SolidQ

Think Big. Move Fast.

Módulo 3: Entendiendo los algoritmos de Machine Learning

Segunda Parte

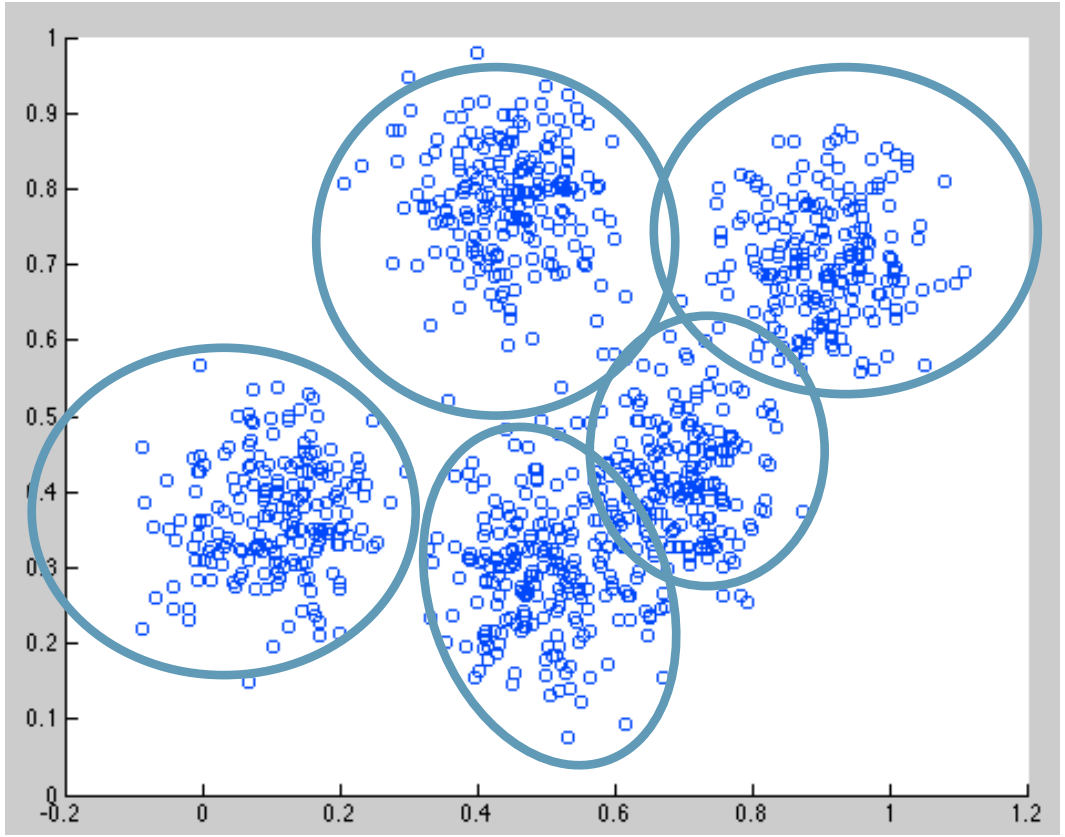
Agenda

- Introducción a los algoritmos
- Datos de Entrenamiento y Prueba
- Clasificación
- Regresión
- Clustering
- Recomendaciones
- Conjuntos de datos No-balanceados
- Como interpretar modelos

