Sesión Nº 1

Aprendizaje No Supervisado

(Traducido de presentaciones de Sriram Sankararaman y Junming Yin)

Plan de la unidad

- **Introducción**
  - **Aprendizaje no supervisado**
  - **¿Qué es el análisis de agrupamientos o clustering?**
  - Aplicaciones de clustering
- Similaridades y Distancias entre datos
- Algoritmos de agrupamiento
  - K-means
  - Modelo de mezcla de normales (GMM)
  - Agrupamiento Jerárquico
- Mapas auto-organizados (redes de Kohonen)

Aprendizaje no supervisado

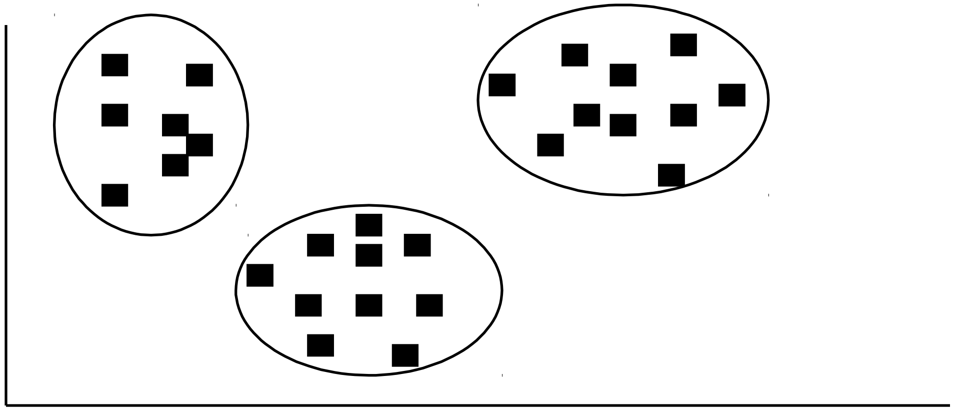
- Recordemos que en el contexto del aprendizaje supervisado (clasificación y regresión), los datos de entrenamiento se representan por $(x_i, y_i)_{i=1\ldots n}$ y el objetivo es “aprender” una función para predecir y dado x
- En el aprendizaje no supervisado, solo se tienen los datos $(x_i)_{i=1\ldots n}$, sin etiquetas. En este caso la idea es inferir algunas propiedades de la distribución de X.

¿Por qué aprendizaje no supervisado?

- Es más fácil contar con datos sin clasificar. Etiquetar datos puede ser costoso.
- Cuando los datos están en un espacio de alta dimensionalidad, este tipo de métodos permite encontrar características de los datos en espacios de menor dimensión que pueden ser suficientes para describir una muestra.
- En los primeros pasos de una investigación, es muy importante desarrollar un análisis exploratorio de los datos. De esta forma se obtiene una idea de la naturaleza y/o estructura de los mismos.
- El análisis de agrupamientos (clustering) o segmentación es un método de aprendizaje no supervisado

¿ Qué es el análisis de agrupamientos?

- El objetivo es descubrir grupos en los datos observados, tales que los datos dentro de un mismo grupo son mas parecidos que aquellos en los otros grupos.
- Se requiere:
  - Una función de similaridad (distancia) entre datos
  - Una función de pérdida para evaluar el agrupamiento efectuado
  - Un algoritmo para optimizar la función de pérdida



Plan de la unidad

- Introducción
  - Aprendizaje no supervisado
  - ¿Qué es el análisis de agrupamientos o clustering?
  - **Aplicaciones de clustering**
- Similaridades y Distancias entre datos
- Algoritmos de agrupamiento
  - K-means
  - Modelo de mezcla de normales (GMM)
  - Agrupamiento Jerárquico
- Mapas auto-organizados (redes de Kohonen)

Segmentación de imágenes

http://people.cs.uchicago.edu/~pff/segment/

Resultados de una búsqueda de la palabra cluster

EisenLab
Commercial use of the ScanAlyze, **Cluster** and/or TreeView executable and/or ... **Cluster** and TreeView are an integrated pair of programs for analyzing and ...
rana.lbl.gov/EisenSoftware.htm - 11k - Cached - Similar pages

Book results for **cluster**
The Linux Enterprise Cluster : build a highly ... - by Karl Kopper - 466 pages
Messier's Nebulae and Star Clusters - by Kenneth Glyn Jones - 456 pages

Searches related to: **cluster**

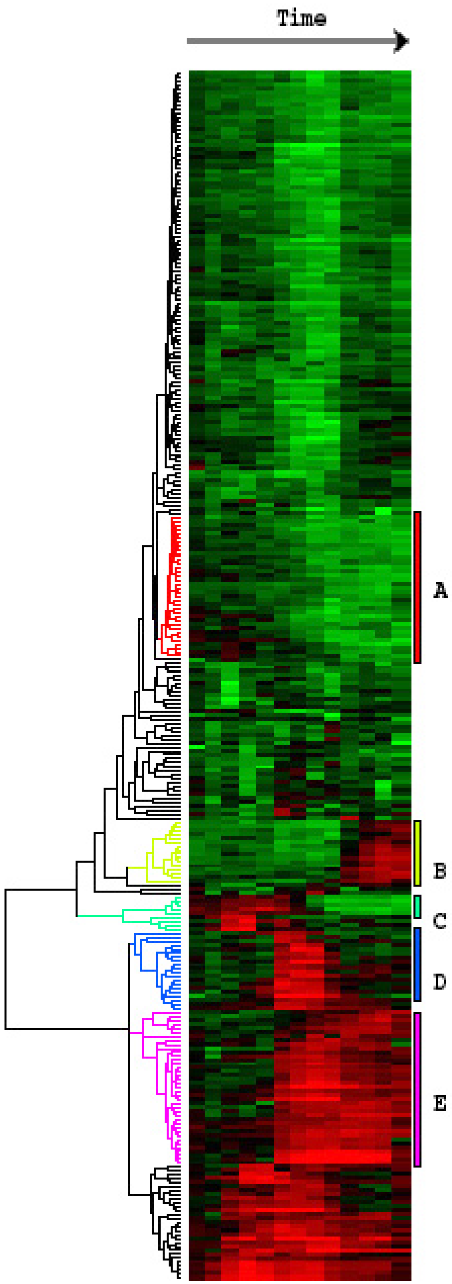
cluster **headache**	cluster **analysis**	**server** cluster	cluster **sampling**
cluster **windows 2003**	**sql** cluster	**oracle** cluster	**clusty**

Gooooooooogle
1 2 3 4 5 6 7 8 9 10 **Next**

cluster Search

Search within results | Language Tools | Search Tips | Dissatisfied? Help us improve |

Clustering de datos de expresión de genes



Eisen et al, PNAS 1998

Algoritmo de compresión de imágenes (Vector quantization)

Bishop, PRML

Plan de la unidad

- Introducción
  - Aprendizaje no supervisado
  - ¿Qué es el análisis de agrupamientos o clustering?
  - Aplicaciones de clustering
- **Similaridades y Distancias entre datos**
- Algoritmos de agrupamiento
  - K-means
  - Modelo de mezcla de normales (GMM)
  - Agrupamiento Jerárquico
- Mapas auto-organizados (redes de Kohonen)

Distancias entre datos

¿Cómo medir la distancia o disimilaridad entre datos?

