Clustering con k-Medias

Jorge Gallego

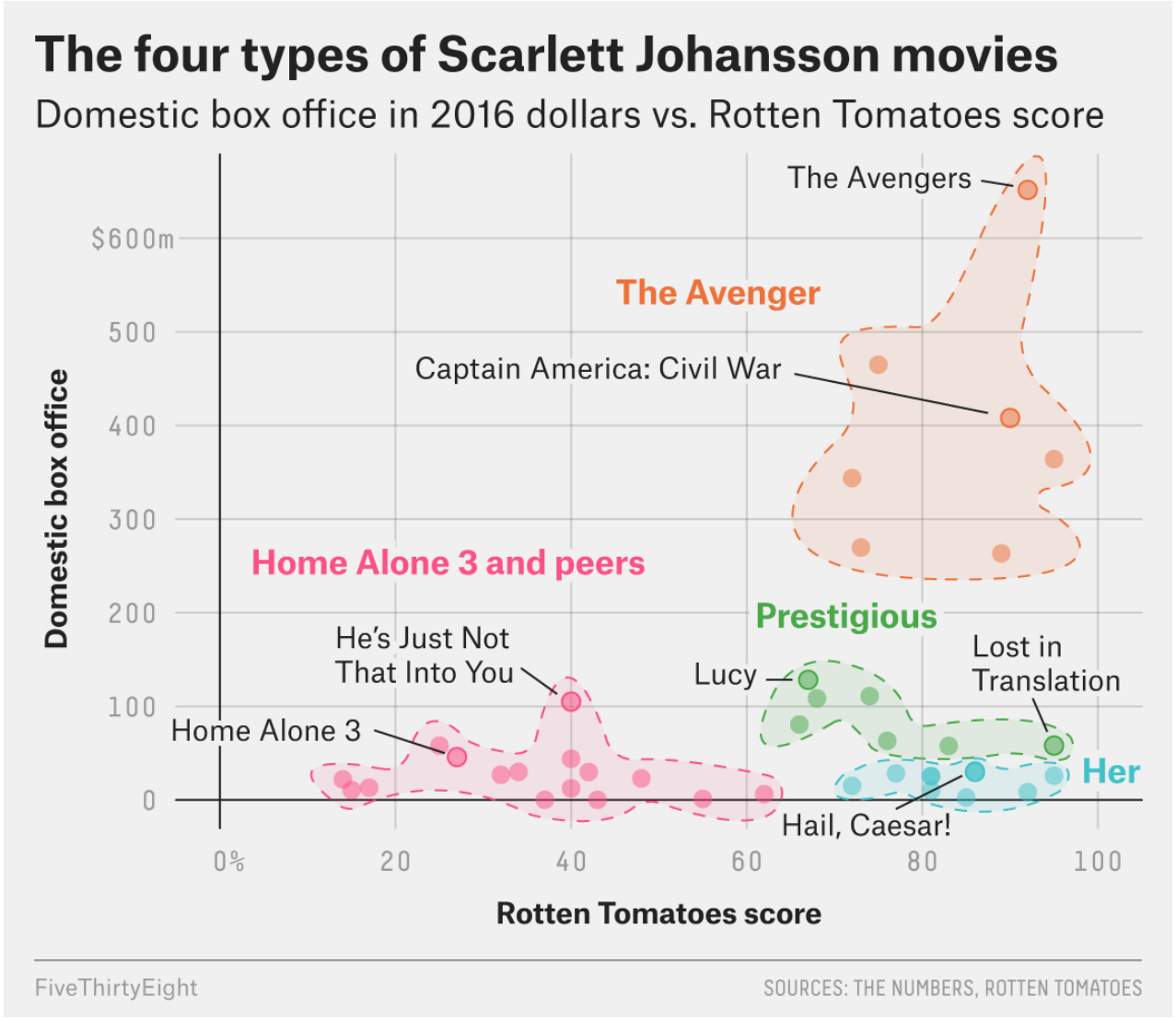
Facultad de Economía, Universidad del Rosario