Clustering

Clustering

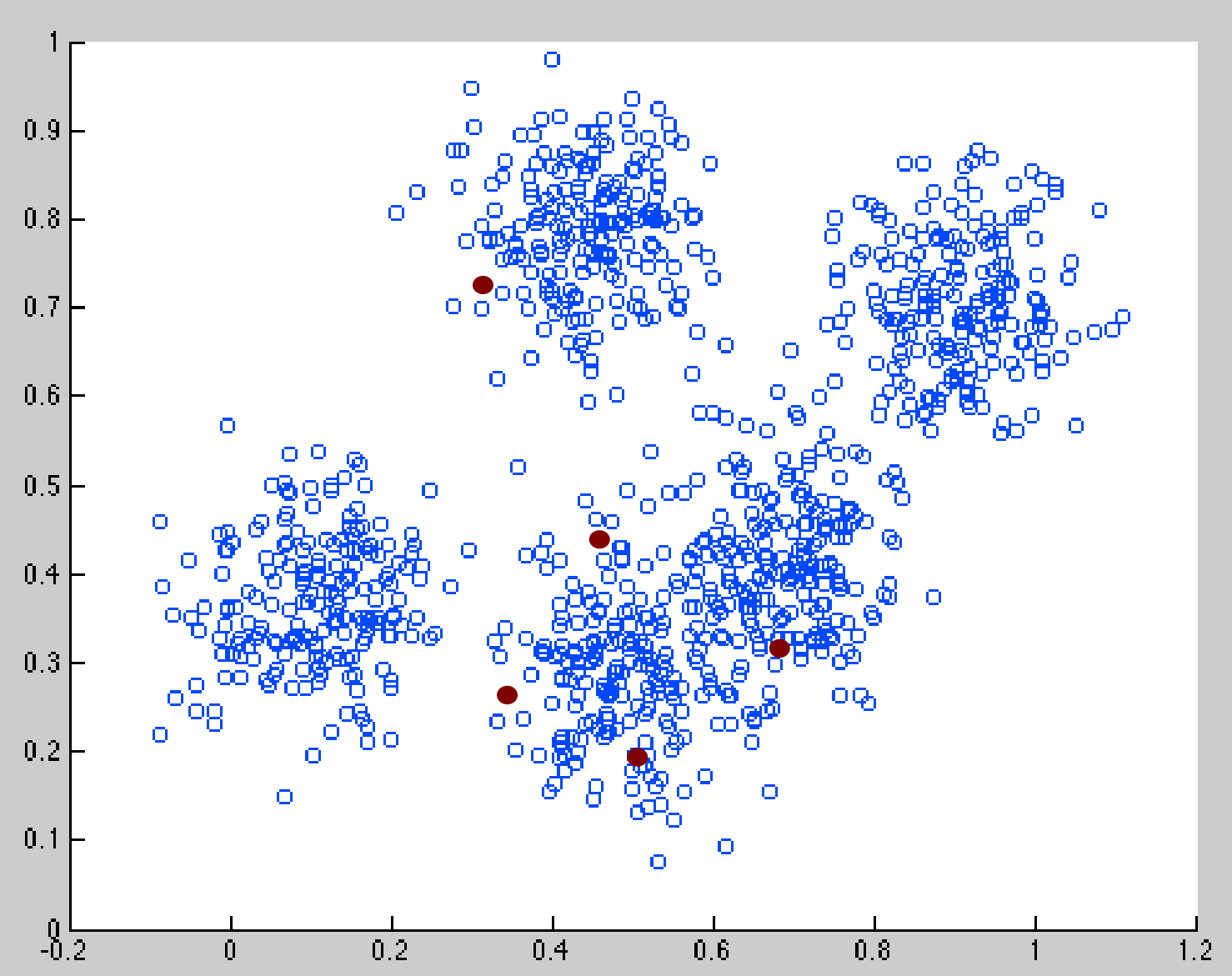
- Datos en un cluster deben de ser similares a otros miembros de ese cluster