- Los resultados del agrupamiento depende de la elección de la distancia o medida de la disimilaridad.
- Usualmente depende del contexto de los datos en estudio.
- Depende de la naturaleza de las características observadas: Cuantitativas, ordinales, categóricas.

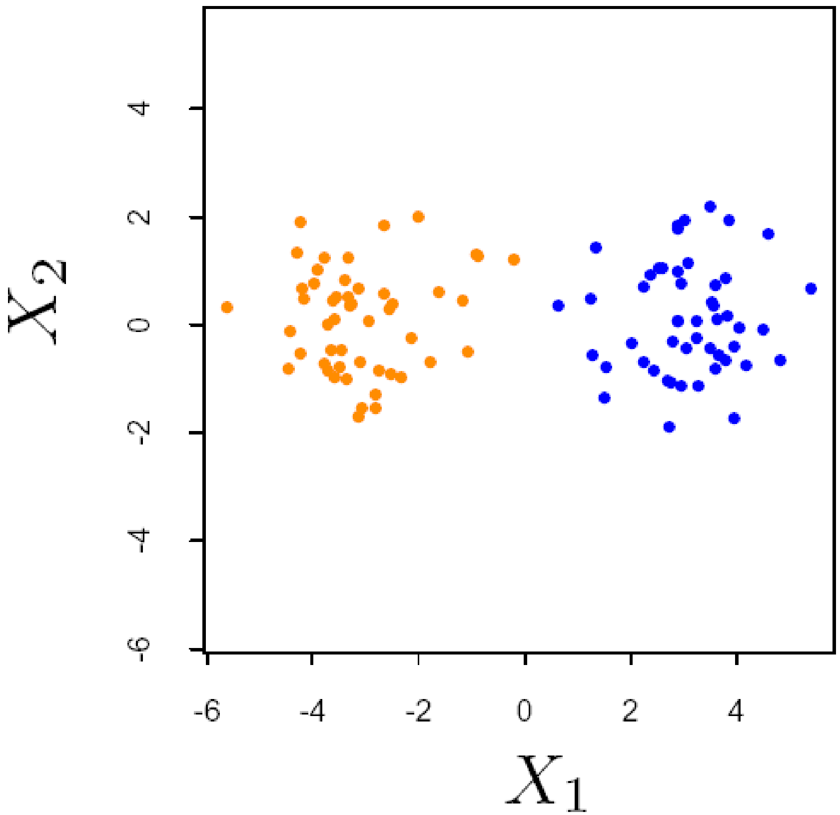
Disimilaridad en base a características

- Habitualmente los datos x_i contienen valores de ciertas características x_{ij}, $j = 1, \ldots, p.$
- Una elección común de disimilaridad es la distancia Euclidiana

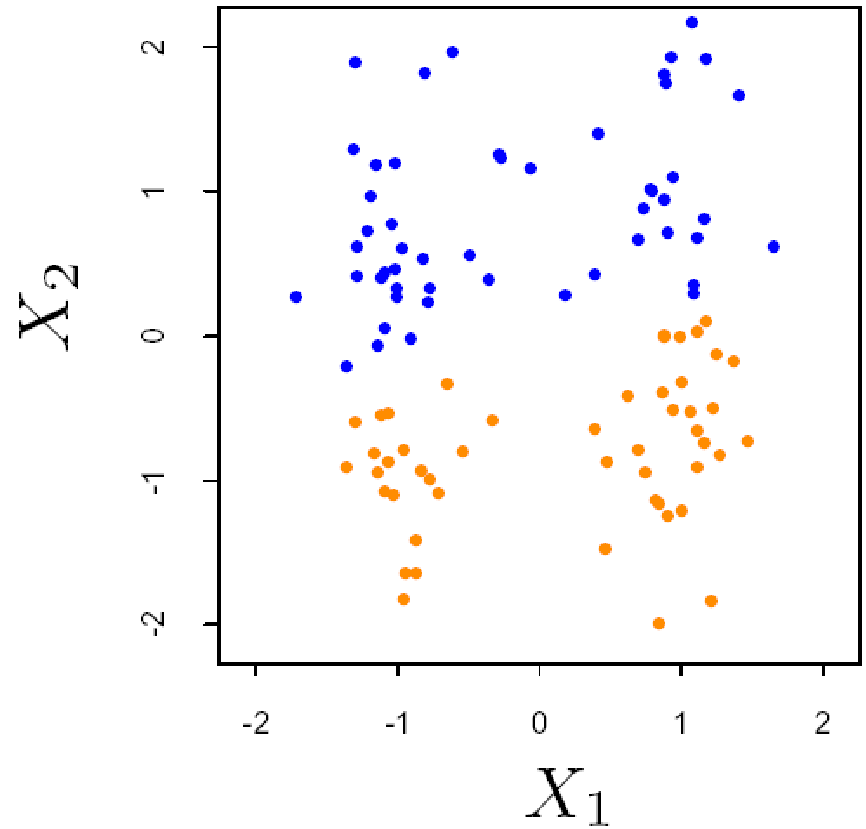
$$D(x_i, x_{i'}) = ||x_i - x_{i'}|| = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}$$

- Los agrupamientos definidos a partir de la distancia Euclidiana son invariantes a traslaciones y rotaciones del espacio de características. No así al escalamiento de los datos.
- Una manera de estandarizar los datos es realizar una traslación y escalamiento de manera que todos tengan media 0 y varianza 1.

La estandarización no siempre es útil!!



Datos simulados, 2-means sin estandarización



Datos simulados, 2-means con estandarización

Plan de la unidad

- Introducción
  - Aprendizaje no supervisado
  - ¿Qué es el análisis de agrupamientos o clustering?
  - Aplicaciones de clustering
- Similaridades y Distancias entre datos
- Algoritmos de agrupamiento
  - **K-means**
  - Modelo de mezcla de normales (GMM)
  - Agrupamiento Jerárquico
- Mapas auto-organizados (redes de Kohonen)

K-means: Idea

- Representar el conjunto de datos en términos de K grupos, cada uno resumido por un prototipo μ_k
- Cada dato es asignado a uno de los K grupos
  - Representado por las responsabilidades $r_{ik} \in \{0, 1\}$ tales $\sum_{k=1}^{K} r_{ik} = 1$ para cada índice de datos i
  - Ejemplo: 4 datos y 3 grupos: $(r_{ik}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

K-means: Idea

- Función de pérdida: La suma de los cuadrados de las distancias desde un dato hasta su prototipo asignado, conocido como la dispersión dentro de los grupos:

dato

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \|x_i - \mu_k\|^2$$

Responsabilidades

prototipos

Minimizando la función de pérdida

- Problema inicial:
  - Si los prototipos son conocidos se pueden asignar las responsabilidades.
  - Si se conocen las responsabilidades, se pueden calcular los prototipos óptimos.
- Se minimiza la función de pérdida con un algoritmo iterativo.
- Otra manera de minimizar la función de pérdida es mediante un enfoque de mezcla y separación.

