CLASIFICACIÓN BASADA EN SIMILARIDADES, K-NN

Abraham Sánchez López
FCC/BUAP
Grupo MOVIS

MOVIS

Facultad de Ciencias de la Computación

El algoritmo de k vecinos más cercanos

- Creo que las cosas simples de la vida son las mejores: jugar Frisbee en el parque, pasear a mi perro, jugar juegos de mesa con mi familia y usar el algoritmo k-NN.
- Algunos profesionales del aprendizaje automático desprecian un poco a k-NN porque es muy simplista.
- De hecho, k-NN es posiblemente el algoritmo de aprendizaje automático más simple, y esta es una de las razones por las que gusta tanto.
- A pesar de su simplicidad, k-NN puede proporcionar un rendimiento de clasificación sorprendentemente bueno y su simplicidad hace que sea fácil de interpretar.
- **Nota**. Recuerda que, debido a que k-NN usa datos etiquetados, es un algoritmo de aprendizaje supervisado.

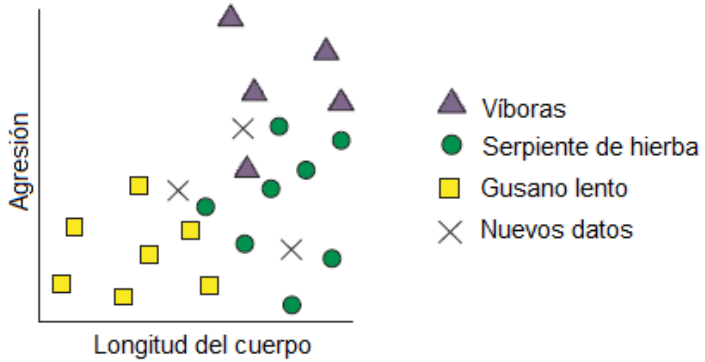
Como funciona el algoritmo, I

- Entonces, ¿cómo aprende k-NN? Bueno, voy a usar serpientes para ayudarme a explicar. Soy de México, donde, algunas personas se sorprenden al saber, tenemos algunas especies nativas de serpientes.
- Dos ejemplos son la culebra y la víbora, que es la única serpiente venenosa de México (hay cosas más interesantes en nuestra amplia geografía).
- Pero también tenemos un lindo reptil sin extremidades llamado gusano lento, que comúnmente se confunde con una serpiente.
- Imagina que trabajas para un proyecto de conservación de reptiles con el objetivo de contar el número de culebras, víboras y gusanos lentos en un bosque.
- Tu trabajo es construir un modelo que te permita clasificar rápidamente los reptiles que encuentres en una de estas tres clases.
- Cuando encuentra uno de estos animales, solo tienes el tiempo suficiente para estimar rápidamente su longitud y una medida de cuán agresivo es hacia ti, antes de que se escape (los fondos son muy escasos para tu proyecto).

Como funciona el algoritmo, II

- Un experto en reptiles te ayuda a clasificar manualmente las observaciones que has realizado hasta ahora, pero decides crear un clasificador k-NN para ayudarte a clasificar rápidamente los especímenes futuros que encuentres. Mira la gráfica de datos antes de la clasificación en la figura 1.

- Cada uno de nuestros casos se grafica contra la longitud del cuerpo y la agresión, y la especie identificada por tu experto se indica mediante la forma del dato.
- Vuelve al bosque y recopila datos de tres nuevos especímenes, que se muestran con las cruces negras.



- Figura 1: Longitud del cuerpo y agresión de los reptiles. Los casos etiquetados para víboras, culebras y gusanos lentos se indican por su forma. Los datos nuevos, sin etiquetar, se muestran con cruces negras.

Como funciona el algoritmo, III

- Podemos describir el algoritmo k-NN (y otros algoritmos de aprendizaje automático) en términos de dos fases:

1. La fase de entrenamiento
2. La fase de predicción

- La fase de entrenamiento del algoritmo k-NN consiste únicamente en almacenar los datos.

- Esto es inusual entre los algoritmos de aprendizaje automático (como aprenderás durante el curso), y significa que la mayor parte del cálculo se realiza durante la fase de predicción.

- Durante la fase de predicción, el algoritmo k-NN calcula la distancia entre cada nuevo caso sin etiquetar y todos los casos etiquetados.

- Cuando se dice “distancia”, nos referimos a su cercanía en términos de las variables de agresión y longitud del cuerpo, ¡no a qué tan lejos en el bosque los encontraste!

Como funciona el algoritmo, IV

- Esta métrica de distancia a menudo se denomina distancia euclidiana, que en dos o incluso tres dimensiones es fácil de visualizar mentalmente como la distancia en línea recta entre dos puntos en un gráfico (esta distancia se muestra en la figura 2). Esto se calcula en tantas dimensiones como estén presentes en los datos.