Mayo 14 de 2019

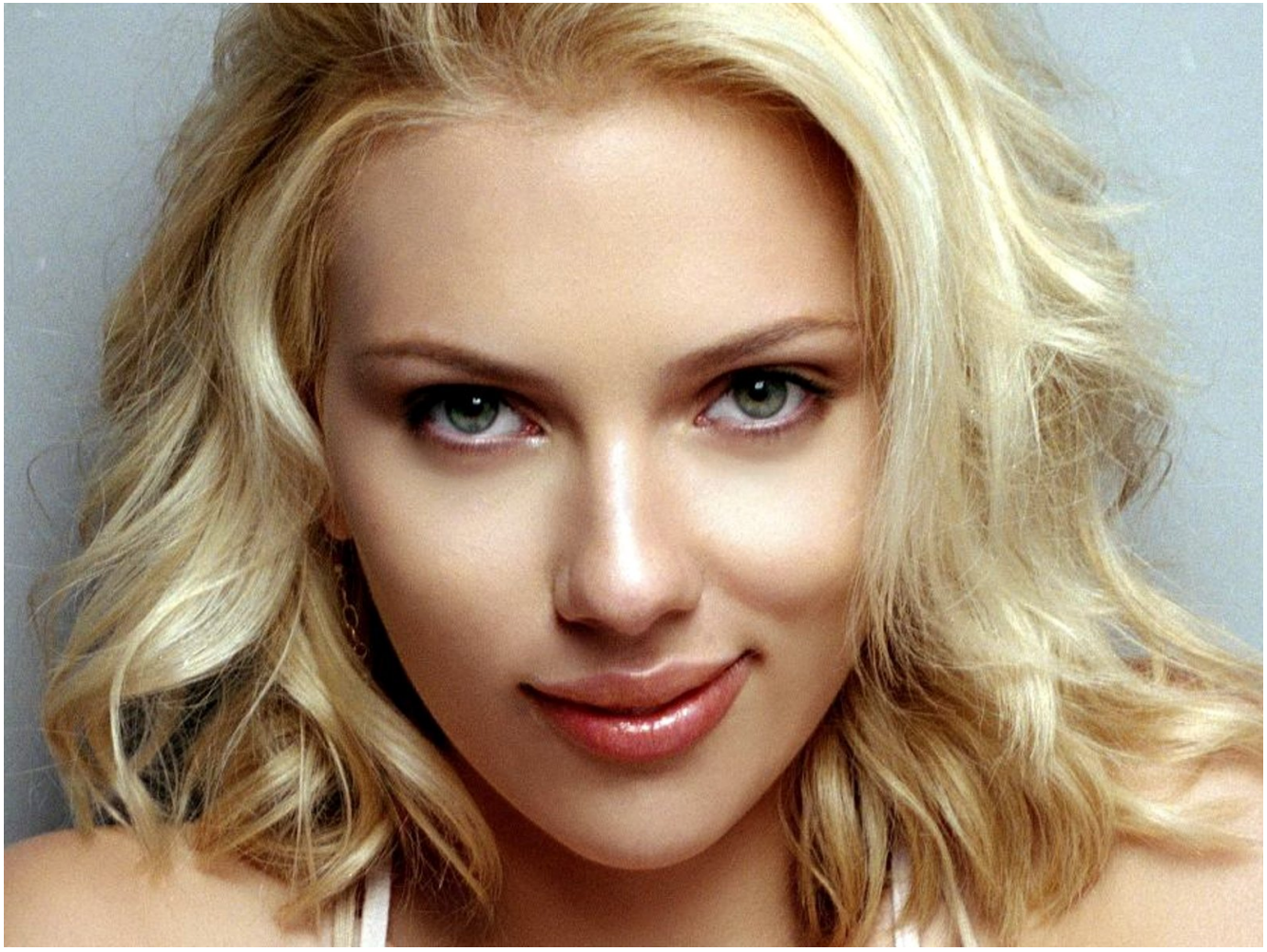
Introducción

- En muchas ocasiones es conveniente agrupar diferentes objetos
- Por ejemplo en marketing, segmentar mercados para saber qué estrategia utilizar en cada uno
- En campañas políticas lo mismo: identificar diferentes tipos de votantes
- Y en muchas otras aplicaciones suele ser relevante encontrar clusters de objetos
- Estudiaremos técnicas de agrupación basadas en algoritmos de machine learning

Películas de Scarlett Johansson



Scarlett Johansson



Clustering

- El clustering es un técnica de aprendizaje no supervisado que consiste en dividir los datos en clusters
- Los clusters son grupos de ítems similares
- El aprendizaje es no supervisado porque no se especifica ex ante cómo deben verse los grupos
- Esta técnica no es para predecir. Es para descubrir patrones
- ¿Cómo determina el computador cuáles son los límites de los grupos?

