29/04/2024

Regresión logística

Crea una recta y le aplica la función sigmoide para romper la linealidad y que se ajuste mejor a los resultados. Se utiliza para problemas de clasificación, en lugar de la regresión lineal que es para problemas de regresión

kNN

Clasificación en base al vecino más cercano (Nearest neighbour). La ‘k’ define el número de vecinos más cercanos. Este algoritmo no tiene apenas entrenamiento por lo que es muy rápido. Sus principales desventajas son los posibles empates que pueden ocurrir si encuentra el mismo número de vecinos. Otro problema es saber el número de ‘k’.

Para utilizar este modelo los datos han de estar normalizados o estandarizados

Para saber la k adecuada se puede utilizar la formula k = raíz(n)

Medidas de distancia

Distancia euclídea: Pitágoras

Distancia manhattan: sumando catetos. Útil para problemas de navegación con el coche

Distancia Minkowski: un punto medio entre euclídea y manhattan

06/05/2024

Modelo SVM

Crea una recta que separa las dos clases. Utiliza la técnica del vector soporte para buscar la línea mas ancha entre las dos clases.

Cuando los datos estas perfectamente separados, se utiliza un separador de margen duro. Se los datos están solapados se debe aplicar un clasificador de margen blando

Datos linealmente no separables

Cuando los datos no se pueden separar linearmente, se utiliza el truco kernel. Se agrega un polinomio para que los datos pasen de ser rectos a una curva

13/05/2024

Arboles de decisión

Modelos predictivos formados por reglas binarias (si/no). La complicado de este modelo es saber como hacer las condiciones. Las condiciones se establecen a través del Código gini. El árbol para de bifurcar cuando el índice Gini es 0.

En ocasiones que el índice Gini sea 0 no es positivo, ya que puede dar sobreajuste. A veces, no es rentable que todos los nodos tengan el índice a 0

Hacer las prediciones

El modelo empieza a evaluar el dato a través de la profundidad 0 y a partir de ahí le aplica las reglas que se aplican

Para calcular el índice Gini se resta 1 a la probabilidad de acertar (cuantas clases aciertan con ese nodo). Es la impureza del nodo

Algoritmo de entrenamiento CART

Coge todos los atributos y hace particiones de los atributos para saber cual es el mejor Gini. Si los atributos son categóricos coge uno de los posibles valores y bifurcas a partir de este.

Es un algoritmo de tipo voraz

Condiciones de parada

El funcionanmiento estándar es bifurcar hasta alcanzar el Gini perfecto en los nodos hoja. Sin embargo, para que generalice mejor y sea mas eficiente a nivel computacional se pueden ajustar los siguientes hiperparametros.

* Max\_depth: profundidad máxima
* Min\_samples\_splitt: no dividir un nodo si tiene menos de un numero dado
* Min simples\_leaf:
* Min\_weight\_fraction\_left
* Max\_features

Los arboles de decisión también se pueden utilizar para problemas de regresión, devolviendo la media de los valores de la hoja

Ensamblaje

Ensamblaje consiste en utilizar varios modelos al mismo tiempo. Se utiliza un sistema de votación, eligiendo lo que la mayoría de modelos decida. Esta técnica es conocida como hard voting.

Otra técnica conocida como soft voting, se queda con la probabilidad mas alta

Baggins y pasting

Se entrenan varios modelos iguales pero con datos de entrenamiento diferentes. Baggins utiliza remplazo (deja una copia) y pasting sin remplazo (el dato se saca del dataset)

Random forest

Ensablamje de arboles de decisión

Extra-Trees

Ensamble de bosques extremadametne aleatorios. Coge todas las características y todas sus validaciones de forma aleatoria. Son bastante mas rápido.

Boosting

La salida de uno es la entrada de otro