K-means

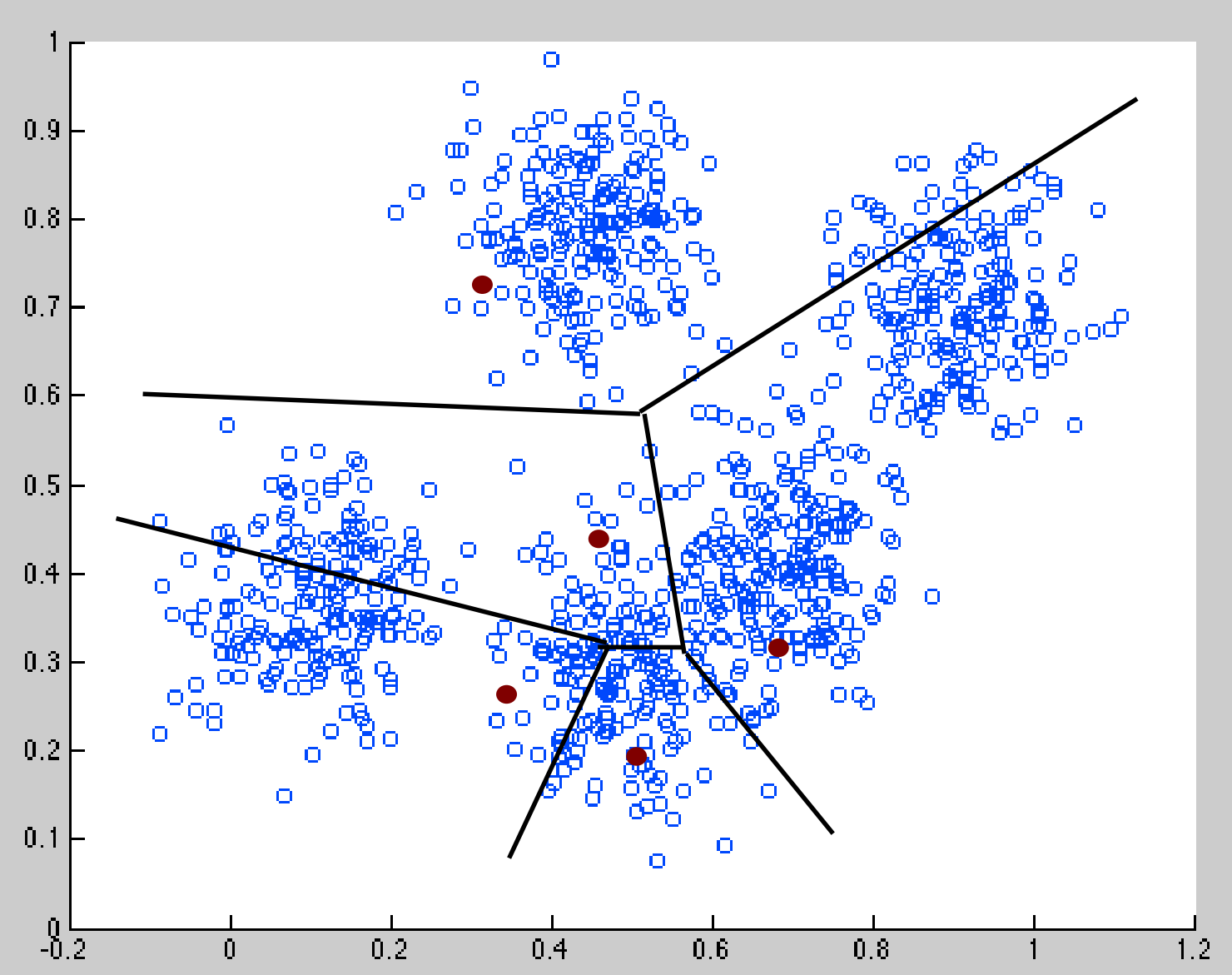
- Como entrada el número de clusters. Inicializa los centros aleatoriamente
- Asigna todos los puntos al centro de cluster más cercano
- Cambia los centros de los clusters para que se encuentren en el centro de sus puntos
- Repite hasta la convergencia

K-Means en acción



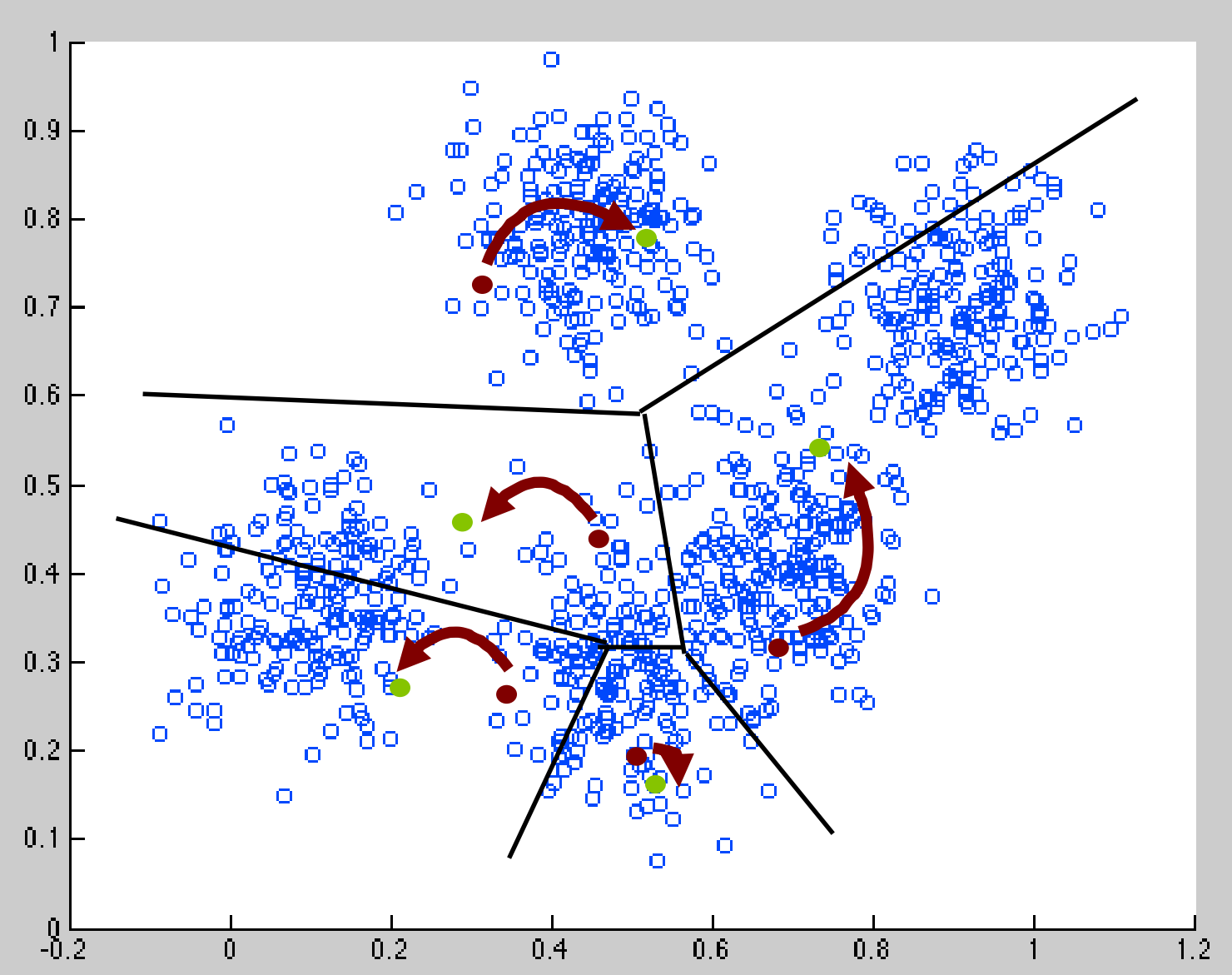
1. Entrada número de clusters, inicializa los centros aleatoriamente
2. Asigna todos los puntos al centro del cluster más cercano
3. Cambia los centros del cluster para que esté en el medio de sus puntos
4. Repite hasta converger