Clustering

- Items dentro de un cluster deben parecerse entre sí y debe ser muy diferentes a los que están por fuera
- Esta tarea es diferente a las de clasificación o predicción numérica
- Antes teníamos modelos que relacionaban características con outcomes o con otras características
- Se describían patrones existentes en los datos

Clustering

- Al hacer clustering creamos nuevos datos
- Ejemplos no categorizados son asignados a un cluster que se infiere de las relaciones entre los datos
- Por esta razón se le llama clasificación no supervisada
- El método clasifica ejemplos no etiquetados

Clustering

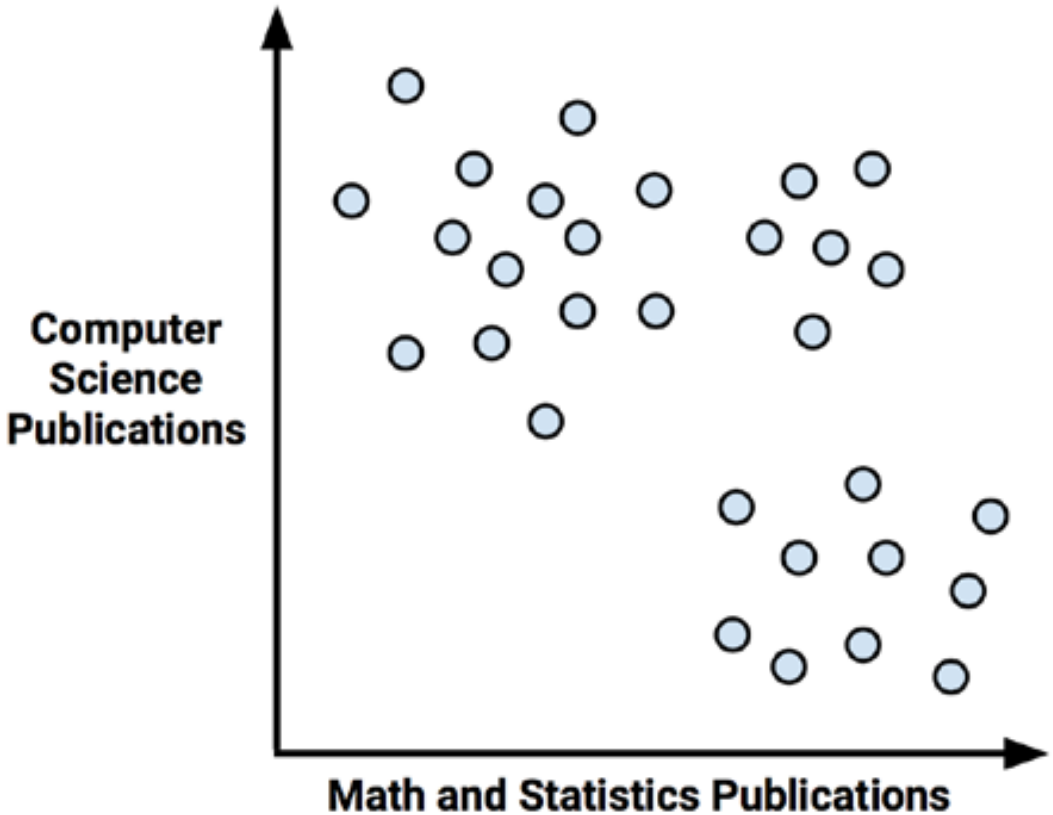
- Los grupos que devuelve el computador no tienen un significado explícito
- La interpretación la debe dar el analista
- Consideremos el siguiente ejemplo
- Se organiza una conferencia en data science
- Queremos sentar a los asistentes según su especialidad:
- Pero ex ante no sabemos a qué áreas pertenecen

Clustering

- ¿Es posible inferir estas categorías a partir de ciertas características de los asistentes?
- Por ejemplo, el tipo de publicaciones que tienen
- Supongamos dos campos de publicación: journals de computer science y de matemáticas y estadística
- Podemos describir a cada invitado según estas dos características

