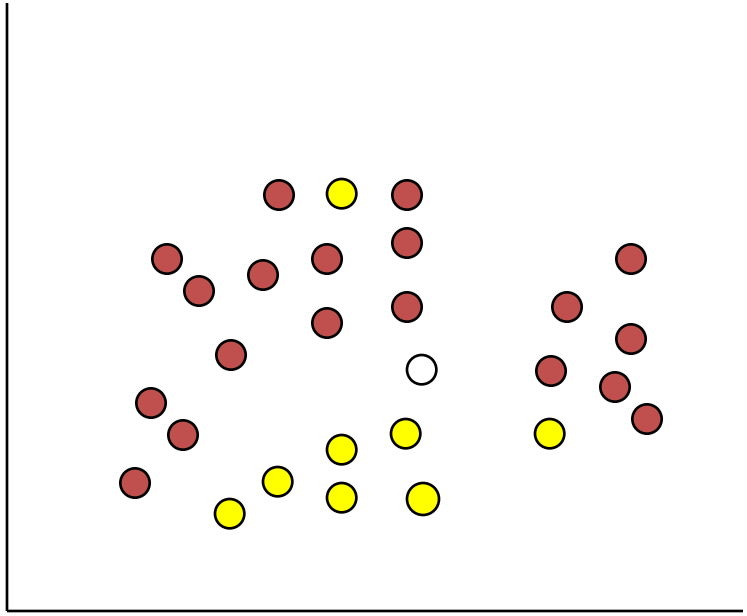
Aplicaciones de Ciencia de la Computación