Minimizando la función de pérdida

Paso E: Fija valores para μ_k y minimiza J con resp. a r_{ik}

- Asigna cada dato al prototipo más cercano

Paso M: Fija valores para r_{ik} y minimiza J con resp. a μ_k

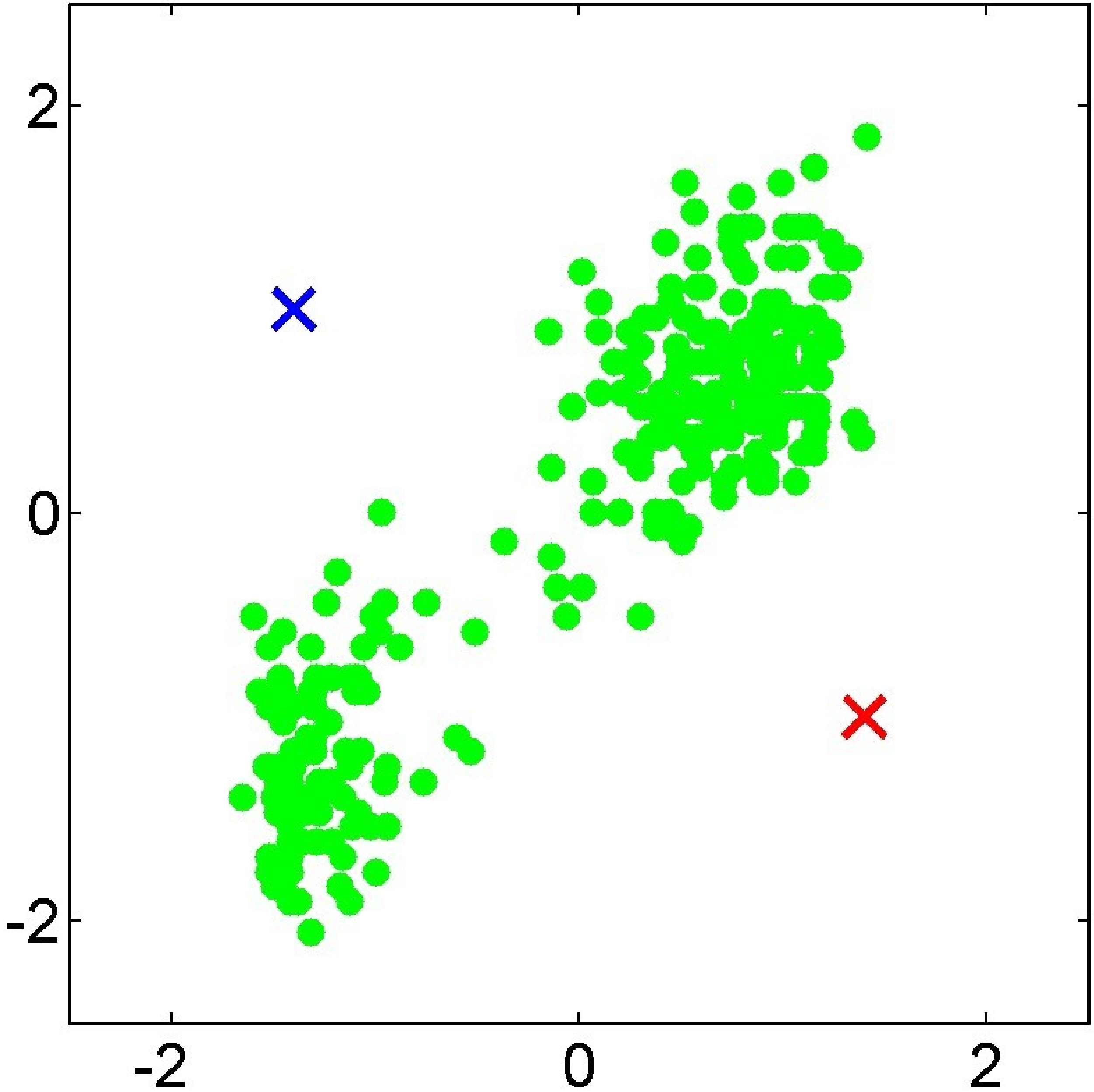
- Esto resulta en:

$$\mu_k = \frac{\sum_i r_{ik} x_i}{\sum_i r_{ik}}$$

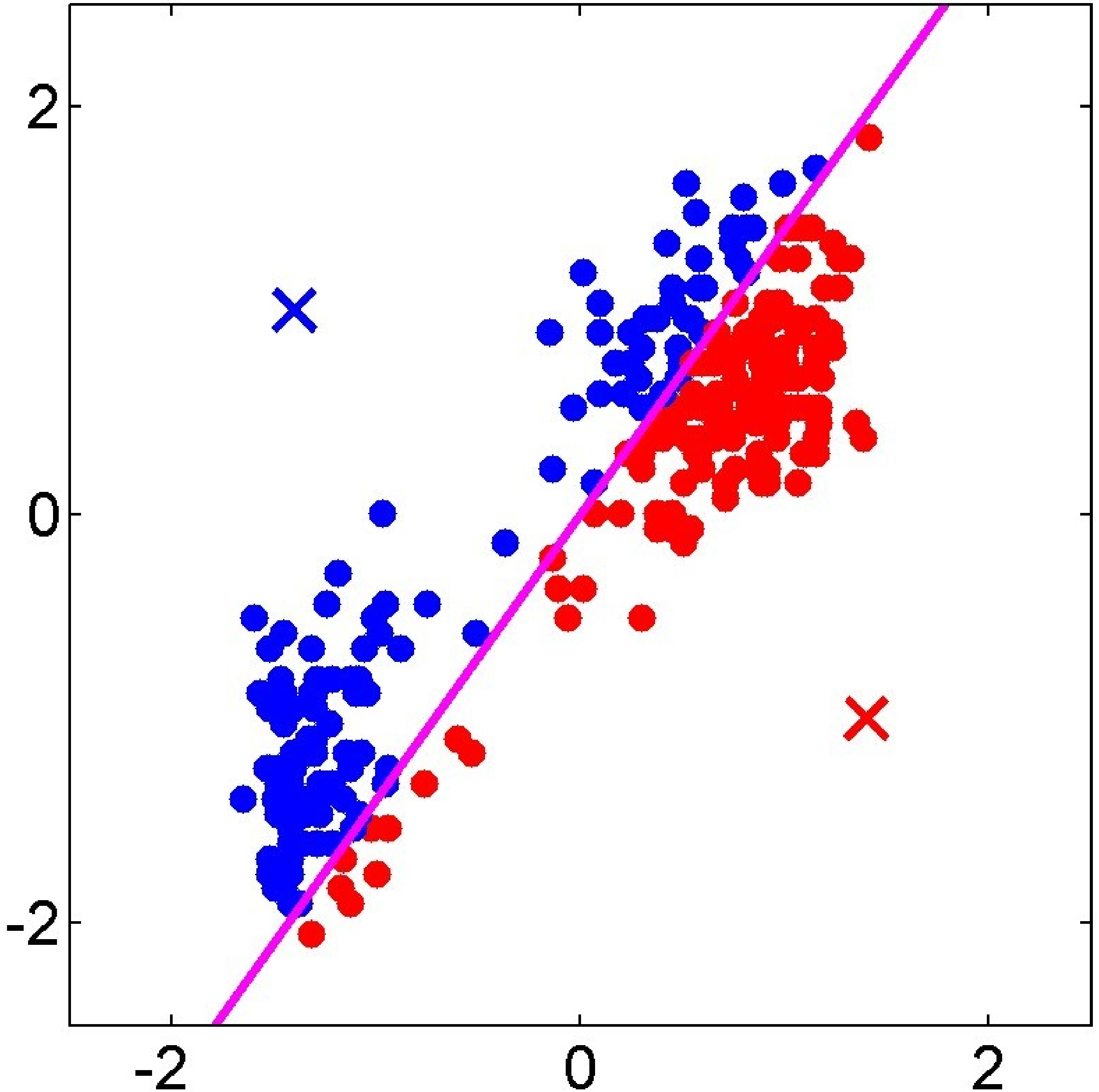
- Cada prototipo μ_k se define como la media de los puntos en el grupo.