Ejemplo: Conferencia de Nerds

Figure: Publicaciones en Campos

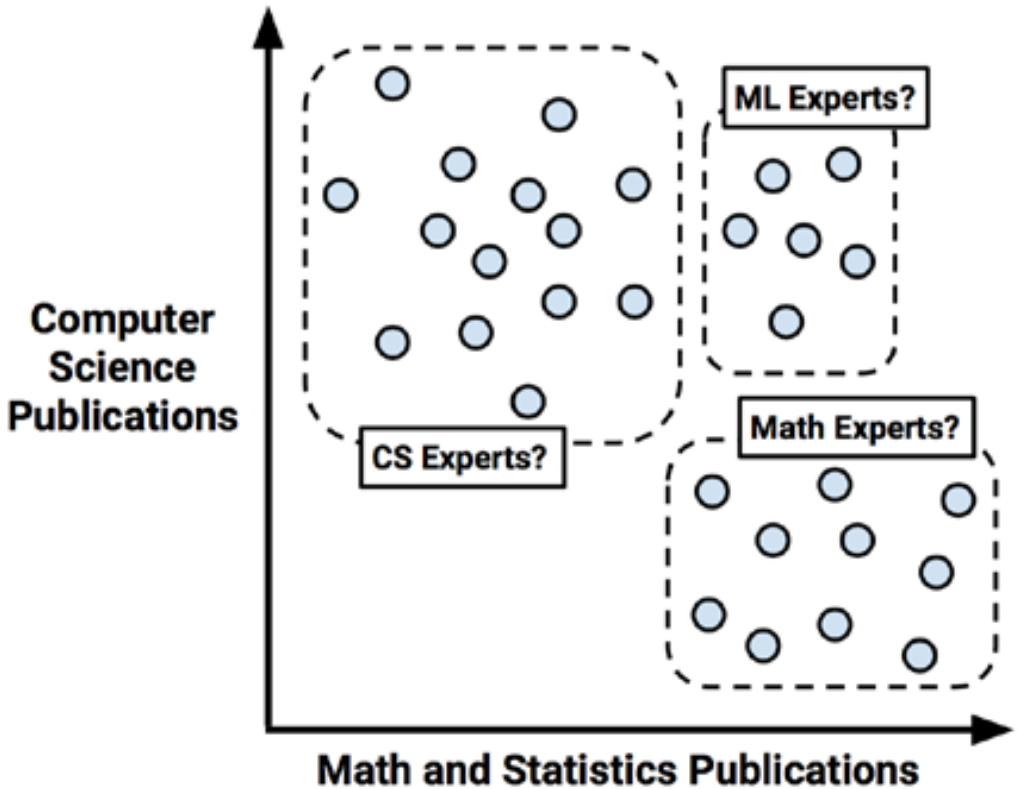


Ejemplo: Conferencia de Nerds

- ¿Es evidente algún patrón?
- ¿Podemos identificar clusters de algún tipo?
- Computer scientists: muchas en computer science y pocas en matemáticas
- Matemáticos y estadísticos: muchas en matemáticas y pocas en computer science
- Machine learning: muchas en las dos

Ejemplo: Conferencia de Nerds

Figure: Publicaciones en Campos



Algoritmo de k-Medias

- ¿Cómo se forman los clusters?
- Existen diferentes algoritmos. El más usado es el de *k-means*
- Tiene una relación estrecha con el algoritmo kNN que vimos para clasificación
- El algoritmo asigna cada uno de los n ejemplos a uno de k clusters
- El número de clusters k se determina ex ante y de manera heurísticas
- Se buscan minimizar las diferencias intra clusters y maximizar las diferencias inter clusters

Algoritmo de k-Medias

El proceso se adelanta en dos fases:

1. Se asignan ejemplos a un conjunto inicial de k clusters
2. Se actualizan las asignaciones al ajustar las fronteras de los clusters de acuerdo con los ejemplos que caen en cada cluster

