marzo de 2022

(Hitz Research Center), Departamento de LSI, EHU

Sist. de Ayuda a la Decisión:
Clasificación: KNN

AitZiber AtutXa
aitziber.atucha@ehu.eus

marzo de 2022

Índice

1. Objetivos de Aprendizaje
2. Intuición trás el Knn
3. Formalización del Knn
4. Calculo del Knn
5. Lectura y visionado de vídeos

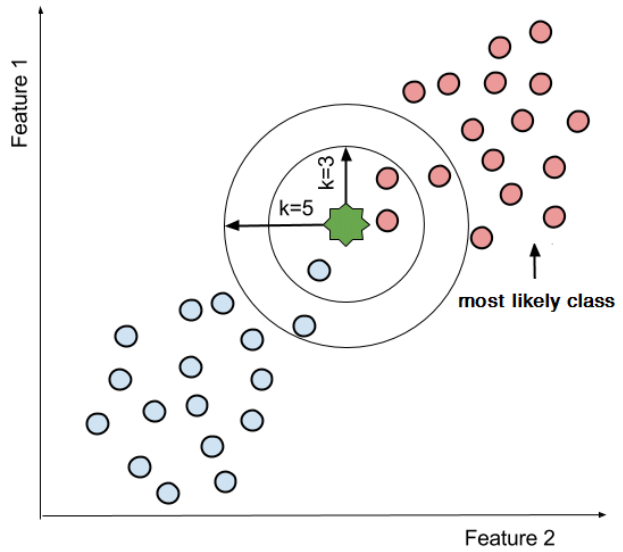
Objetivos de Aprendizaje

- Ser capaces de describir y explicar sus características: algoritmo de clasificación, Lazy y no paramentrizable.
- Ser capaces de formalizar matemáticamente el algoritmo.
- Ser capaces de generar un modelo básico de clasificación empleando el algoritmo KNN.
- Ser capaces de realizar un barrido de hiperparámetros en el KNN.
- Ser capaces de entender el significado de los hiperparámetros **k:** número de vecinos, **d:** distancia y **w:** peso asociado a cada vecino.

Intuición trás el Knn

Te conoceré por tus parecidos (vecinos): Algoritmo Geométrico

- Recoger las k instancias más próximas
- Votar por mayoría



Pregunta.

¿Deberían todos los votos tener el mismo peso?

□

imagen: Mohammad Reza

Formalización del Knn

Recordemos: $\mathcal{D}^{train}, \mathcal{D}^{dev}, \mathcal{D}^{test}$

$$\mathcal{X}^{train} = \{x^{\{1\}}, x^{\{2\}}, ...x^{\{n\}}\}, \mathcal{X}^{dev} = \{x^{\{n+1\}}, x^{\{n+2\}}, ...x^{\{l\}}\},$$
$$\mathcal{X}^{test} = \{x^{\{l+1\}}, x^{\{l+2\}}, ...x^{\{z\}}\}$$

n es tamaño de la muestra train, l-n+1 el tamanño del dev, z-l-n+1
(i.e. núm. instancias de entrenamiento, de desarrollo, de test)

$$Y^{train} = \{y^{d\{1\}}, y^{d\{2\}}, ...y^{d\{n\}}\}, Y^{dev} = \{y^{d\{n+1\}}, y^{d\{n+2\}}, ...y^{d\{l\}}\}$$
$$Y^{test} = \{y^{d\{l+1\}}, y^{d\{l+2\}}, ...y^{d\{z\}}\}$$

Las Y contienen los valores reales a *acertar*
Obviamente Y^{train} contiene n valores porque cada instancia de $\mathcal{X}^{train}$ tendrá su clase *real* asociada, lo mismo para el dev y test

$$\mathcal{D}^{train} = \{(x^{\{1\}}, y^{d\{1\}}), (x^{\{2\}}, y^{d\{2\}}), ...(x^{\{n\}}, y^{d\{n\}})\}$$
$$\mathcal{D}^{dev} = \{(x^{\{n+1\}}, y^{d\{n+1\}}), (x^{\{n+2\}}, y^{d\{n+2\}}), ...(x^{\{l\}}, y^{d\{l\}})\}$$
$$\mathcal{D}^{test} = \{(x^{\{l+1\}}, y^{d\{l+1\}}), (x^{\{l+2\}}, y^{d\{l+2\}}), ...(x^{\{z\}}, y^{d\{z\}})\}$$