Aprendizaje de Máquina: Clasificación KNN

Prof. Dr. César A. Beltrán Castañón
cbeltran@pucp.pe

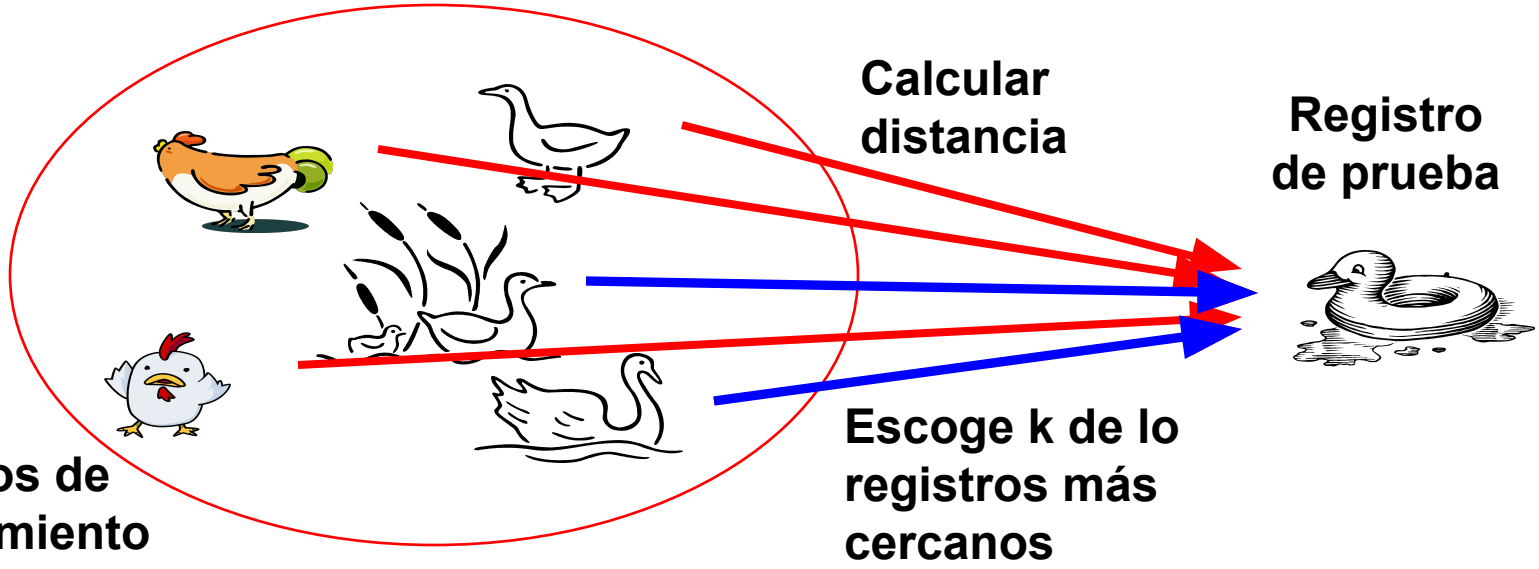
Clasificación KNN



Dado un conjunto de puntos de las clases, ● ●
cuál es la clase de un nuevo punto ○ ?

Clasificación KNN

Idea básica: Si camina como pato, parpa como pato, probablemente sea un pato.



Medidas de similitud (Distancia)

Un concepto importante en el aprendizaje supervisado (clasificación) y no supervisado (segmentación) es el concepto de similitud:

- La razón de este uso es que, intuitivametne, datos similares tendrán clases/grupos similares. ¿Cómo se mide la similitud?

- DISTANCIA inversa a SIMILITUD.

- Los métodos de similitud (o de distancia) se basan en almacenar los ejemplos vistos, y calcular la similitud/distancia del nuevo caso con el resto de ejemplos.

Medidas de similitud (Distancia)

- Muchísimas formas de calcular la distancia:
  - ***Distancia Euclídea:*** $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
  - ***Distancia de Manhattan:*** $\sum_{i=1}^{n}|x_i - y_i|$
  - ***Distancia de Chebychev:*** $max_{i=1..n}|x_i - y_i|$
  - ***Distancia del coseno:***
    cada ejemplo es un vector y
    la distancia es el coseno del ángulo que forman

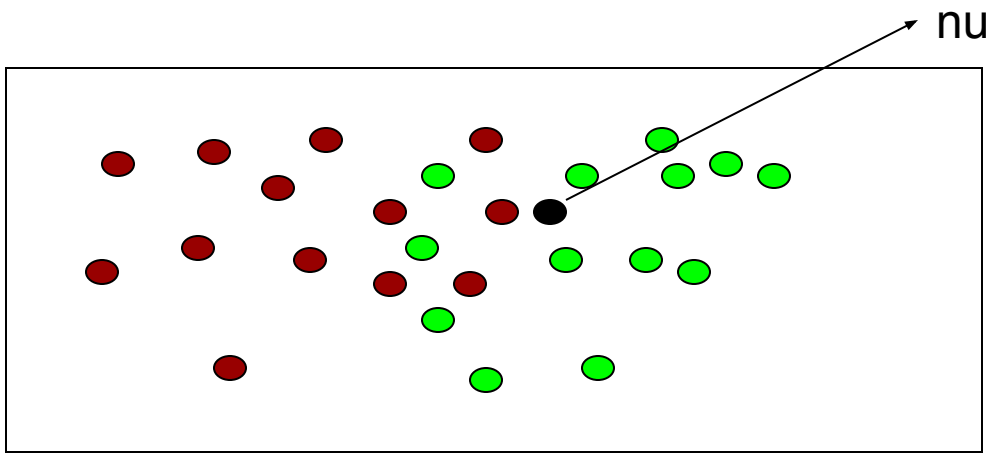
Valores Continuos (conveniente normalizar entre 0-1 antes)

