Ejercicio para entregar 1

Minería de datos II-Universidad Francisco de Vitoria

Pablo Hidalgo (pablo.hidalgo@ufv.es)

En este ejercicio puedes profundizar en el camino emprendido en clase sobre algunas de las propiedades del modelo de k vecinos más cercanos.

1 Normas

Algunas de las sesiones tendrán asociados unos ejercicios para entregar. **A lo largo de la segunda parte de la asignatura deberás entregar al menos uno de estos ejercicios para entregar**. Servirán para la nota de clase.

Aunque puedes entregar más de uno, **la nota final será la mediana de los ejercicios entregados.**

Cada apartado del ejercicio tendrá (*), (**), (***) o (****). Para obtener una determinada nota, deberás responder correctamente a todos los ejercicios marcados de forma que

- (*): para alcanzar una nota en $[0, 5]$. Serán ejercicios de aplicación directa de lo visto en clase. Se consideran ejercicios con los contenidos mínimos que hay que conocer para aprobar la asignatura.
- (**): para alcanzar una nota en $(5, 7]$. Son ejercicios en los que se debe demostrar que, además de conocerse, se entienden los conceptos clave.
- (***): para alcanzar una nota en $(7, 9]$. Se evaluarán críticamente los conceptos.
- (****): para obtener una nota en $(9, 10]$. Se pedirá ir un paso más allá tanto en los conceptos como en la búsqueda de comandos para R.

No se puede alcanzar una nota superior si no se han contestado bien a (casi) todos los ejercicios marcados de la nota inferior.

Puedes utilizar las funciones dentro del script `aux.R` que hemos utilizado en clase.

2 Formato de entrega

La entrega se debe realizar en un **documento `.pdf`** antes de la fecha límite que aparece en Canvas.

BAJO NINGUNA CIRCUNSTANCIA SE ACEPTARÁ UNA ENTREGA QUE HAYA SIDO SUBIDA FUERA DE LA FECHA LÍMITE.

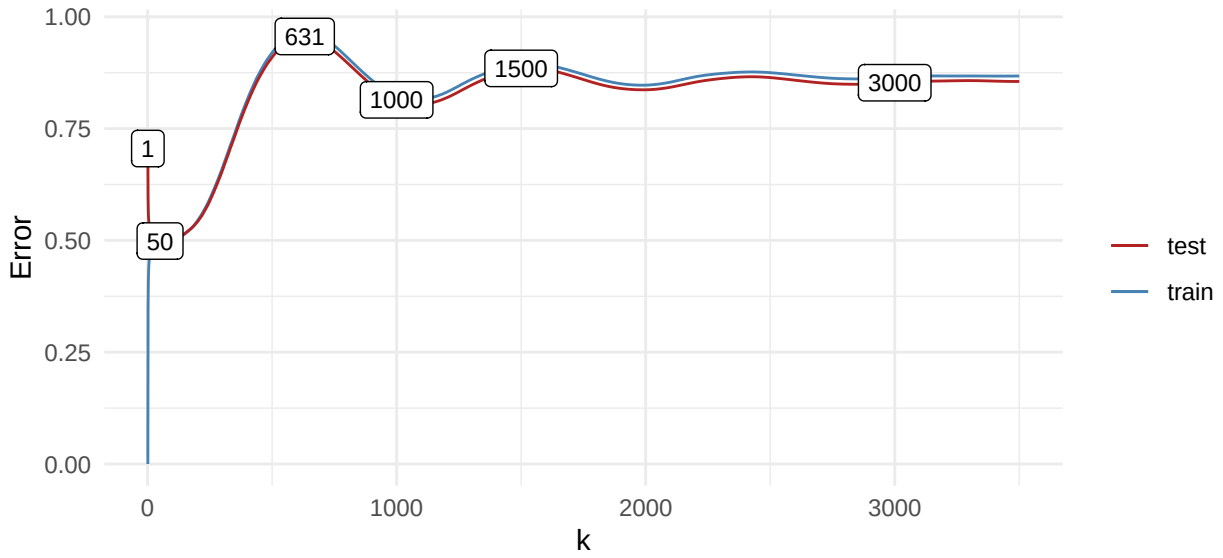
En el documento debes contestar a las preguntas lo más detalladamente posible apoyándote en los gráficos siempre que sea necesario. Aunque para realizar la práctica debes utilizar R, **en el documento no incluyas ninguna parte de código.**

3 Ejercicios

1. (*) Carga los dos conjuntos de datos `train.csv` y `test.csv`. Representa gráficamente el conjunto de datos de entrenamiento. El gráfico debe tener como título "Representación gráfica del conjunto de train". Describe el patrón que siguen los datos.