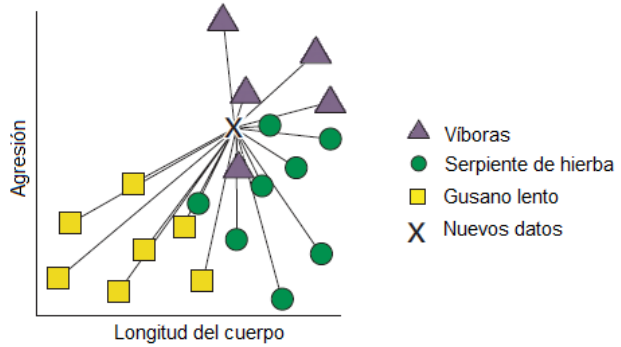
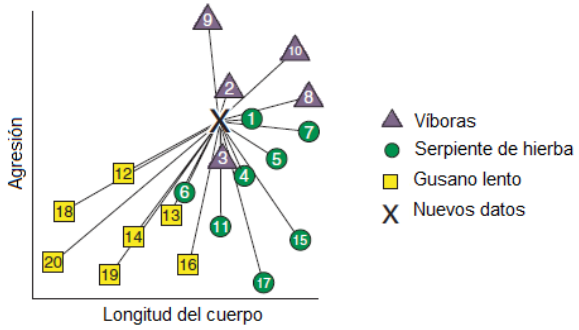


Figura 2. El primer paso del algoritmo k-NN: calcular la distancia. Las líneas representan la distancia entre uno de los casos sin etiquetar (la cruz) y cada uno de los casos etiquetados.

- A continuación, para cada caso sin etiquetar, el algoritmo clasifica a los vecinos desde el más cercano (el más similar) hasta el más lejano (el menos similar). Esto se muestra en la figura 3.

Figura 3. El segundo paso del algoritmo k-NN: clasificar a los vecinos. Las líneas representan la distancia entre uno de los casos sin etiquetar (la cruz) y cada uno de los casos etiquetados. Los números representan la distancia clasificada entre el caso sin etiquetar (la cruz) y cada caso etiquetado (1 = más cercano).

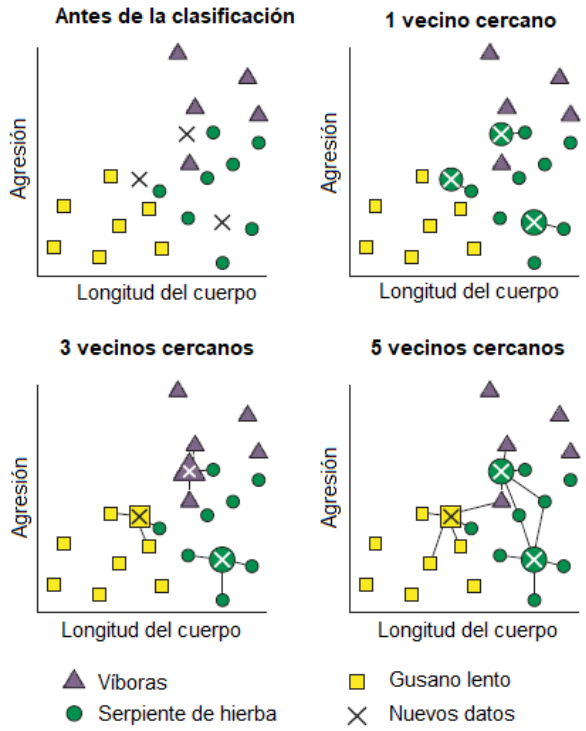


Como funciona el algoritmo, V

- El algoritmo identifica los casos etiquetados con k (vecinos) más cercanos a cada caso no etiquetado. k es un número entero especificado por nosotros (cubriremos posteriormente cómo elegimos k).
- En otras palabras, encuentra los casos etiquetados con k que son más similares en términos de sus variables al caso no etiquetado. Finalmente, cada uno de los casos del vecino más cercano "vota" en qué clase pertenecen los datos no etiquetados, según la clase propia del vecino más cercano.
- En otras palabras, cualquier clase a la que pertenezcan la mayoría de los k-vecinos más cercanos es como se clasifica el caso sin etiqueta.
- **Nota**. Debido a que todos sus cálculos se realizan durante la fase de predicción, se dice que k-NN es un aprendiz perezoso (*lazy learner*).
- Analicemos la figura 4 y veamos esto en la práctica. Cuando establecemos k en 1, el algoritmo encuentra el único caso etiquetado que es más similar a cada uno de los elementos de datos no etiquetados.

Como funciona el algoritmo, VI

- Cada uno de los reptiles sin etiqueta es el más cercano a un miembro de la clase de serpientes de hierba, por lo que todos están asignados a esta clase.