Idealmente $\mathcal{D}^{train} \cap \mathcal{D}^{dev} \cap \mathcal{D}^{test} = \emptyset$

Formalización del Knn

k: número de vecinos con derecho a voto

$x^{\{t\}} \in \mathcal{D}^{train} \wedge \hat{x} \in [\mathcal{D}^{dev} | \mathcal{D}^{test}]$

$x^{\{t\}} = \{x_1^{\{t\}}, x_2^{\{t\}}, ..., x_n^{\{t\}}\}$ n: num. atrib. de cada instancia

$d(x^a, x^b)$ es la distancia entre la instancia a y b

Algorithm 1 Predicción-Knn(k,d$_{card}$, $\mathcal{D}^{train}$, $\hat{x}$)

$Aux = []$
for each $x^{\{t\}} \in \mathcal{D}^{train}$ **do**
  $Aux = Aux \oplus < d(x^{\{t\}}, \hat{x}), c >$#almacena dist. y clase de $x^{\{t\}}$
end for
$ordenar(Aux)$
$C\_Clases = []$ #cont. de clases
for $k = 1$ **to** K **do**
  $actualizar(C\_Clases, c^{\{k\}})$
end for
$clasePred = max(C\_Clases)$

¿Cómo calculamos la distancia?

$x_i^{\{t\}}$ representa el *i-abo* atributo de la instancia $x^{\{t\}}$

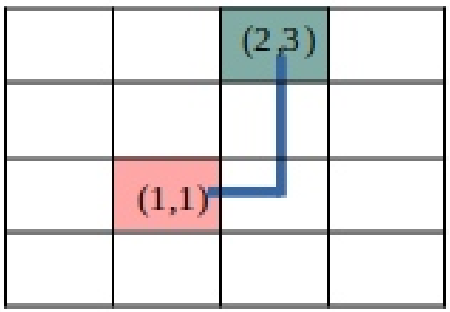
- Distancia de Minkowski $\left(\sum_{i=1}^{n} |x_i^{\{t\}}, \hat{x}_i|^p \right)^{\frac{1}{p}}$
  - No negatividad: d($x^{\{t\}}, \hat{x}$) >=0
  - Identidad: d($x^{\{t\}}, \hat{x}$)=0 si solo si $x^{\{t\}} == \hat{x}$
  - Simetría: d($x^{\{t\}}, \hat{x}$)=d($\hat{x}, x^{\{t\}}$)
  - Desigualdad triangular: d($x^{\{t\}}, \hat{x}$)+d($\hat{x}, z$) >=d(x,z)

¿Cómo calculamos la distancia?

Cuando la p=1, hablamos de Distancia de Manhatan. Nota[2]

$\sum_{i=1}^{n} |x_i^{\{t\}}, \hat{x}_i|$ **nota** valor abs |...|

- los puntos son (1,1) y (2,3)
- distManh: $(2-1)+(3-1)=3$



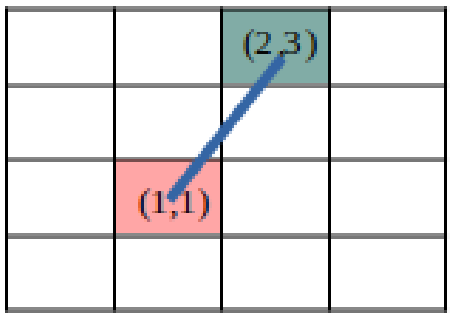
[2] $x_i^{\{t\}}$ representa el *i-abo* atributo de la instancia $x^{\{t\}}$

¿Cómo calculamos la distancia?

Cuando la p=2, hablamos de Distancia Euclídea

$$\left(\sum_{i=1}^{n} |x^{\{t\}}, \hat{x}|^2 \right)^{\frac{1}{2}}$$

- los puntos son (1,1) y (2,3)
- distEucl:
  $\sqrt{(2-1)^2 + (3-1)^2} = \sqrt{5} = 2{,}23$



¿Cómo calculamos la distancia?

Ejercicio: ¿Cúal es la clase del siguiente libro?