Valores Continuos. No es necesario normalizar

Clasificación KNN

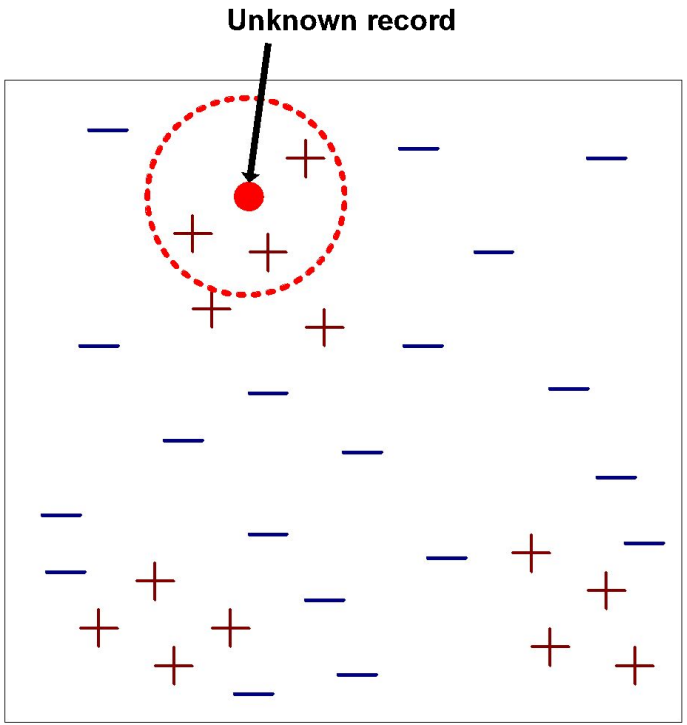
- Votación dentro de los k nearest neighbors

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$



K= 1: brown
K= 3: green

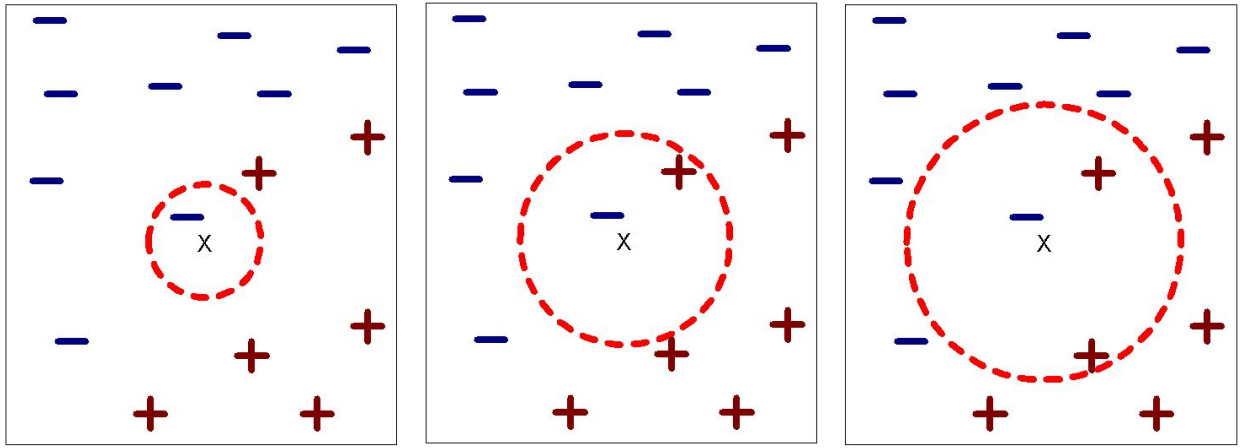
Clasificación KNN



- Requiere tres cosas
  - El conjunto de registros almacenados
  - Medida de distancia
  - El valor de *k*, el número de vecinos más próximos a recuperar

- Para clasificar un elemento desconocido:
  - Calcular la distancia a los otros elementos de entrenamiento
  - Identificar los k vecinos más próximos
  - Usar la etiqueta clase de los vecinos más próximos para determinar la clase del elemento desconocido (ej. por votación)