K-Means en acción



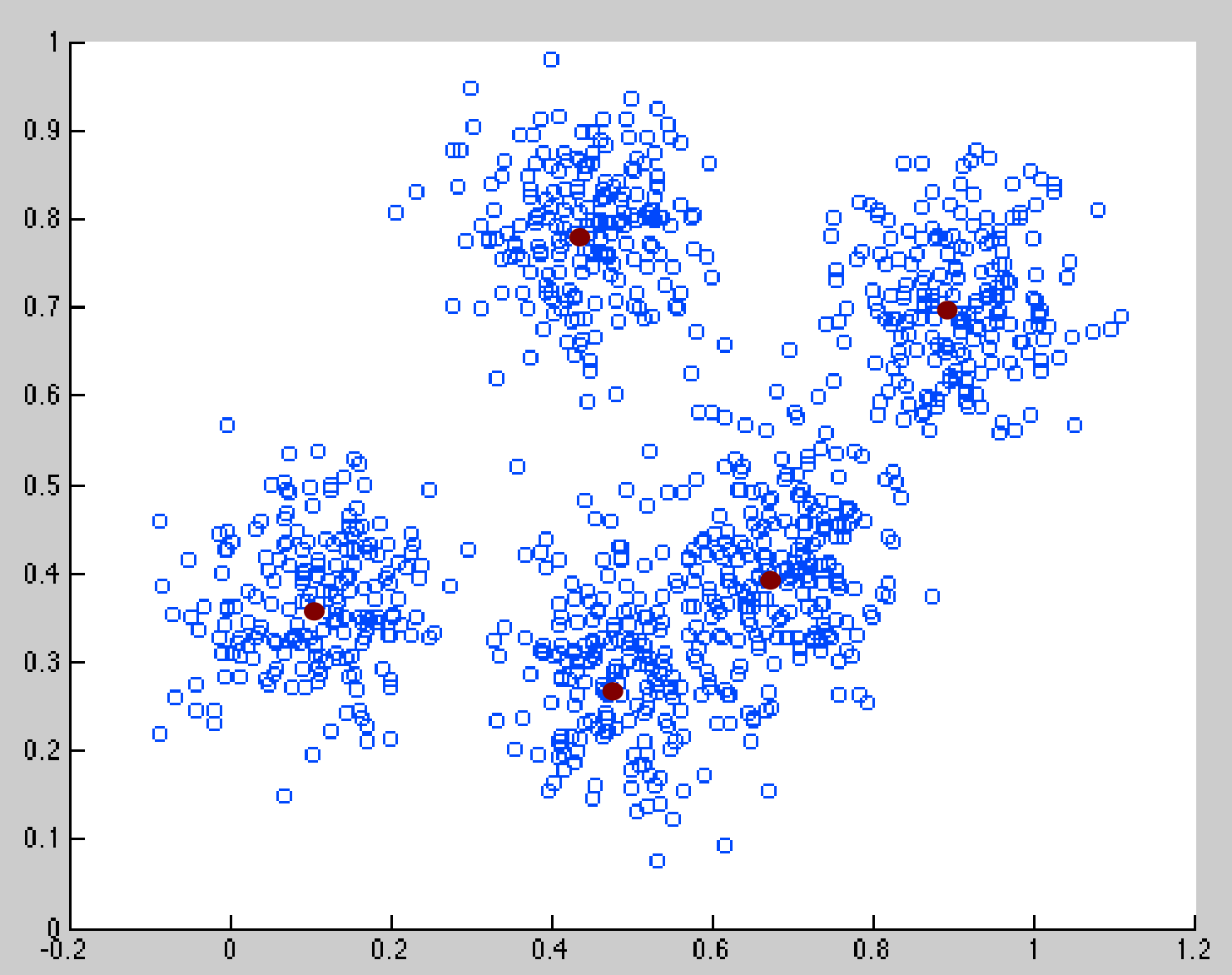
1. Entrada número de clusters, inicializa los centros aleatoriamente
2. Asigna todos los puntos al centro del cluster más cercano
3. Cambia los centros del cluster para que esté en el medio de sus puntos
4. Repite hasta converger