La convergencia está garantizada ya que hay un número finito de definiciones de las responsabilidades.

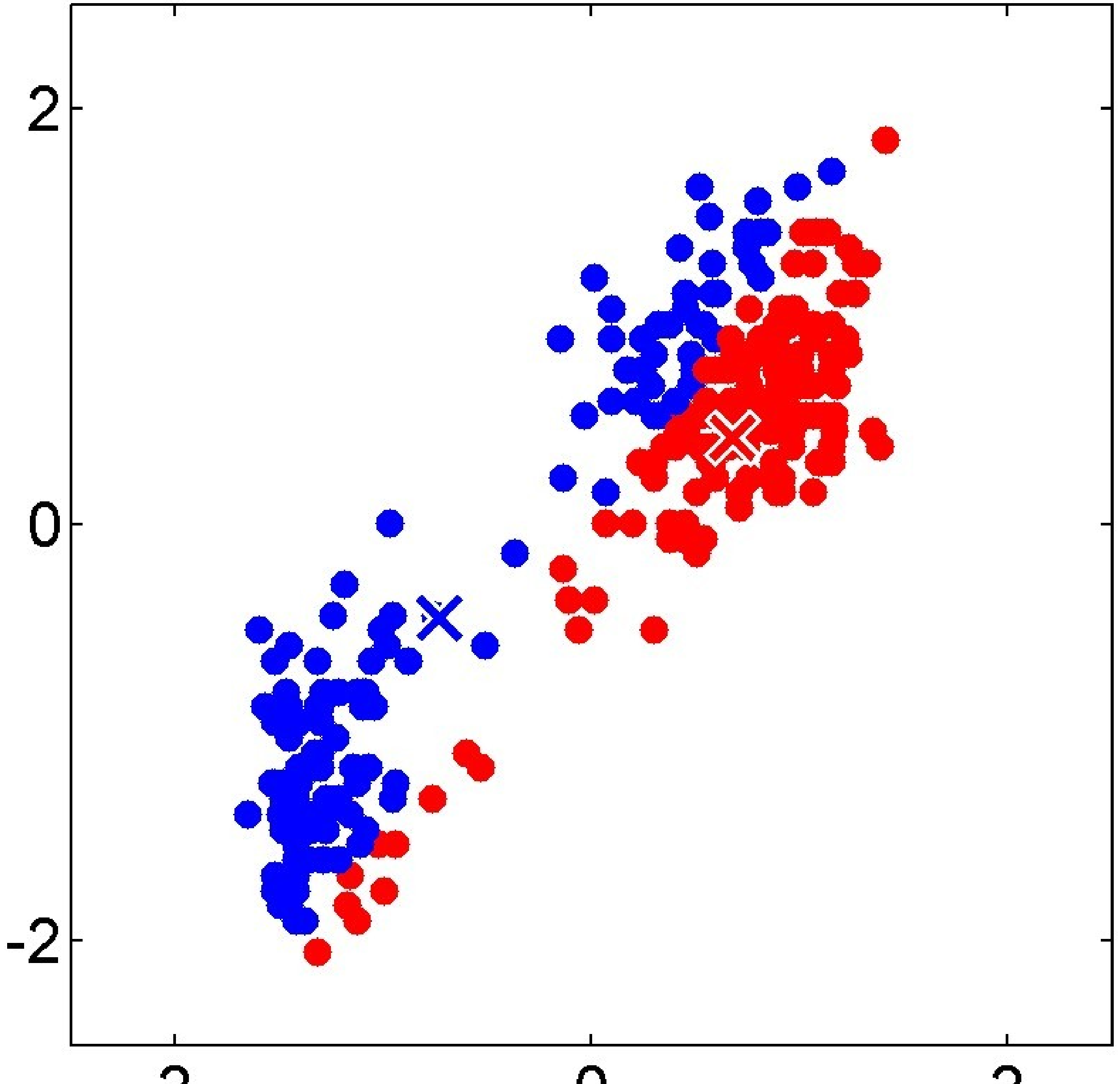
Encuentra mínimos locales, por lo que se debiera correr el algoritmo con distintos valores iniciales.