Definición de Nearest Neighbor

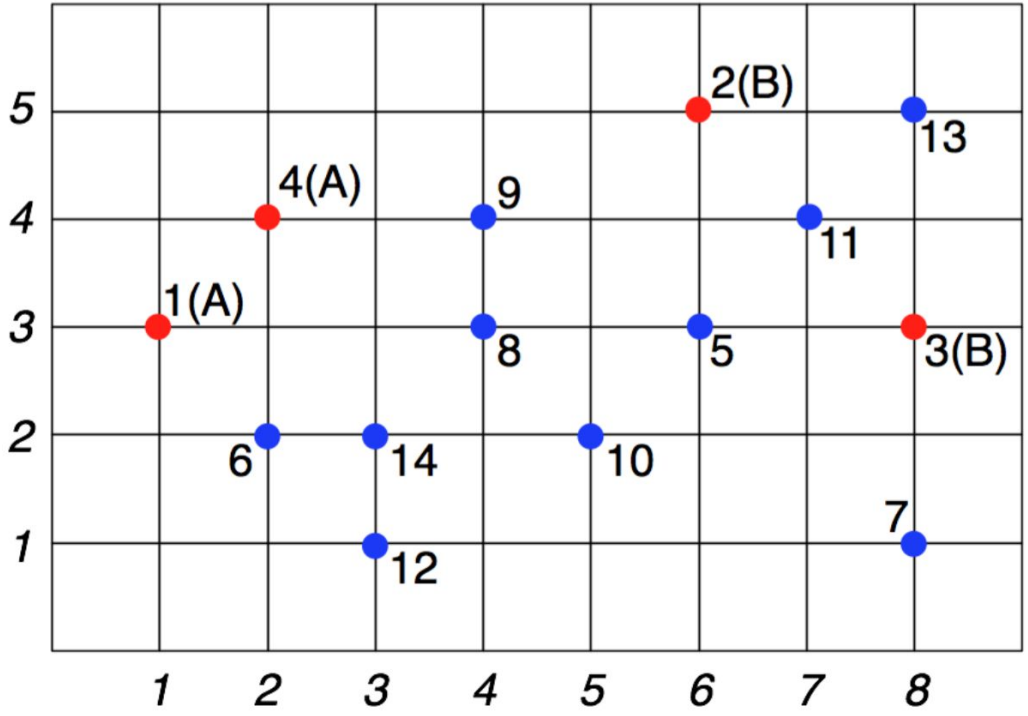


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors de un elemento x son los puntos de datos que tienen las k distancias más próximas a x

KNN - Ejemplo

En la figura siguiente, los puntos representan un conjunto de vectores de dimensión 2. Estos pertenecen a dos clases denominadas A y B. El orden de selección de los vectores está indicado por los índices situados al lado de cada punto. Los puntos 1 al 4 ya han sido clasificados (en rojo), por consiguiente, el interés es de clasificar los puntos restantes (a partir del 5, en azul). Aplicar el método kNN para k = 3.



Solución: 5B – 6A – 7B – 8A – 9A – 10A – 11B – 12A – 13B – 14A

KNN - Ejemplo

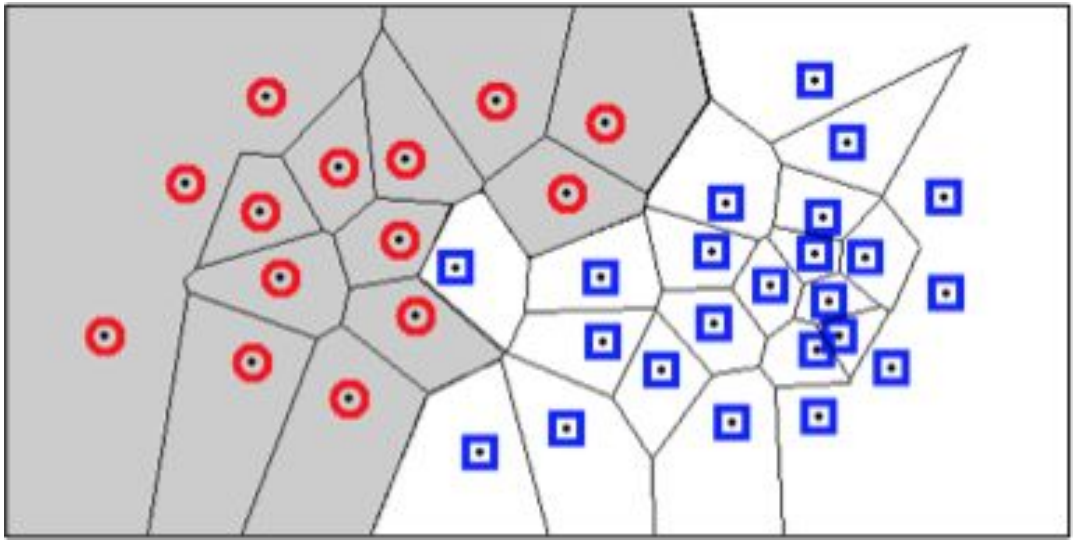
- En la figura anterior (u otra), demuestre con un ejemplo que el resultado de la clasificación depende del orden de cómo son leídos los datos
- ¿Qué puede pasar si k es par? Muestre este caso especial con un ejemplo en la figura anterior (u otra)