Identifica la clase del libro $\hat{x} = (230, 10)$ dada la siguiente muestra de entrenamiento, k=1 y d=Manhatan:

- Para ello comienza por dibujar las instancias en un plano

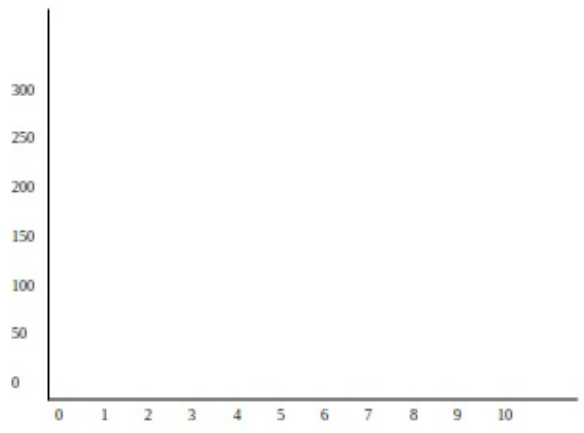
BestSeller	FreqCrime	NumPag
0	1	250
0	0	200
1	9	150
1	10	200
0	1	200

Cuadro: Muestra de entrenamiento

□

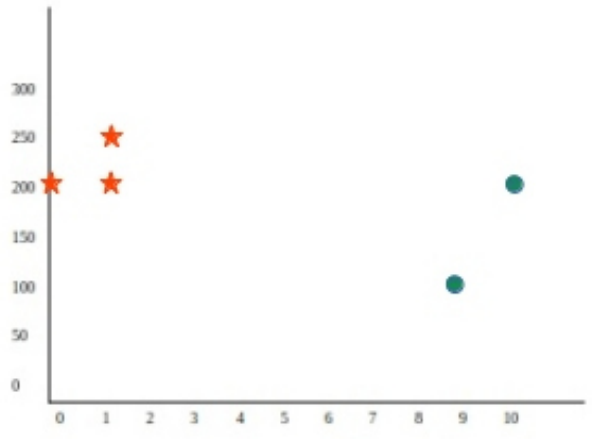
¿Cómo calculamos la distancia?

Añadir las instancias de entrenamiento en el gráfico

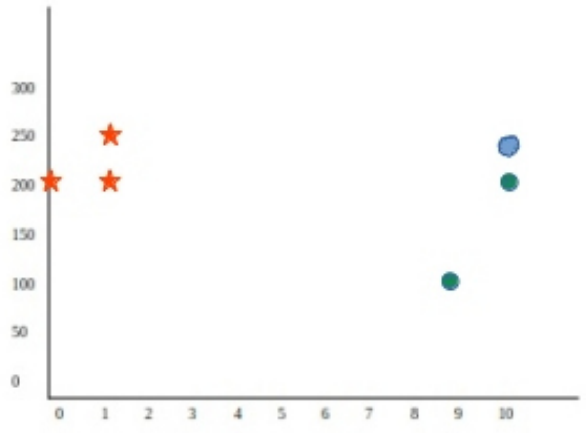


¿Cómo calculamos la distancia?

Ahora, añadir la instancia $x^{\{t\}}$ en el anterior gráfico:



¿Cómo calculamos la distancia?



Pregunta:

¿No véis ningún problema con la escala? □

¿Cómo calculamos la distancia?

Ejercicio: ¿Cúal es la clase del siguiente libro?

Calcula la clase del libro $\hat{x} = (230, 10)$ dada la siguiente muestra de entrenamiento, k=1 y d=Manhatan:

BestSeller	NumPag	FreqCrime	DistManh($\hat{x}$)
0	250	1	??
0	200	0	40
1	100	10	130
1	200	9	??
0	200	1	??

Cuadro: Muestra de entrenamiento

La clase de $\hat{x} = (230, 10)$ con k=1 y sin escalar sería 0

□

¿Cómo calculamos la distancia?

Ejercicio: ¿Cúal es la clase del siguiente libro?

Calcula la clase del libro $\hat{x} = (230, 10)$ dada la siguiente muestra de entrenamiento, k=1 y d=Manhatan:

BestSeller	NumPag	FreqCrime	DistManh($\hat{x}$)
0	250	1	29
0	200	0	40
1	150	10	130
1	200	9	31
0	200	1	39

Cuadro: Muestra de entrenamiento

La clase de $\hat{x} = (230, 10)$ con k=1 y sin escalar sería 0

□

¿Cómo calculamos la distancia?

Ejercicio: ¿Qué pasa si escalamos empleando z-core?