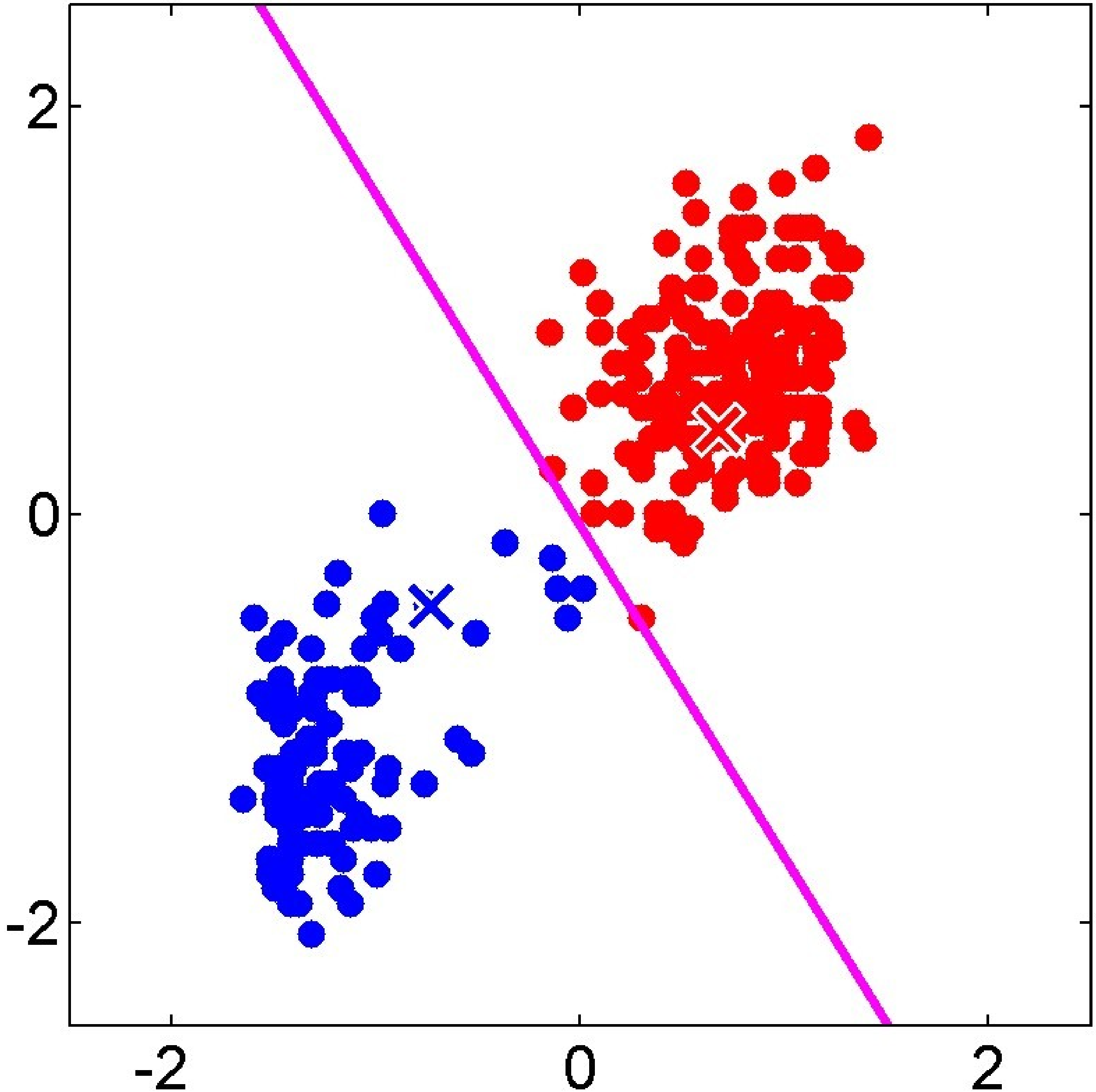
2
0
-2
-2
0
2



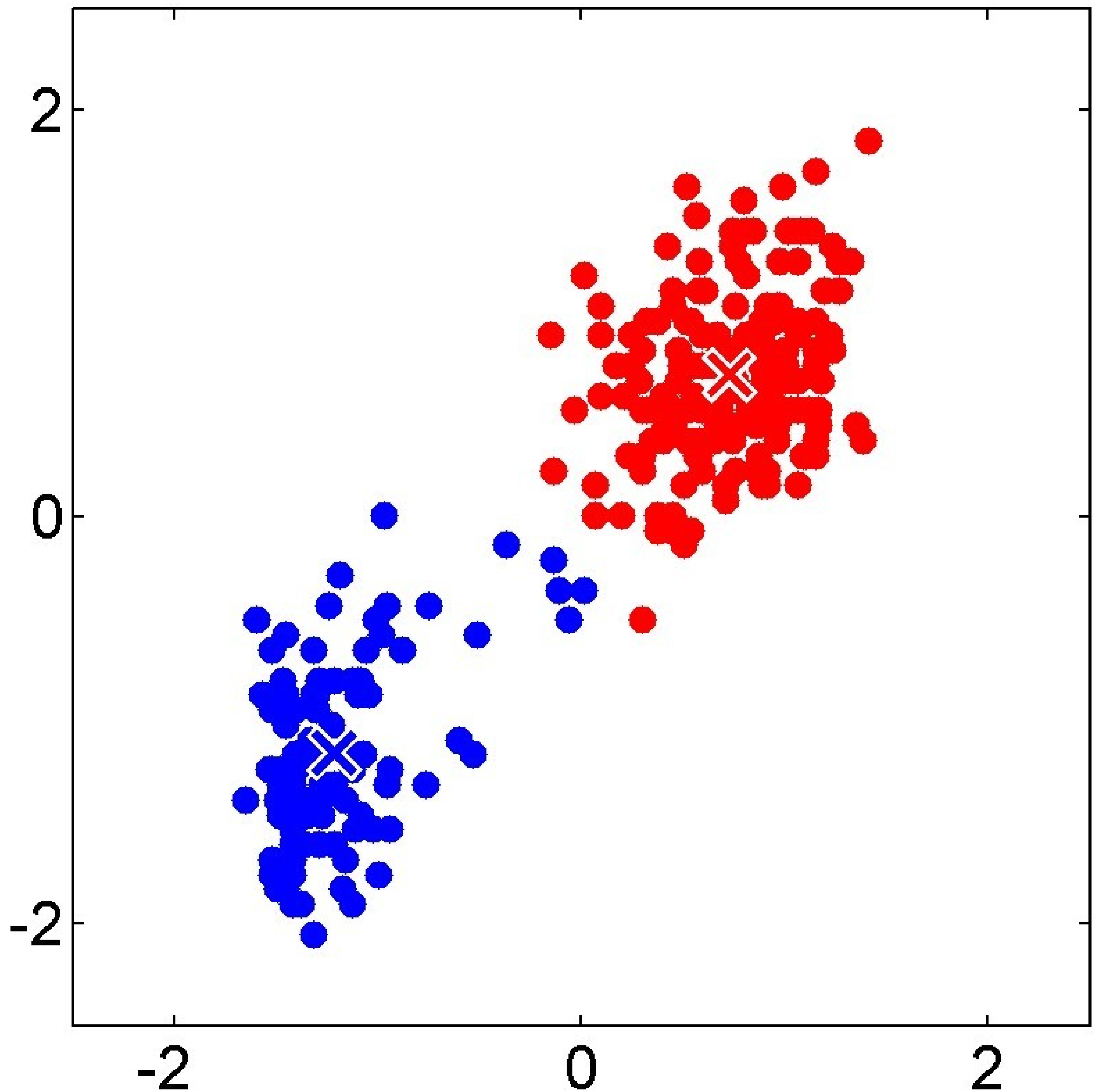
2
0
-2
-2
0
2

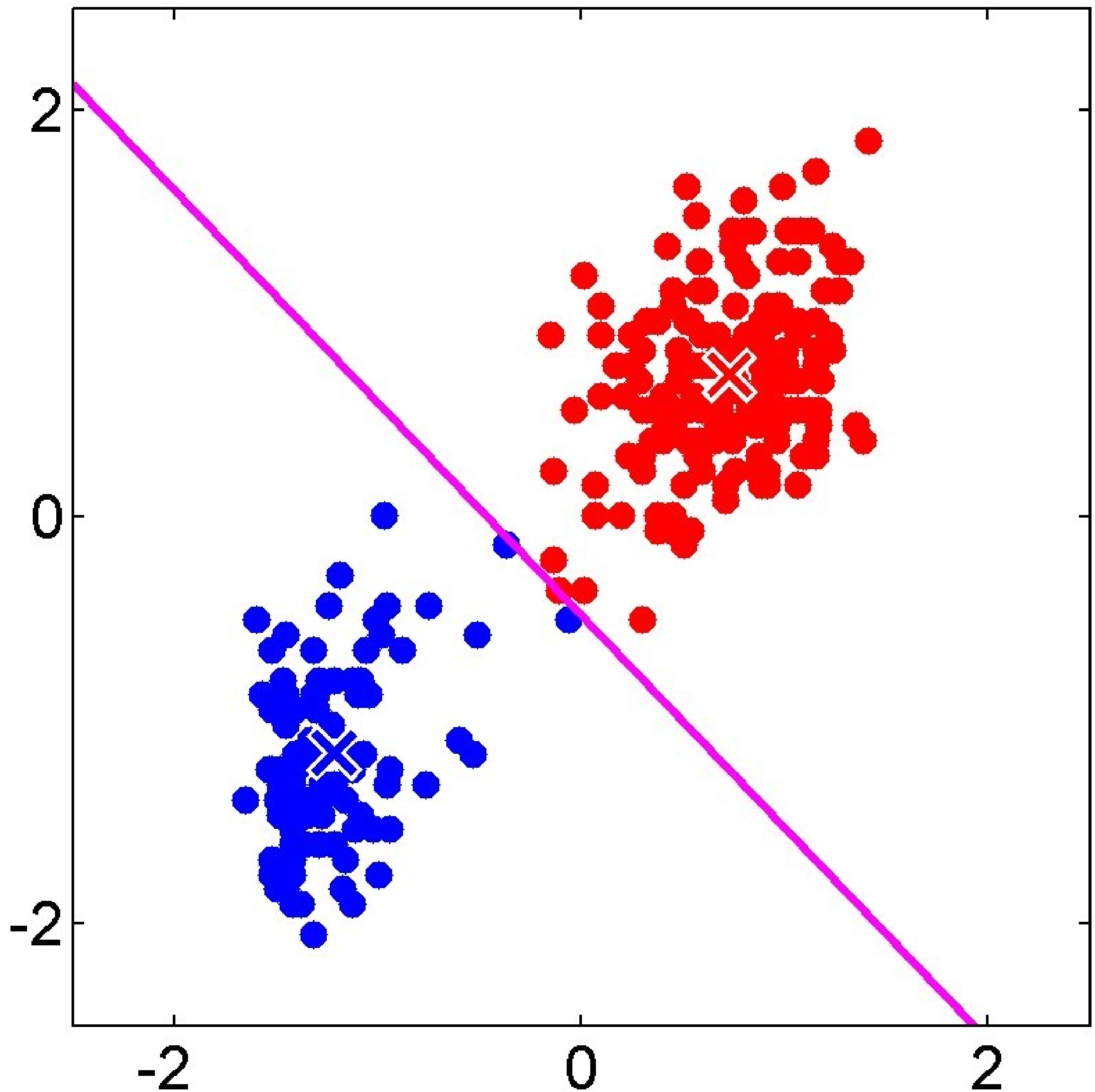


2
0
-2
-2
0
2

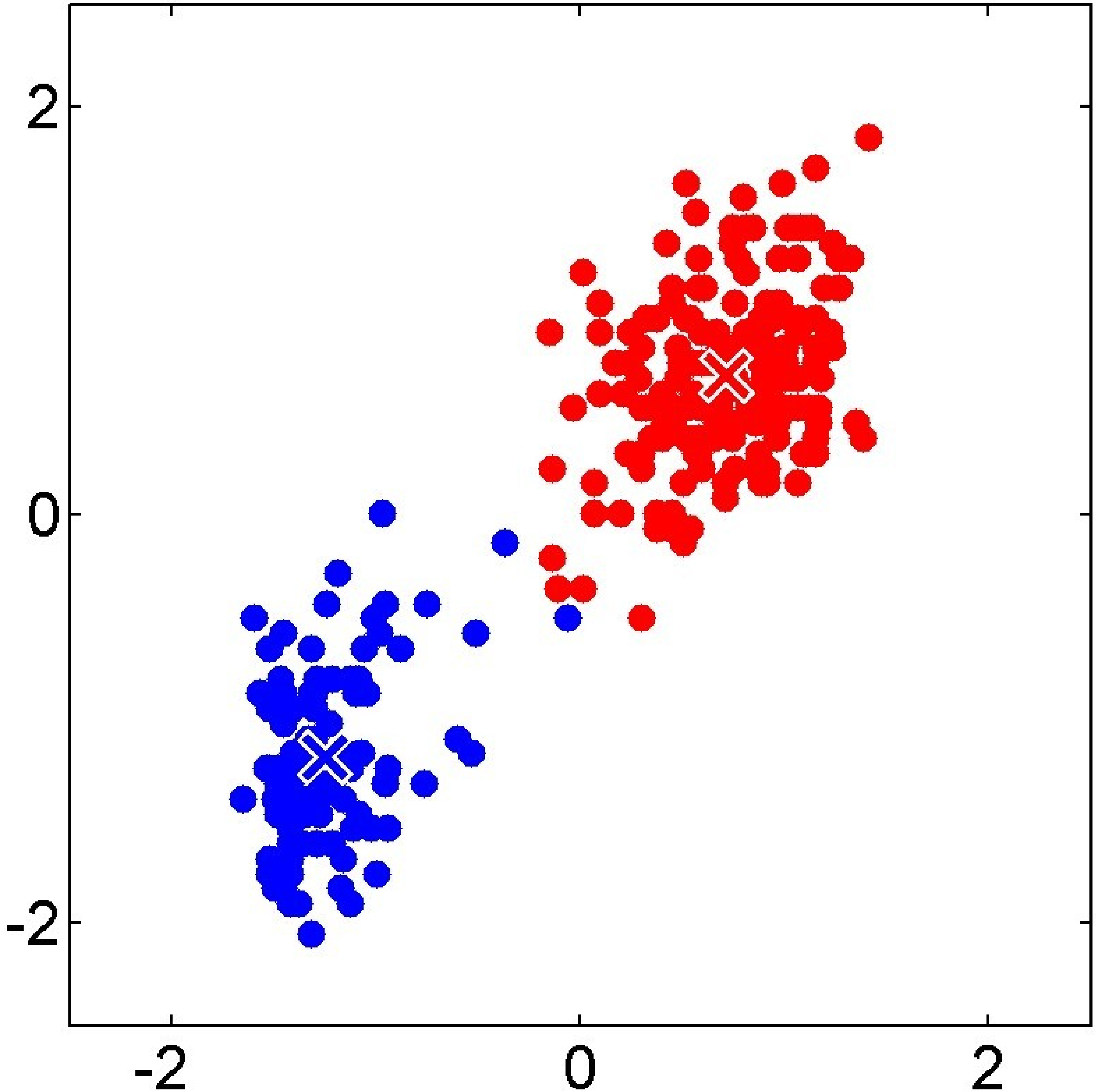


2
0
-2
-2
0
2

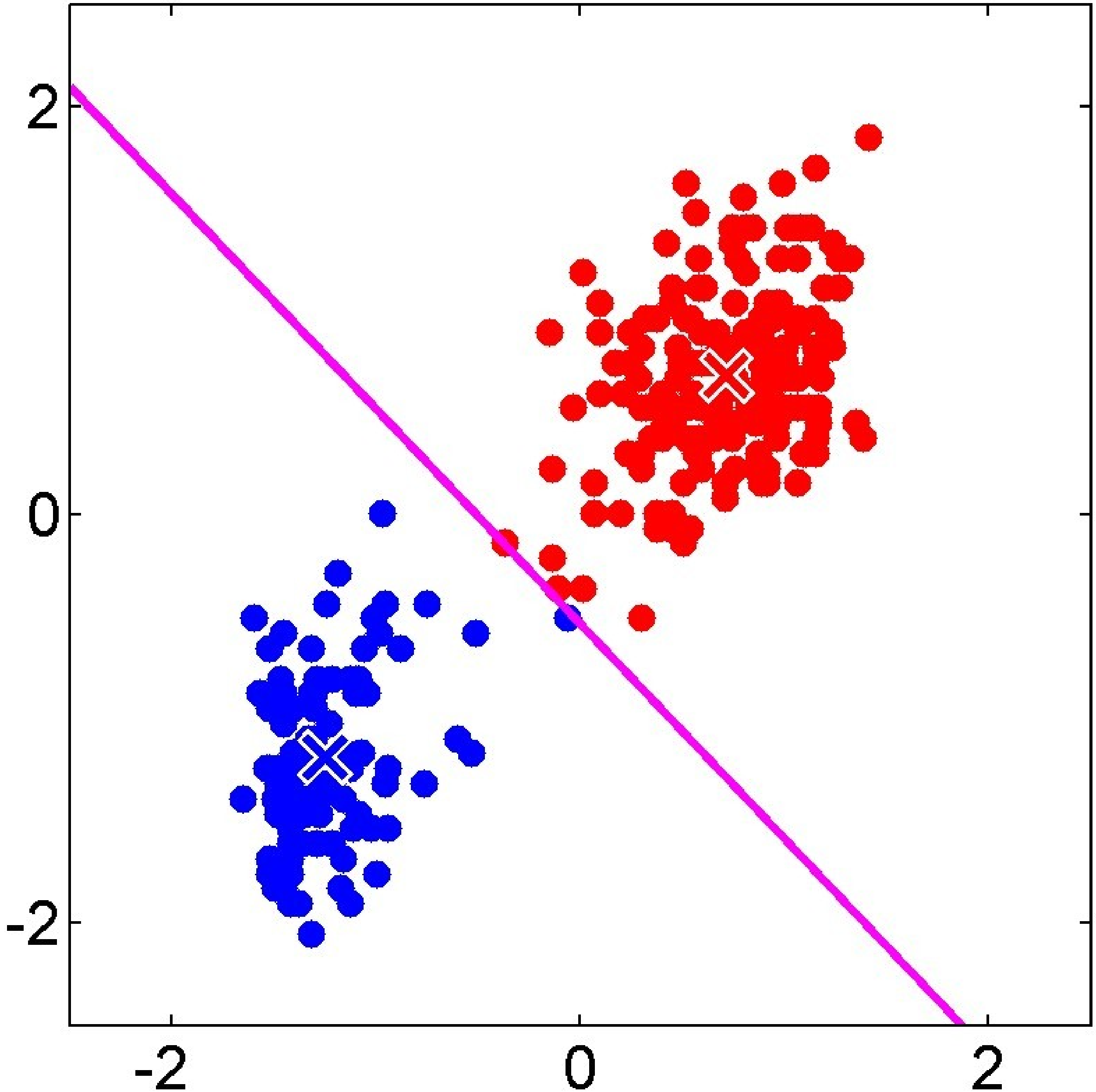

2
0
-2
-2
0
2

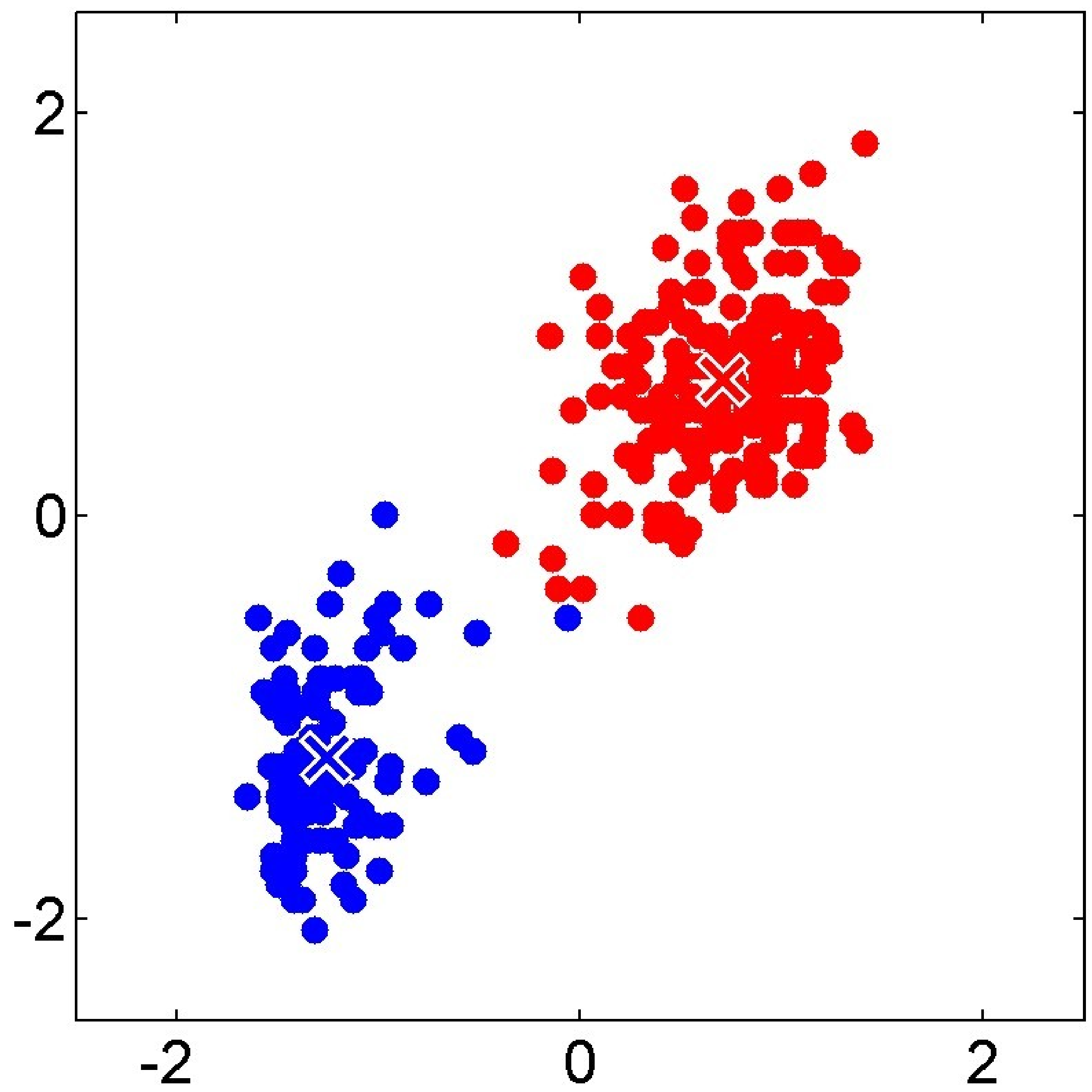


2
0
-2
-2
0
2



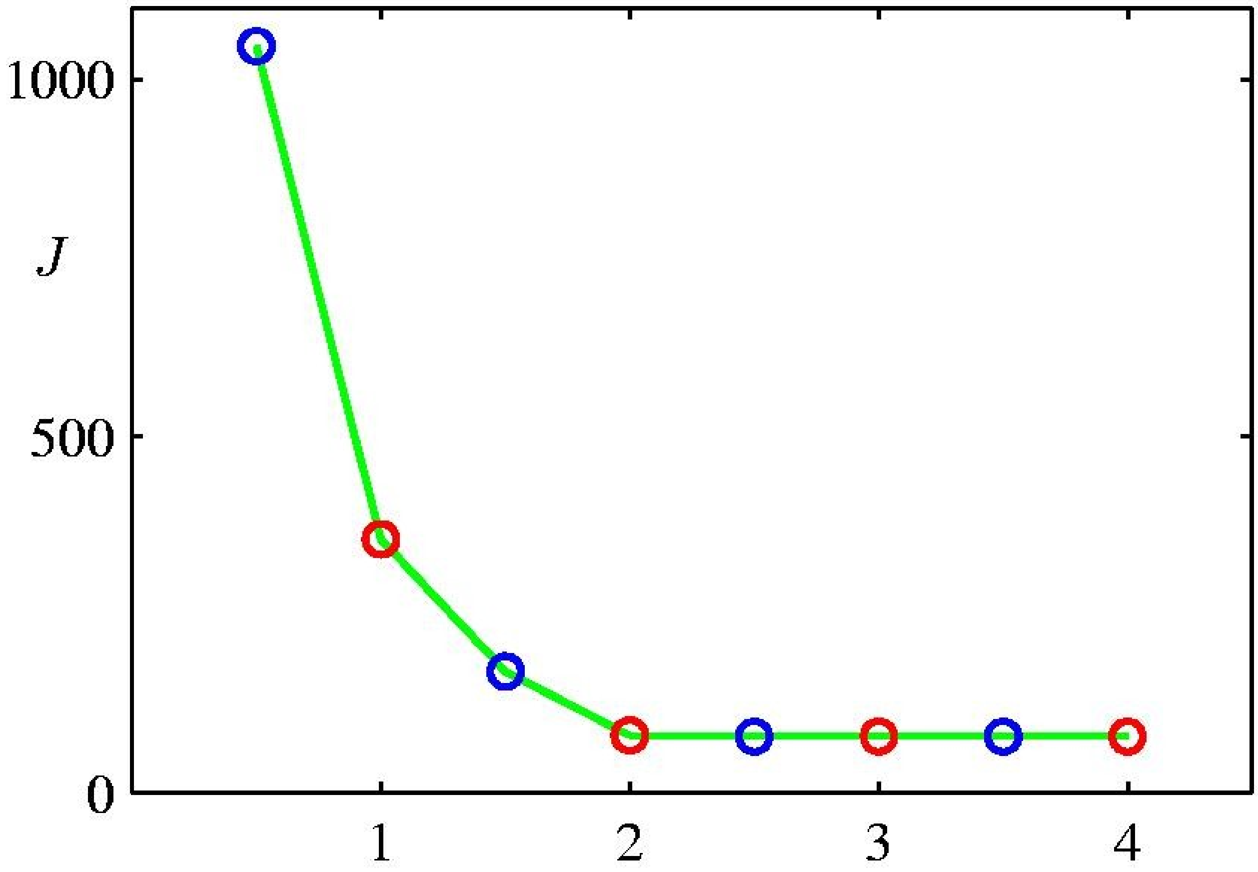
2
0
-2
-2
0
2


2
0
-2
-2
0
2



2
0
-2
-2
0
2

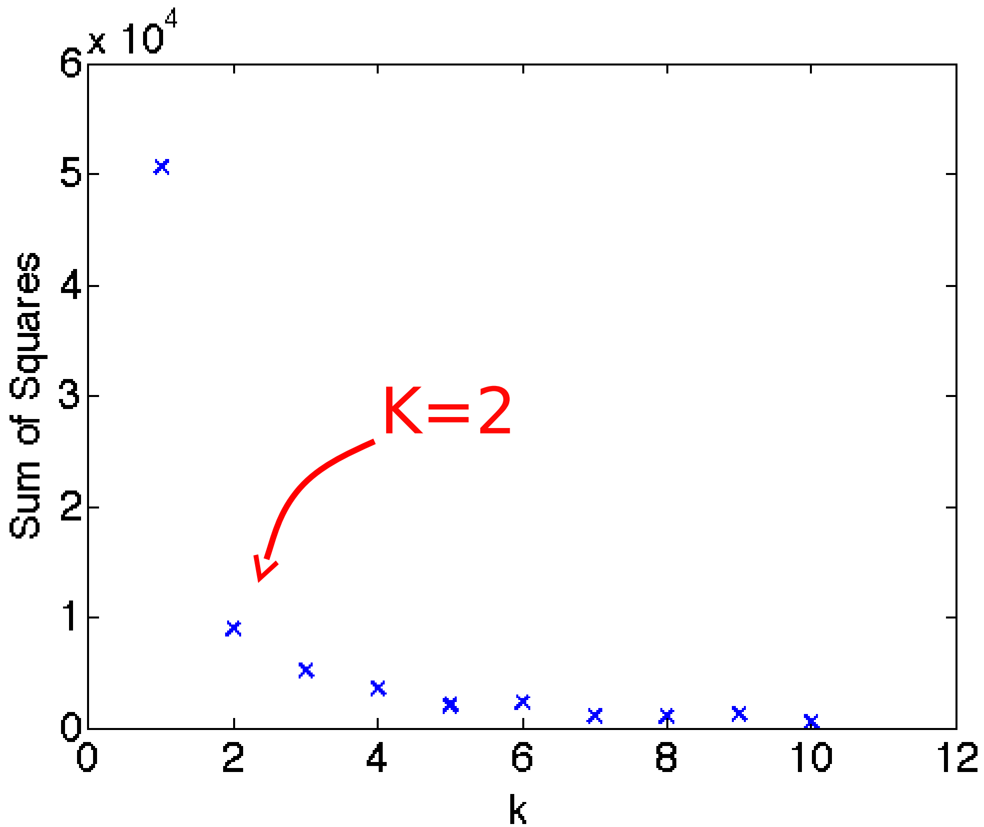
La función de pérdida después de cada paso E y M



¿Cómo escoger *K*?