K-Means en acción



1. Entrada número de clusters, inicializa los centros aleatoriamente
2. Asigna todos los puntos al centro del cluster más cercano
3. Cambia los centros del cluster para que esté en el medio de sus puntos
4. Repite hasta converger

K-Means in action

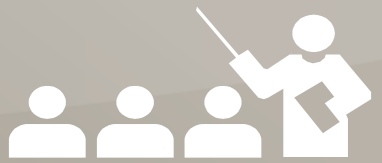


1. Entrada número de clusters, inicializa los centros aleatoriamente
2. Asigna todos los puntos al centro del cluster más cercano
3. Cambia los centros del cluster para que esté en el medio de sus puntos
4. Repite hasta converger

K-Means

- Algoritmo muy popular de clustering que es eficiente desde un punto de vista de cómputo
- Realiza una minimización alterna en la función de coste
- No siempre minimiza esa función de coste (suele ser habitual realizer multiples replicas para obtener una Buena solución)
- Podemos utilizar la función de coste para evaluar si una replica es mejor que la otra
- Podemos utilizar la función de coste para decider el número de clusters
- No funciona bien para clusters que muy no-esféricos

Demo 03 E – Clustering



K-means

Recomendadores

Recomendadores

- Recomendar los ítems más populares
- Usar un clasificador para hacer una recomendación
- Algoritmos de Recomendación
  - Basados en Contenido
  - Filtrado colaborativo

User-Based Collaborative Filtering

- Calcular el score de Carmen para Alien utilizando valoraciones similares de otros usuarios

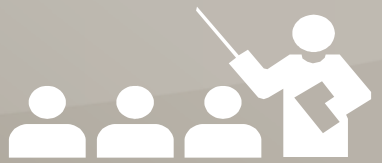
	Alien	Bug's Life	Cars	Dark Knight
Carmen	?	4	4	1
Joseph	5	4	-	2
Leonore	1	1	-	3
Esmerelda	2	3	1	-

Item-Based Collaborative Filtering

- Calcular el score de Joseph para Dark Knight utilizando ratings similares de items

	Alien	Bug's Life	Cars	Dark Knight
Carmen	2	-	1	1
Joseph	5	4	2	?
Leonore	4	-	3	3
Esmerelda	-	4	1	-