SOLUCIÓN:

En el gráfico anterior, si el vector lanzado en la posición 10 hubiera sido lanzado en 8, entonces ´ este hubiera sido clasificado como B y no como A, puesto que los 3 vecinos más próximos ya habrían sido clasificados como 5(B), 6(A) y 7(B)

Si k = 2 y el punto 9 habría sido tirado en la posición 4, este tendría como vecinos a 4(A) y 3(B). ¿Cuál clase elegir? el algoritmo elegiría uno de los dos a azar

Región de decisión con diagrama Voronoi para 1-NN



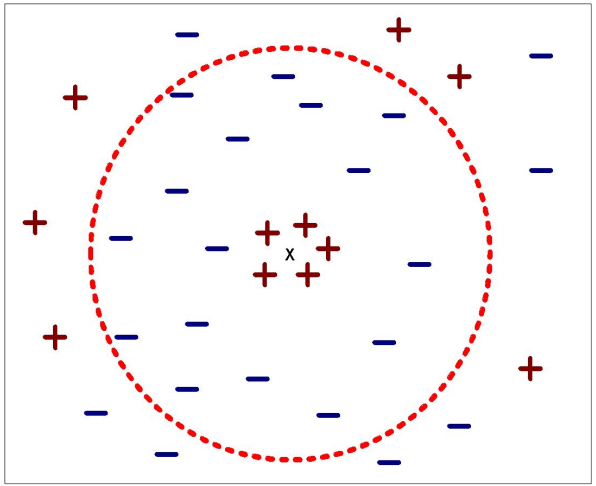
Cada celda contiene un ejemplo, y cada ubicación dentro de la celda es la más próxima a la de otro ejemplo.

Un diagrama de Voronoi divide el espacio en tale celdas

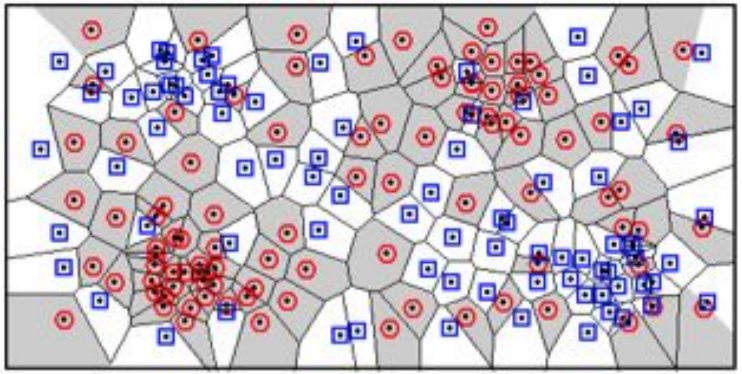
Clasificación Nearest Neighbor

Escogiendo el valor de k:

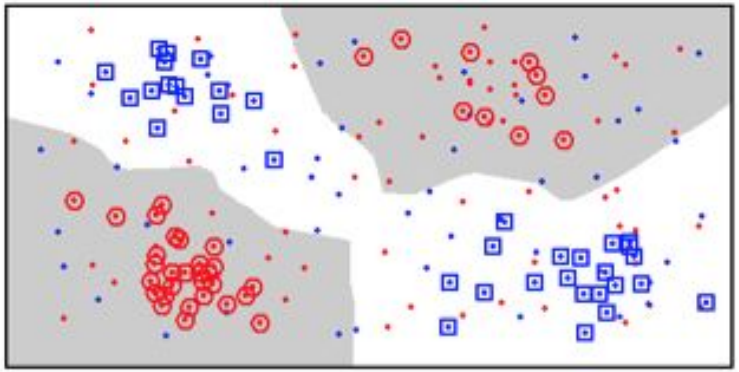
- Si k es muy pequeño, sensible a ruido
- If k is muy grande, se podrían considerar puntos de otras clases.



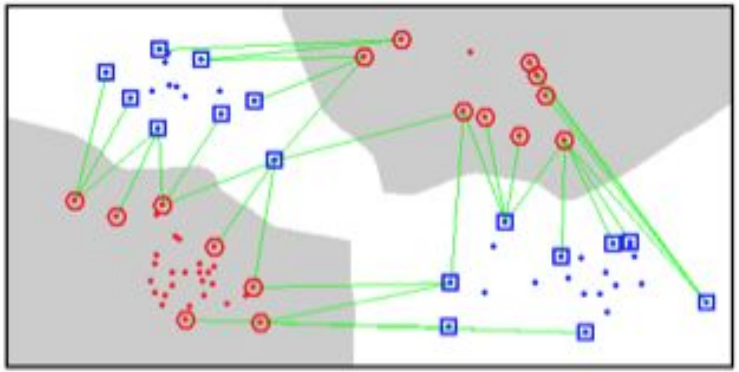
KNN, regiones de decisión complejas



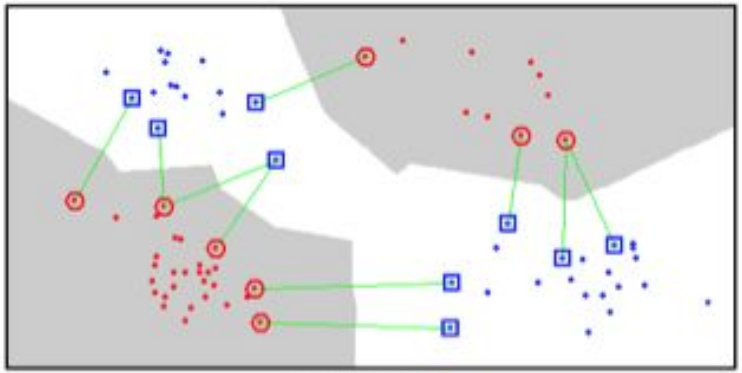
Original data: (99,85)



Multi-edit: (39,36)



Delaunay edited: (16,16)



Gabriel edited: (8,9)

KNN, prós y contras

- k-NN es un clasificador perezoso (lazy learners)
- A diferencia de los clasificadores ansiosos (eager learners) como árboles de decisión y sistemas basados en reglas, no construye los modelos explícitamente
- Clasificar elementos nuevos es relativamente caro
- Basado en conocimiento local (“hace que la data hable por ellos”), no es un modelo global para decidir.
- k-NN depende de la función de distancia; la calidad de esta es crítica para la performance del clasificador
- Es capaz de crear regiones de decisión complejas, consistente en las regiones formadas por los lados del diagrama de Voronoi.
- k-NN no calcula regiones de decisión, a diferencia de árboles de decisión y SVM
- k-NN obtiene buenas tasas de predicción y es muy popular en diferentes áreas como text data mining y recuperación de información.

KNN, implementación en sklearn

In [15]:
```
# machine learning
from sklearn.neighbors import KNeighborsClassifier
```

In [16]:
```
knn = KNeighborsClassifier(n_neighbors = 3)
knn.fit(X_train, y_train)

y_pred = knn.predict(X_test)
print(accuracy_score(y_test,y_pred)*100)
```

```
60.8938547486
```

Más información, consulte el manual de sklearn

http://scikit-learn.org/stable/modules/neighbors.html#classification