Calcula la clase del libro $\hat{x} = (230, 10)$ escalado (0,73,1) dada la siguiente muestra de entrenamiento, k=1 y d=Manhatan:

BestS	FCrime	NumPg	CriEsc	PgEsc	DManh($x^{\{t\}}$)
0	1	250	-0,66	1,09	2,03
0	0	200	???	????	???
1	10	150	1,20	-1,64	2,58
1	9	200	???	?	???
0	1	200	-0,66	0,18	2,21

Cuadro: Muestra de entrenamiento

La clase de $\hat{x} = (230, 10)$ con k=1 y escalando sería 1

□

MediaPg 190 y Std 54,77 y MediaFrecCrime 4,2 y Std 4,8

¿Cómo calculamos la distancia?

Ejercicio: ¿Qué pasa si escalamos empleando z-core?

Calcula la clase del libro $\hat{x} = (230, 10)$ escalado (0,73,1) dada la siguiente muestra de entrenamiento, k=1 y d=Manhatan:

BestSeller	NumPag	FreqCrime	DistManh($\hat{x}$)
0	1,09	-0,66	2,03
0	0,18	-0,87	2,42
1	-1,64	1,20	2,58
1	0,18	1	0,54
0	0,18	-0,66	2,21

Cuadro: Muestra de entrenamiento

La clase de $\hat{x} = (230, 10)$ con k=1 y escalando sería 1

□

¿Cómo calculamos la distancia?

Ejercicio: ¿Cúal es la clase del siguiente libro?

Calcula la clase del libro $\hat{x} = (230, 10)$ dada la siguiente muestra de entrenamiento, k=3 y d=Manhatan:

BestSel	NPg	FqCrime	NPgEsc	CriEsc	DistManh
0	250	1	????	????	????
0	200	0	????	????	????
1	150	10	-0,822	1,341	1,81
1	200	9	????	????	????
0	200	1	????	????	????
0	100	1	-1,957	-0,853	5,14
1	190	7	0,085	-0,853	3,10
1	200	7	0,312	-0,853	2,87

Cuadro: Muestra de entrenamiento

La clase de $\hat{x} = (230, 10)$ con k=3 y escalando sería $0, 9930, 1, 34$ 1

□

MediaPg: 186,25 y Std: 44,058 MediaFreqCrime: 4,5 y Std: 4,1

¿Cómo calculamos la distancia?

Ejercicio: ¿Cúal es la clase del siguiente libro?

Calcula la clase del libro $\hat{x} = (230, 10)$ dada la siguiente muestra de entrenamiento, k=3 y d=Manhatan:

BestSel	NPg	FqCrime	NPgEsc	CriEsc	DistManh
0	250	1	1,446	-0,853	2,65
0	200	0	0,312	-1,097	3,11
1	150	10	-0,822	1,341	1,815
1	200	9	0,312	1,097	0,92
0	200	1	0,312	-0,853	2,87
0	100	1	-1,957	-0,853	5,14
1	190	7	0,085	-0,853	3,10
1	200	7	0,312	-0,853	2,87

Cuadro: Muestra de entrenamiento

La clase de $\hat{x} = (230, 10)$ escalado $0,9930, 1,34$ con k=3 y escalando sería 1

¿Cómo calculamos la distancia?

Preguntas par ala reflexión.

¿Dirías que el escalado es primordial en las técnicas geométricas?

¿Crees que el Knn asigna la misma importancia a todos los atributos?

¿Funcionaría con valores faltantes?

¿Crees que es importante que K sea impar?

□

Mejorando el Knn

- La clase estrella sería la predicha con k=5
- Asignando pesos inversos a d a los k vecinos su voto no sería uniforme sino ponderado
- El resultado varía

Fase de Entrenamiento: Obtención hipótesis

Parametros en dataiku y en la librería Sklearn de Python

Dataiku:

- K: número de vecinos que votan
- Distance Weighting: clickable
- p: si 1 manhatan, si 2 euclidea

Sklearn:

- n_neighbours: es la k
- weights: uniform (todos igual), distance (inversamente proporcional a la distancia)
- p: igual que en Dataiku

Lecturas y vídeos recomendados

- "Data Minning: Practical Machine Learning Tools and Techniques"Witten et. al
- Visionado del vídeo del
  https://www.youtube.com/watch?v=mpU84OJ5vdQ