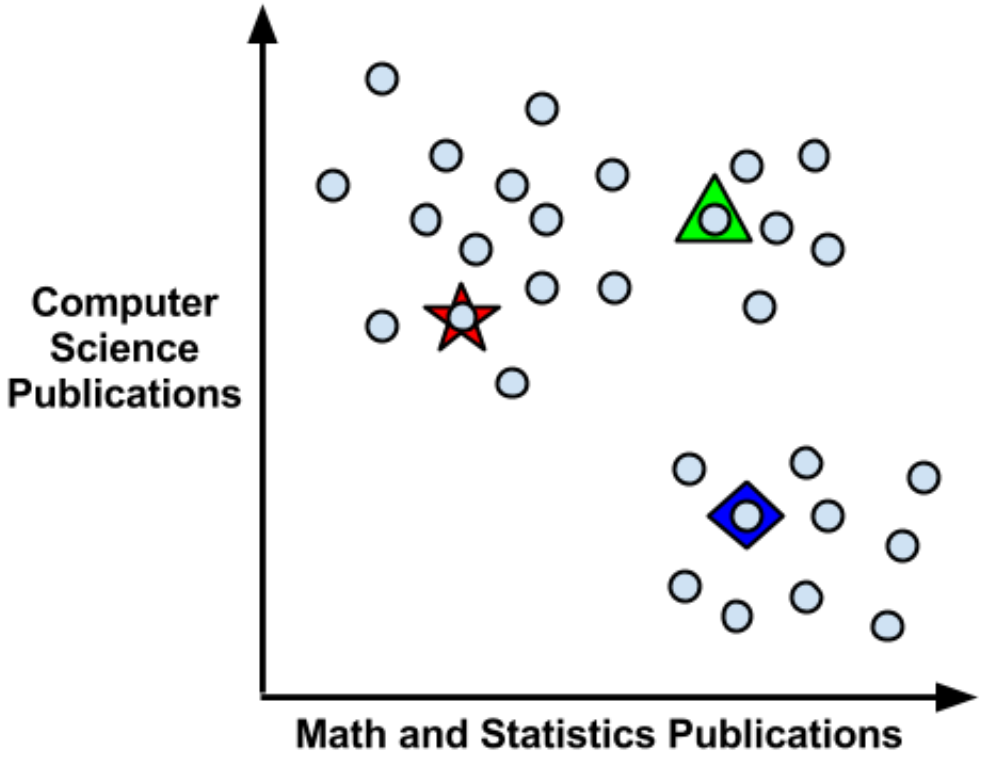
El proceso de ajuste se lleva a cabo hasta que no haya ganancias adicionales

Algoritmo de k-Medias

- Volvamos al ejemplo anterior
- Aleatoriamente se escogen k puntos para funcionar como centros de cada cluster
- En este ejemplo de manera predeterminada, por simple lógica, elegimos $k = 3$
- Pero la elección del número de clusters es un tanto heurística