- En algunos casos es conocido a priori por la naturaleza del dominio del problema.
- Generalmente, es estimado a partir de los datos de acuerdo a una heurística específica.
- La función de pérdida *J* generalmente decrece cuando K crece.
- Idea: Asuma que K^* es el número correcto
  - Se asume que para $K<K^*$ cada grupo estimado contiene un subconjunto de los verdaderos grupos subyacentes.
  - Para $K>K^*$ algunos grupos naturales están separados
  - Asi, se asume que para $K<K^*$ la función de costo decrece sustancialmente, mientras que posteriormente el decrecimiento no es significativo.

¿Cómo escoger *K*?



- A partir de K=2 el decrecimiento disminuye sustancialmente, por lo que este es el valor escogido.

Inicializando K-means

- K-means converge a un óptimo local
- Los grupos definidos dependen de los valores iniciales
- Algunas heurísticas
  - Escoger aleatoriamente K datos como prototipos.
  - Estrategia voraz (greedy). Escoger el prototipo $i+1$ tal que esté lo mas alejado posible de los prototipos $\{1,\ldots,i\}$

Limitaciones de K-means

- Asignación determinista de los datos a cada grupo.
  - Pequeñas variaciones en los datos pueden significar un cambio de grupo
  - Solución: reemplazar asignación determinista por asignación de probabilidades (GMM)
- Asume grupos esféricos y cada grupo con la misma probabilidad.
  - Solución: GMM
- Grupos arbitrarios con diferentes valores de K
  - A medida que K crece, la pertenencia a los grupos cambia de manera arbitraria, los grupos no están siempre anidados
  - Solución: Agrupamiento jerárquico
- Sensibilidad a los outliers
  - Solución: uso de otras funciones de pérdida.

Referencias

- Hastie, Tibshirani and Friedman, The Elements of Statistical Learning, Capítulo 14
- Bishop, Pattern Recognition and Machine Learning, Capítulo 9