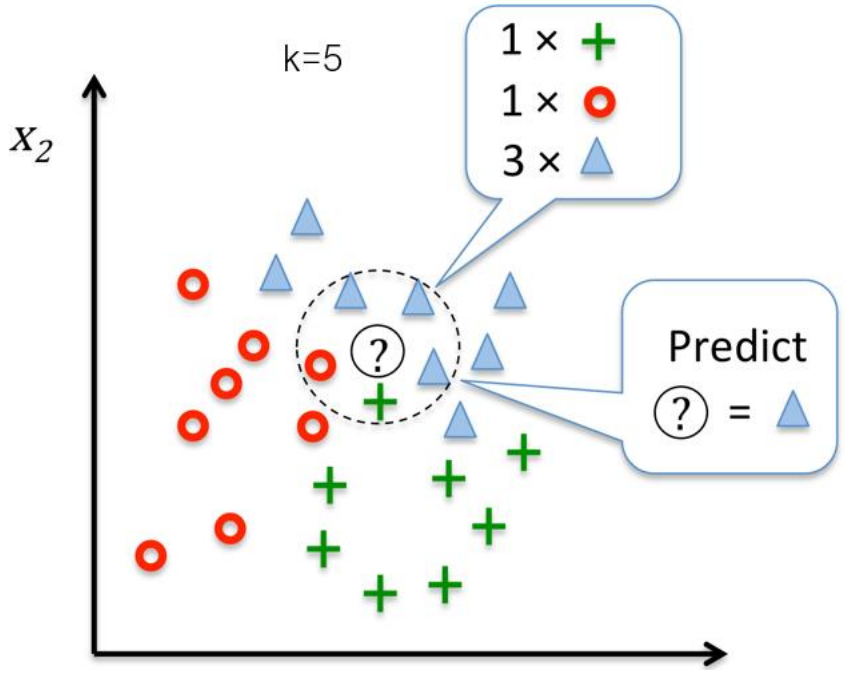
- Figura 4. El paso final del algoritmo k-NN: identificar los k-vecinos más cercanos y obtener el voto mayoritario. Las líneas conectan los datos sin etiquetar con sus vecinos uno, tres y cinco más cercanos. El voto de la mayoría en cada escenario está indicado por la forma dibujada debajo de cada cruz.

Como funciona el algoritmo, VII

- Cuando establecemos k en 3, el algoritmo encuentra los tres casos etiquetados que son más similares a cada uno de los elementos de datos no etiquetados.
- Como puedes ver en la figura, dos de los casos sin etiquetar tienen vecinos más cercanos que pertenecen a más de una clase.
- En esta situación, cada vecino más cercano “vota” por su propia clase y gana el voto de la mayoría.
- Esto es muy intuitivo porque si una sola serpiente de hierba inusualmente agresiva resulta ser el vecino más cercano a una víbora que aún no ha sido etiquetado, será superado en votos por las víboras vecinas en los datos.
- Esperemos que ahora puedas ver cómo esto se extiende a otros valores de k. Cuando establecemos k en 5, por ejemplo, el algoritmo simplemente encuentra los cinco casos más cercanos a los datos sin etiquetar y toma el voto mayoritario como la clase del caso sin etiquetar.
- Ten en cuenta qué, en los tres escenarios, el valor de k afecta directamente cómo se clasifica cada caso sin etiquetar.

Como funciona el algoritmo, VIII

- **Consejo**. ¡El algoritmo k-NN en realidad se puede usar tanto para problemas de clasificación como de regresión!, pero la única diferencia es que, en lugar de tomar el voto de la clase mayoritaria, el algoritmo encuentra la media o la mediana de los valores de los vecinos más cercanos.



¿Qué pasa si se empata la votación?, I

- Puede suceder que todos los k-vecinos más cercanos pertenezcan a diferentes clases y que la votación resulta en un empate. ¿Qué sucede en esta situación?
- Bueno, una forma en que podemos evitar esto en un problema de clasificación de dos clases (cuando los datos solo pueden pertenecer a uno de dos, grupos mutuamente excluyentes) es asegurar que elijamos números impares de k.
- De esta manera, allí siempre será un voto decisivo.
- Pero, ¿qué pasa en situaciones como nuestro problema de clasificación de reptiles, donde tenemos más de dos grupos?
- Una forma de lidiar con esta situación es disminuir k hasta que se pueda obtener una mayoría de votos ganadora.
- Pero esto no ayuda si un caso sin etiqueta es equidistante entre sus dos vecinos más cercanos.

¿Qué pasa si se empata la votación?, II

- En cambio, un enfoque más común (y pragmático) es asignar casos aleatoriamente sin mayoría de votos a una de las clases.
- En la práctica, la proporción de casos que tienen lazos entre sus vecinos más cercanos es muy pequeño, por lo que tiene un impacto limitado en la precisión de clasificación del modelo.
- Sin embargo, si tienes muchos vínculos en tus datos, tus opciones son las siguientes:
- Elegir un valor diferente de k.
- Agregar una pequeña cantidad de ruido a los datos.
- Considerar el uso de un algoritmo diferente!