Demo 03 F- Recomendadores

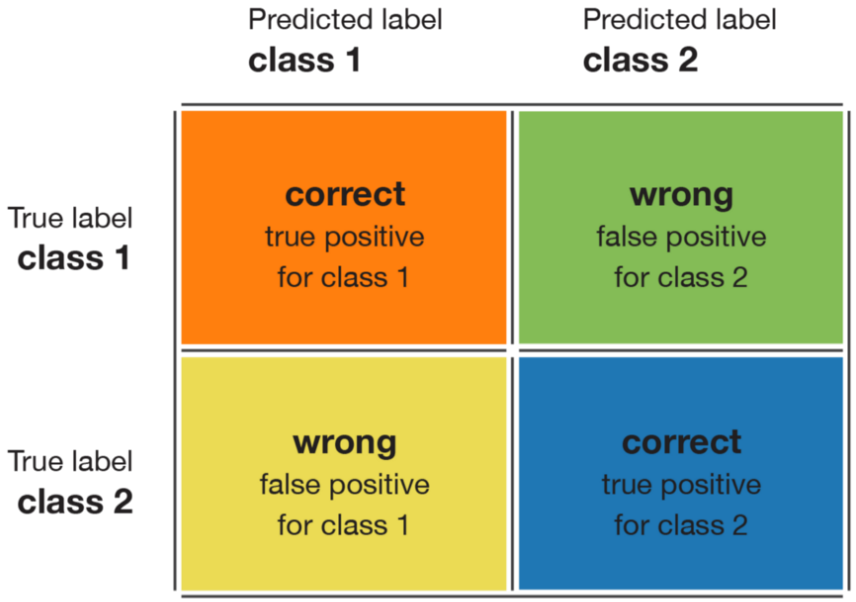


Conjuntos de datos No-balanceados

Validación: Detectar un “mal comportamiento”

- Matriz de Confusión
- ROC y AUROC
- Soluciones
  - Trabajar el conjunto de datos
    - Menos muestras
    - Más muestras
    - Generar valores sintéticos (SMOTE)
  - ¿Cambiar la pregunta?

Matriz de Confusión

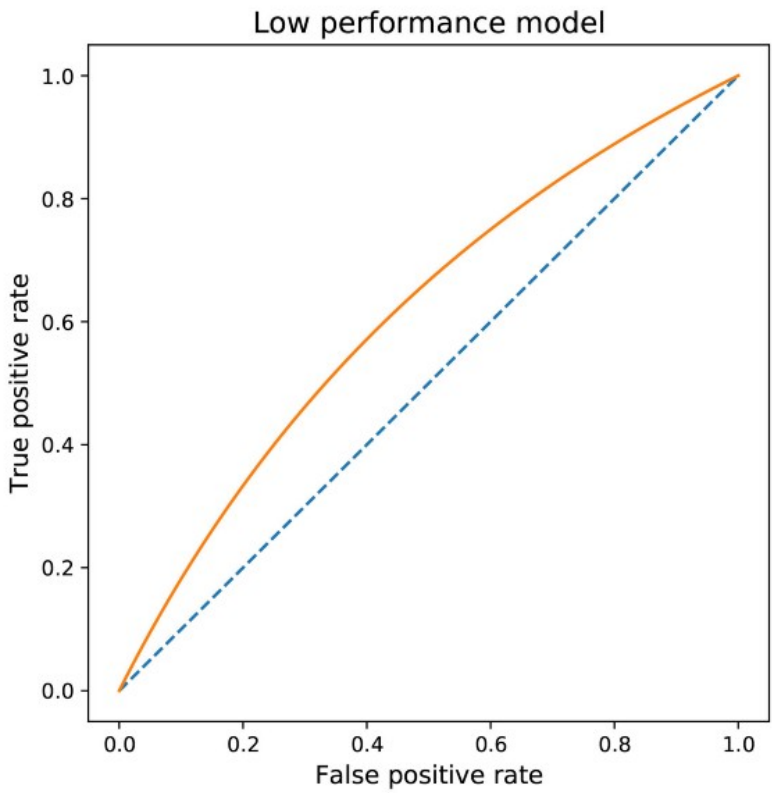


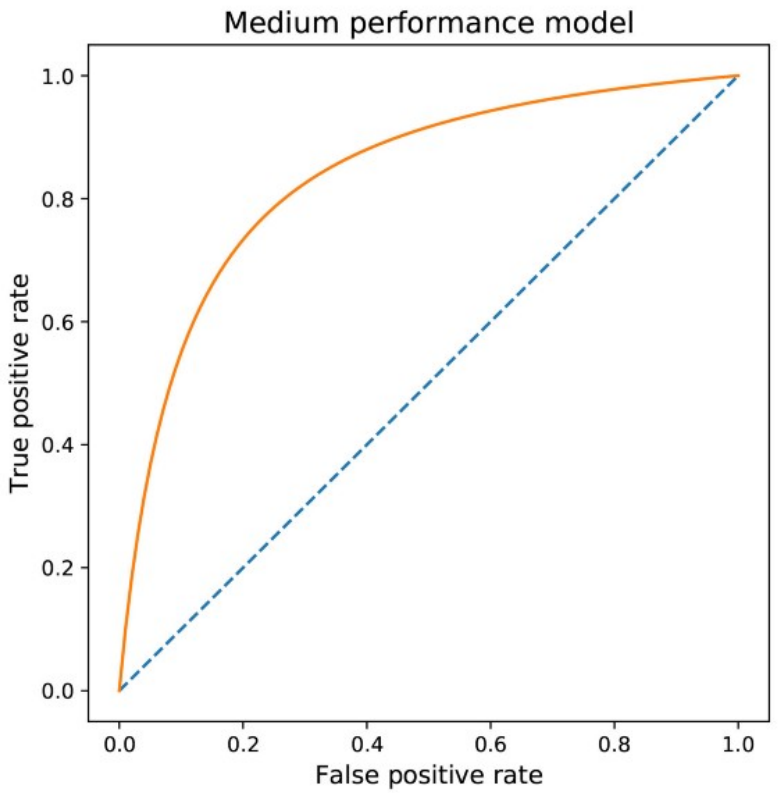
$$\text{accuracy} = \frac{\blacksquare + \blacksquare}{\blacksquare + \blacksquare + \blacksquare + \blacksquare}$$