Ejemplo: Conferencia de Nerds

Figure: Elección de centros



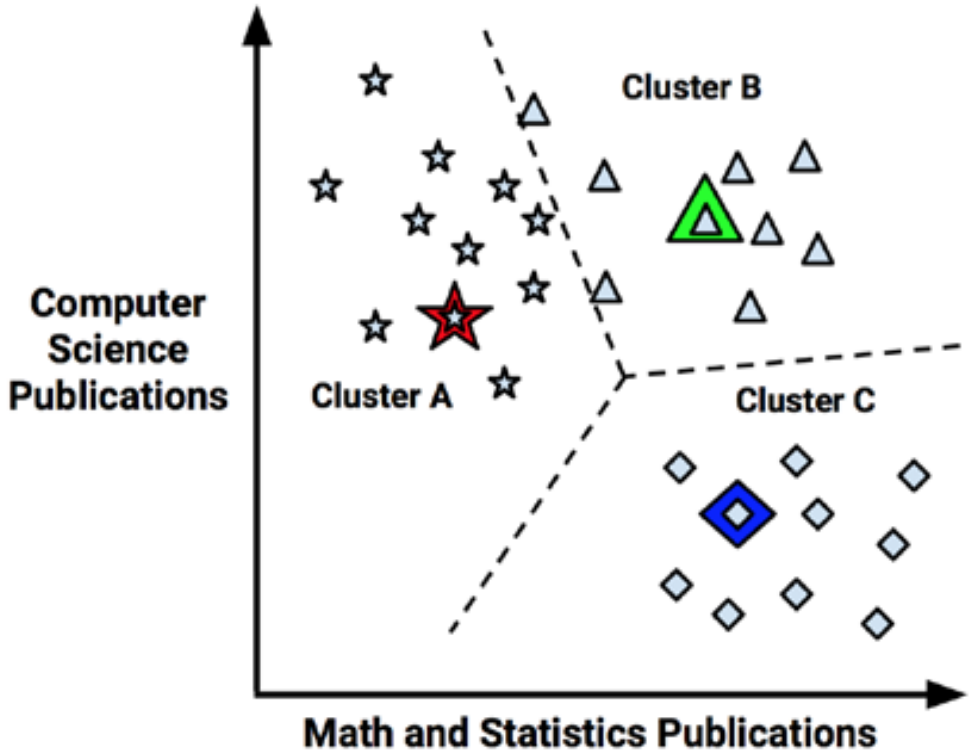
Algoritmo de k-Medias

- Cada ejemplo es asignado al cluster cuyo centro sea el más cercano
- Pare esto se usa alguna función de distancia
- Recordemos que la más usada es la distancia euclídea

$$dist(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Ejemplo: Conferencia de Nerds

Figure: Elección de centros

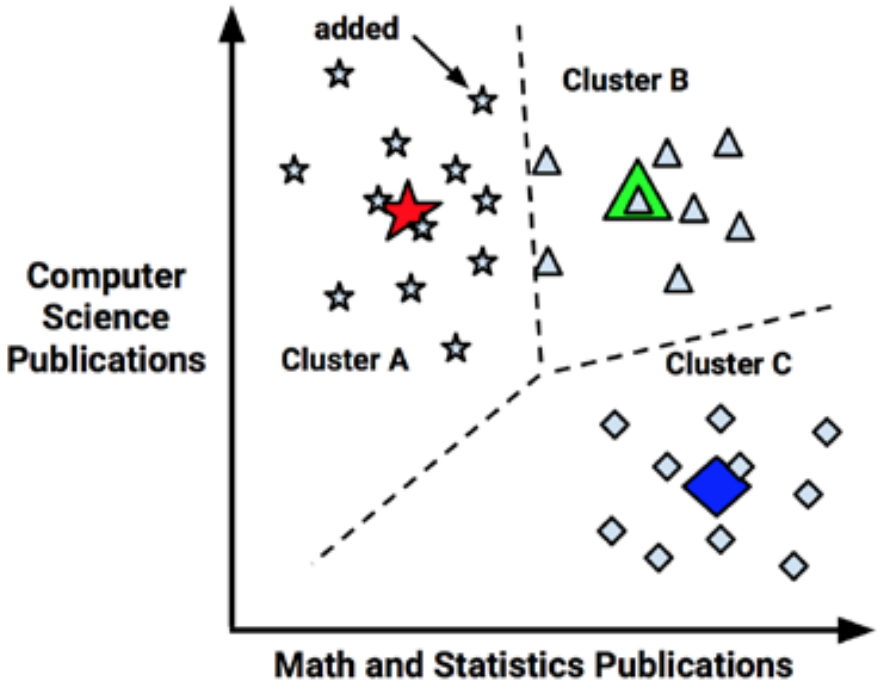


Algoritmo de k-Medias

- Este es el diagrama de Voronoi
- Indica las áreas que están más cerca a un centro que a los otros
- Esta es la primera fase del algoritmo
- Sigue la fase de ajuste. Se cambia el centro de cada cluster para formar los centroides
- El centroide es la posición promedio de los objetos de cada cluster

