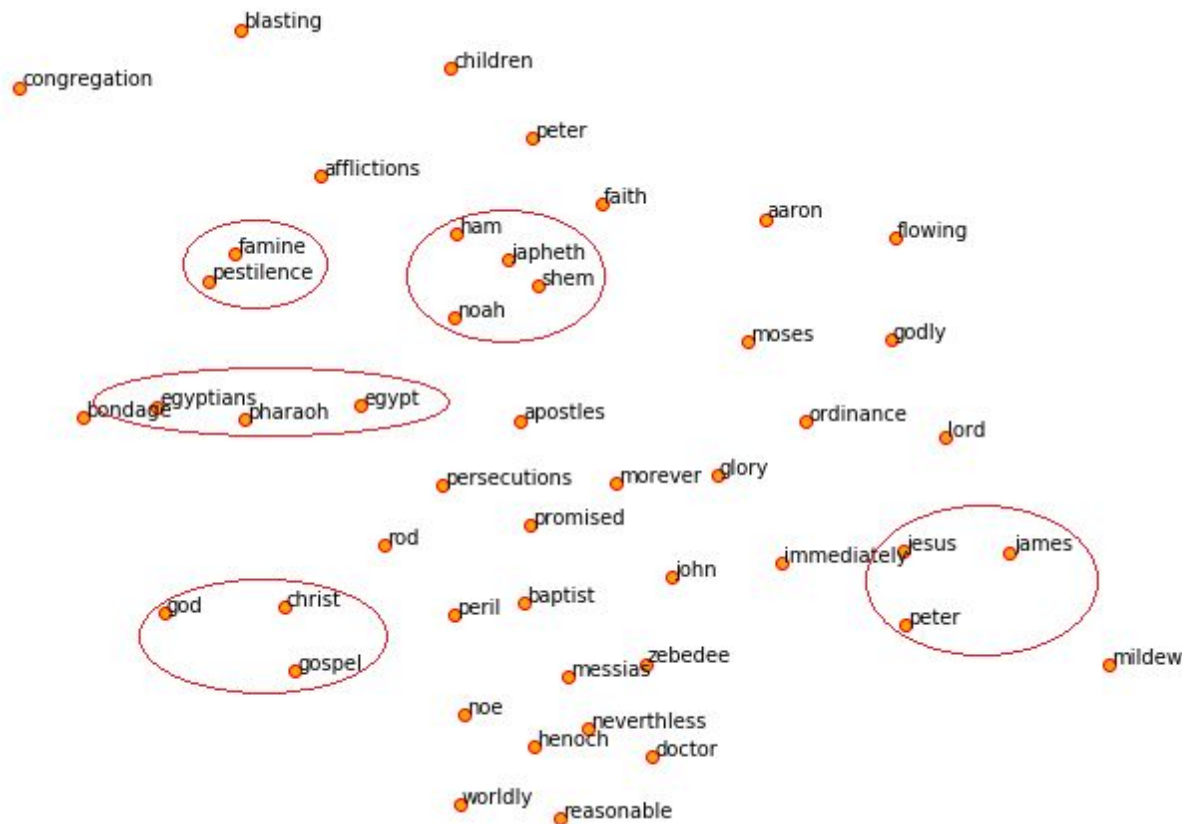
Tarea de pretexto para embeddings

semántica de las palabras



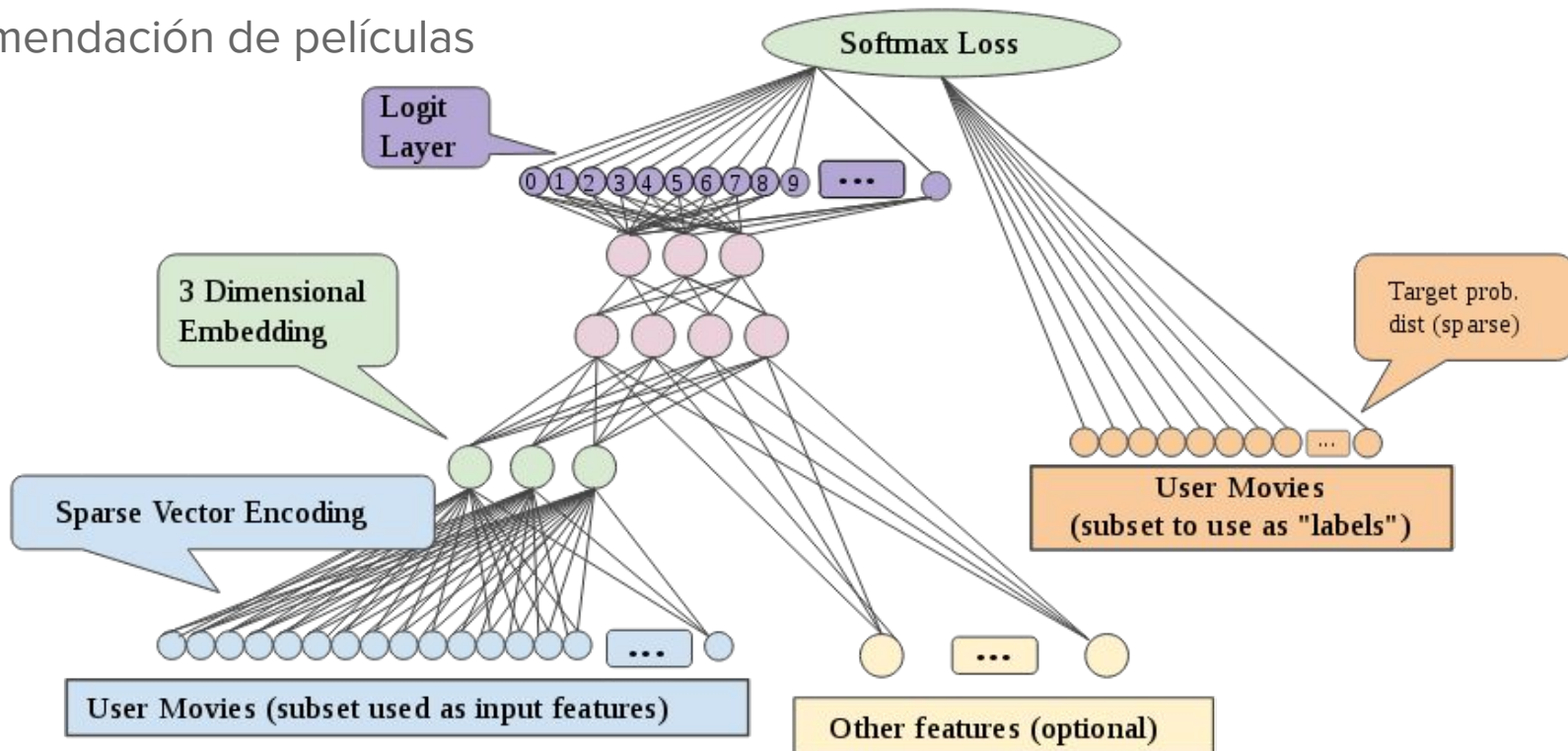
Tarea de pretexto para embeddings

semántica de las palabras



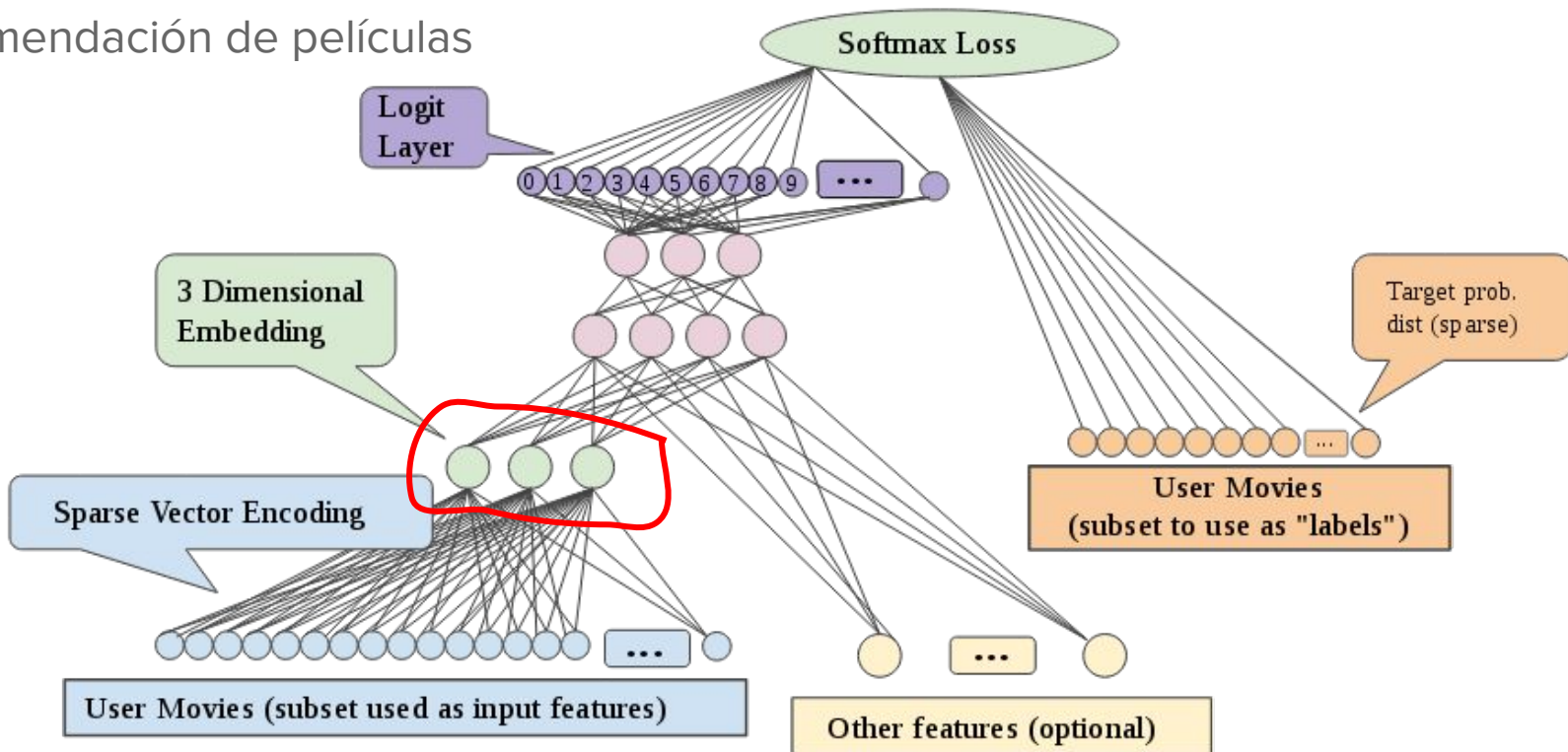
Tarea de pretexto para embeddings

recomendación de películas



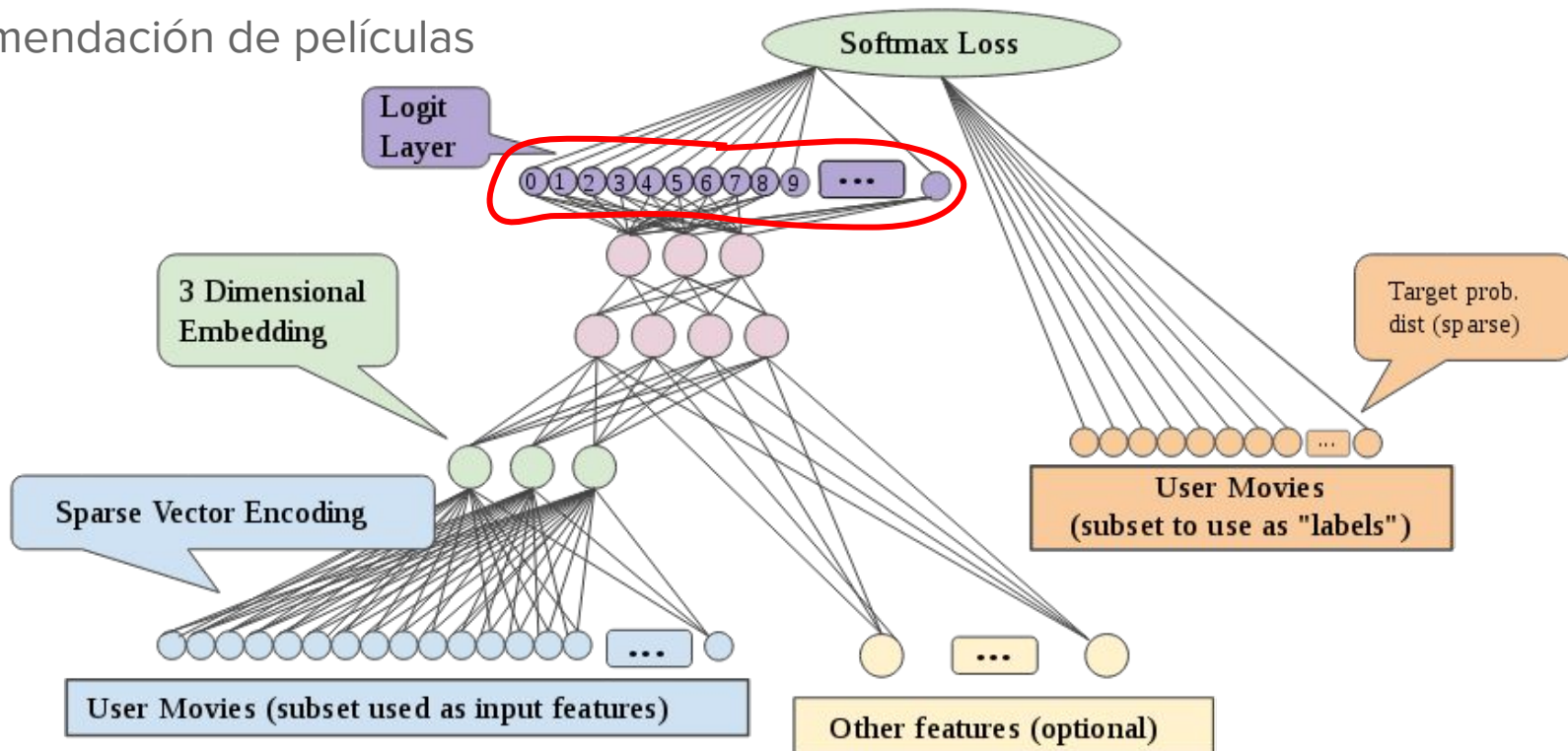
Tarea de pretexto para embeddings

recomendación de películas



Tarea de pretexto para embeddings

recomendación de películas



Transfer learning

Contexto de aplicación:

- Tenemos datos etiquetados del Dominio A
 - Tenemos pocos o ningún dato etiquetado del Dominio B
 - El Dominio A y el Dominio B tienen algunos puntos en común
1. Aprender un modelo en el Dominio A
 2. Usar ese modelo en el Dominio B
 - posiblemente, reentrenar (fine-tuning) el modelo del Dominio A con algunos ejemplos del Dominio B

Supervisado → No supervisado

Usar datos etiquetados para mejorar algoritmos no supervisados

- Clustering with rules
- Constrained Clustering
- Reglas de asociación con clase
- K-nn con etiquetas de usuarios, etiquetas de items
- Etiquetas sobre los datos

Evaluación

Evaluación

Reservar parte de los datos para evaluación (test)

→ tenemos pocos datos!

- reservar datos para evaluación es costoso
- la evaluación es todavía más anecdótica

Qué podemos hacer?

- cross-validation
- monitoreo manual de los datos nuevos
- graficar cómo evoluciona la distribución de población alrededor de los testigos