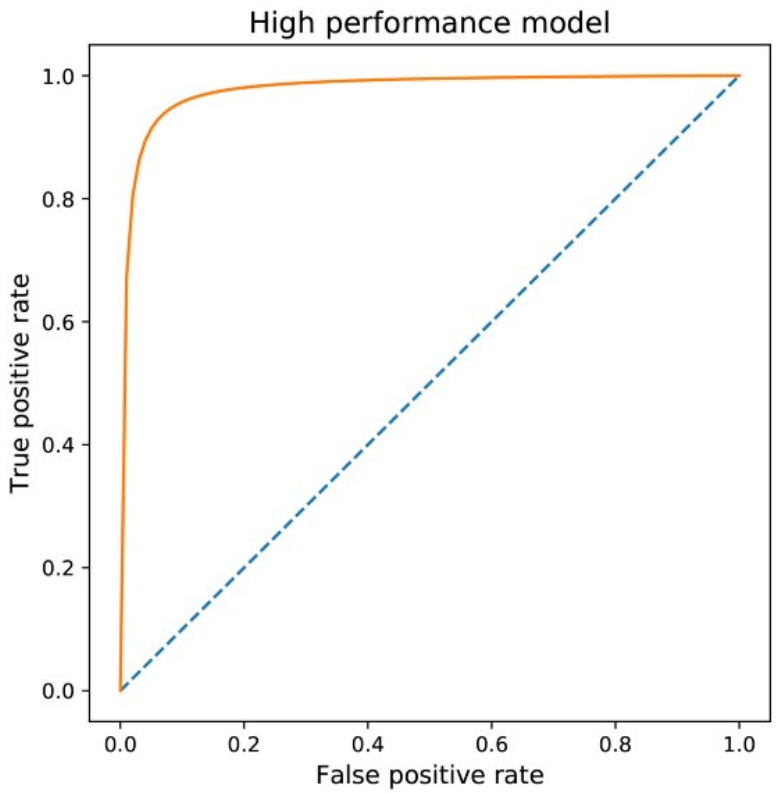
$$\text{class 1 precision} = \frac{\blacksquare}{\blacksquare + \blacksquare} \qquad \text{class 2 precision} = \frac{\blacksquare}{\blacksquare + \blacksquare}$$

$$\text{class 1 recall} = \frac{\blacksquare}{\blacksquare + \blacksquare} \qquad \text{class 2 recall} = \frac{\blacksquare}{\blacksquare + \blacksquare}$$