Ejemplo: Conferencia de Nerds

Figure: Ajuste de los centroides

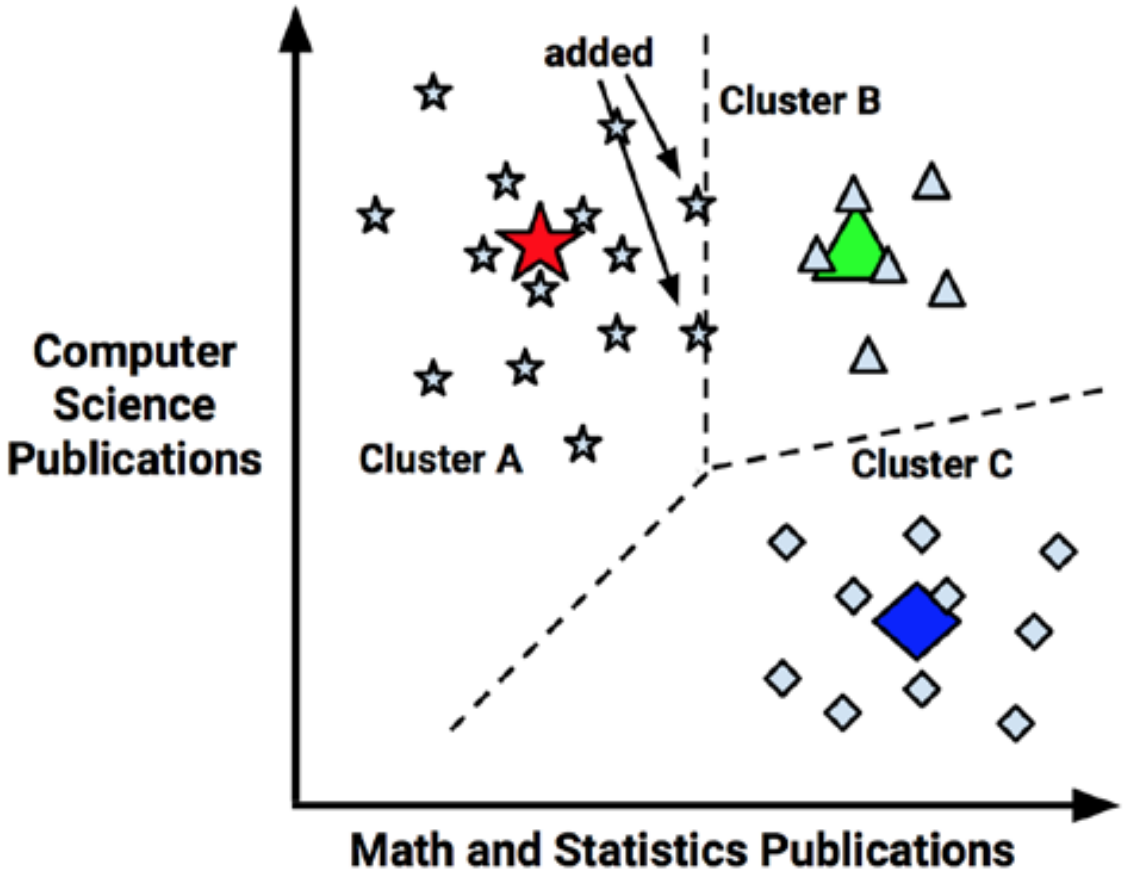


Algoritmo de k-Medias

- Con los nuevos centros se recalculan los clusters
- Hay objetos que puede que cambien de cluster porque ahora están más cerca de otro centro
- Y estos cambios alteran la posición del centroide
- El proceso se repite iterativamente
- Hasta que no haya cambios en la composición de los clusters

Ejemplo: Conferencia de Nerds

Figure: Ajuste de los centroides



Algoritmo de k-Medias

- Al terminar el proceso, se pueden reportar los resultados de dos formas
- Indicando cuáles son los clusters
- O reportando las coordenadas de los centroides de los clusters
- Cualquiera de los dos métodos sirve al propósito de ilustrar los grupos resultantes

Número Apropiado de Clusters

- ¿Cuántos clusters elegir?
- El algoritmo es sensible no solo al número de clusters
- También a quiénes sean los clusters iniciales
- Por eso estas elecciones son cruciales
- k muy alto mejora la homogeneidad al interior de los clusters; pero se corre el riesgo de sobrestimar el modelo

Número Apropiado de Clusters

- Lo ideal es tener una creencia a priori sobre el número de grupos que debería haber
- Si clasificáramos películas, el número de grupos podría ser el número de categorías según la academia
- A veces otras condiciones determinan el número de clusters. Por ej., el número de mesas en la conferencia
- En marketing, puede ser el número de campañas distintas que se pueden hacer
- Si no hay ningún prior, una regla de dedo pulgar es usar $k = \sqrt{n/2}$

Número Apropiado de Clusters

- Otro método es el del codo (*elbow method*)
- Sabemos que al aumentar k aumenta (disminuye) la homogeneidad (heterogeneidad) dentro de los grupos
- Pero no podemos aumentar indefinidamente k
- Nos detenemos cuando la ganancia marginal sea muy baja

Número Apropiado de Clusters

Figure: Método del Codo