Las siguientes preguntas se refieren al comportamiento que tendría un modelo de KNN entrenado sobre `train.csv`.

Si calculamos los RMSE para el conjunto de entrenamiento y test para cada valor de k obtenemos el siguiente gráfico:



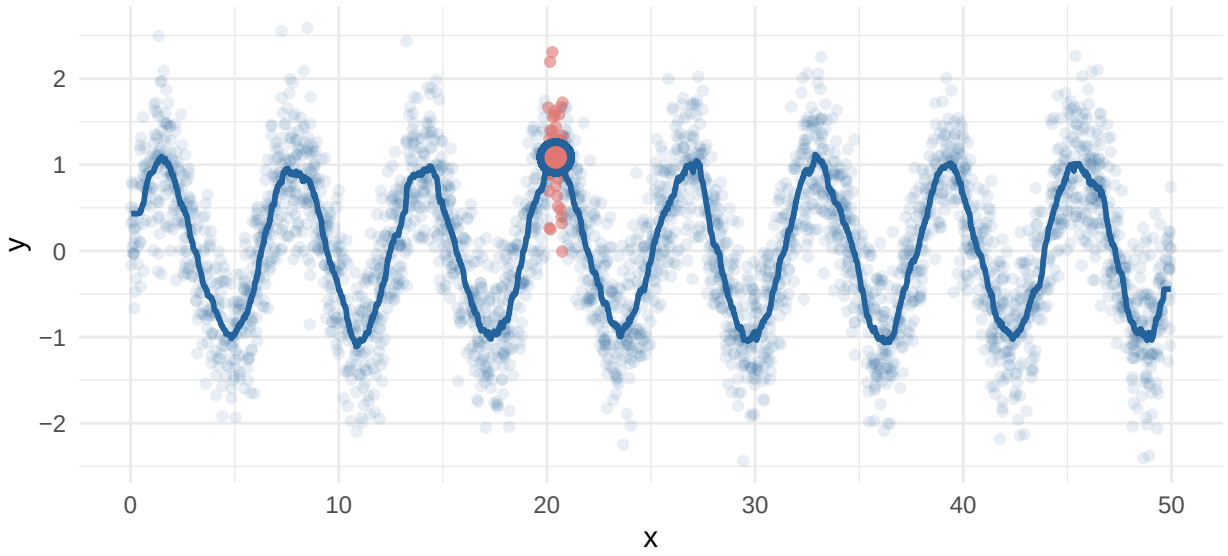
3. (*) Describe el gráfico anterior indicando en qué valores de k el modelo presenta *overfitting* o *underfitting*. ¿Por qué el gráfico termina en $k = 3500$?

4. (**) Calcula el RMSE cometido en train con $k = n$. Calcula la desviación típica de la variable objetivo Y en el conjunto de train. ¿Por qué son prácticamente iguales?

En el gráfico de las curvas de error vemos algunas particularidades. En el modelo que hemos visto en clase las curvas siempre tenían una tendencia estrictamente creciente. Sin embargo, en este gráfico se aprecian *montañas* en los errores. En el gráfico están marcadas algunas zonas interesantes. Los números que aparecen en los recuadros se corresponden con el valor de k.

Podemos representar cómo se comporta el modelo en función de los vecinos. En el gráfico siguiente se muestra:

- La **línea azul** representa la predicción que daría el modelo de k vecinos con $k = 50$ para los valores de la variable predictora X.
- En color coral se muestran los 50 vecinos más cercanos al punto resaltado ($X = 20.42$) y que son los puntos utilizados para realizar la predicción (recuerda que se hace simplemente la media).



El gráfico anterior se realiza mediante la sentencia `plot_neighbors(50, train_x, train_y, x = 20.42)` (la función `plot_neighbors()` está dentro de `aux.R`).

5. (**) Representa el mismo gráfico anterior para los puntos resaltados en el gráfico de las curvas. ¿Las predicciones de qué gráficos son manifiestamente erróneas?¿Por qué?

6. (***) A la vista de los gráficos del apartado anterior, ¿por qué aparecen esas *montañas* en el gráfico de las curvas de error?

7. (****) Construye un conjunto de test compuesto de 2000 observaciones de forma que:

    1. La variable predictora X sea un valor aleatorio entre -10 y 60 (utiliza la función `runif()`).
    2. La variable objetivo Y se calcule como `y <- sin(x) + error` siendo error un valor de la distribución normal de media 0 y desviación típica 0.5.

8. (****) Utiliza la función `info_knn()` de `aux.R` para representar la predicción para el nuevo conjunto de test a partir de los datos de entrenamiento utilizando un k vecinos con $k = 50$. Fíjate en cómo es la predicción para $x < 0$ y $x > 50$. ¿Por qué sucede esto?