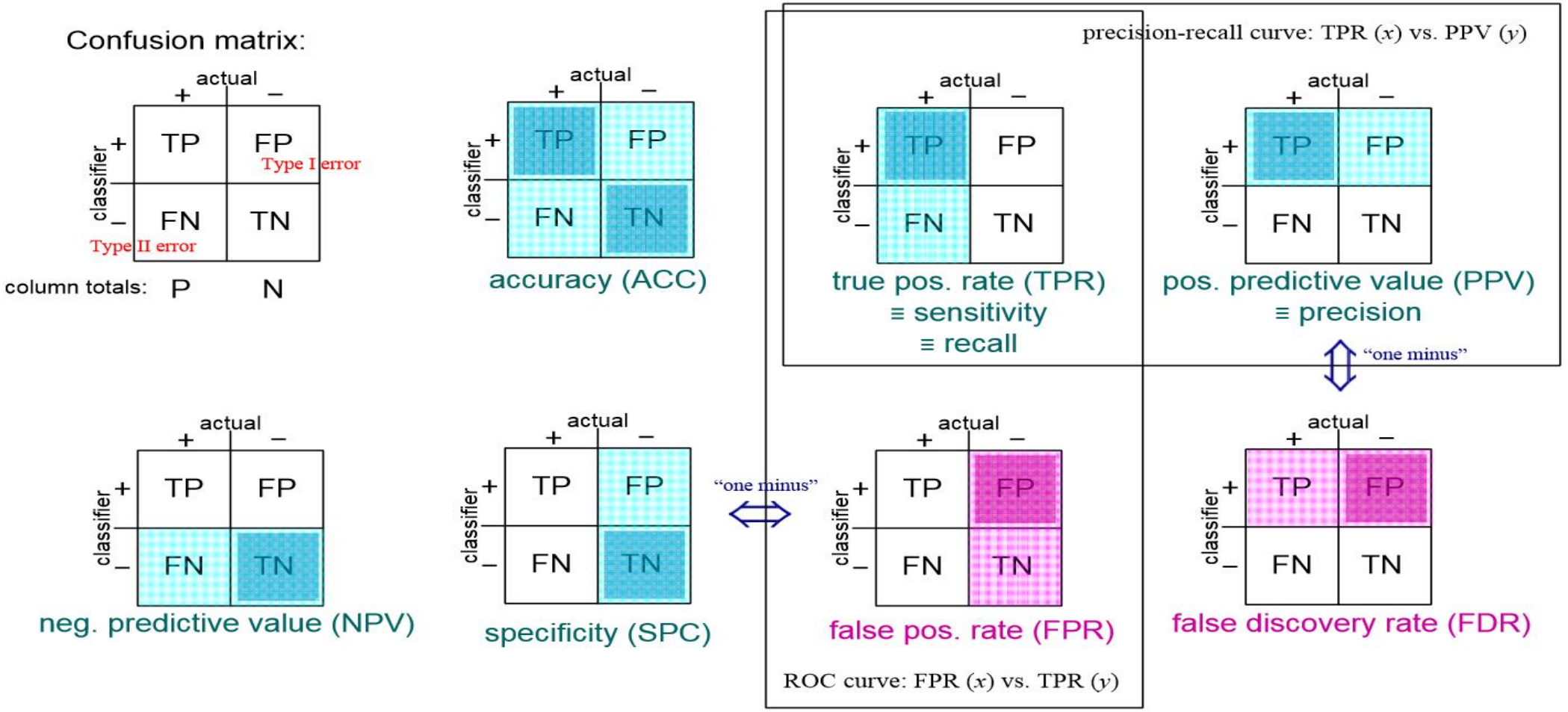
Curva ROC







"Cheat sheet" on accuracy, precision, recall, TPR, FPR, specificity, sensitivity, ROC, and all that stuff!
William H. Press, ver 1.0, 3/29/08



value (between 0 and 1) = numerator / denominator
numerator = dark color shade
denominator = dark + light color shade
blue: value 1 is good
pink: value 0 is good

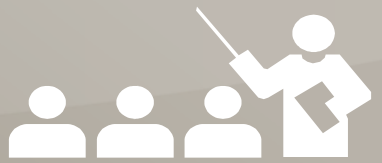
Map points from ROC to Precision-Recall or vice-versa: (TPR same values in both)

$$\mathrm{PPV} = \frac{\mathrm{P}\ \mathrm{TPR}}{\mathrm{P}\ \mathrm{TPR} + \mathrm{N}\ \mathrm{FPR}} \qquad \textit{(ROC to P-R)}$$

$$\mathrm{FPR} = \frac{\mathrm{P}\ (1 - \mathrm{PPV})\ \mathrm{TPR}}{\mathrm{PPV}\ (\mathrm{P} + \mathrm{N}\ \mathrm{TPR})} \qquad \textit{(P-R to ROC)}$$

Fuente: http://numerical.recipes/

Demo 03-G Evaluación de Modelos



Evaluación de Modelos

Como interpretar modelos

La necesidad de interpretar el modelo

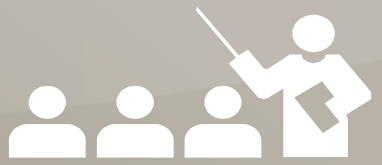
Motivaciones

- Identificar y mitigar sesgos
- Explicación del contexto del problema
- Mejorar la generalización y el rendimiento
- Razones éticas y legales

Cómo interpretar el modelo

- Importancia de las características
  - Generalised Linear models(GLM)
  - Random forest y SVM’s
  - Deep Learning
- LIME

Demo 03-G Intepretando Modelos



Importancia de Características

LIME

www.solidq.com

info@solidq.com