Análisis Avanzado de Datos

Taller 1

Instrucciones

- La calificación se dará sobre 100 puntos y el trabajo se desarrolla preferiblemente en parejas.
- El trabajo podrá ser recibido de manera posterior a la fecha acordada con una penalización en la nota como sigue:
  1. Entregar el trabajo 1 día después: la calificación obtenida se dará sobre 80 de los 100 puntos totales. Nota máxima: 80 puntos.
  2. Entregar el trabajo 2 días después: la calificación obtenida se dará sobre 60 de los 100 puntos totales. Nota máxima: 60 puntos.
  3. Entregar el trabajo 3 días después: la calificación obtenida se dará sobre 40 de los 100 puntos totales. Nota máxima: 40 puntos.
  4. Entregar el trabajo 4 días después: la calificación obtenida se dará sobre 20 de los 100 puntos totales. Nota máxima: 20 puntos.
- Una persona de la pareja debe enviar el informe o la URL de su repositorio publico junto a la página web de su trabajo al correo andresn.lopez@urosario.edu.co **el día 05 de Mayo antes de la media noche**. Un minuto posterior a media noche de la fecha acordada es considerado como el día siguiente a la entrega y será calificado de acuerdo al punto anterior.

Consideraciones Estas consideraciones se verán reflejadas positivamente en la nota en forma de bonificación **en caso de presentar una nota por debajo de la nota máxima**:

- El formato de entrega del trabajo puede ser un archivo word guardado en formato pdf, pero se recomienda a los estudiantes a realizar un cuaderno en rmd: vea un ejemplo: https://anlopezl.github.io/AED/NB1_E.html
- Se invita a los estudiantes a versionar su trabajo en git y trabajar en equipo usando la plataforma. Si se trabaja individualmente, este punto no será considerado en la bonificación.
- El estudiante puede publicar su trabajo mediante github pages y enviar la URL correspondiente.

Problema 1 - 80 pts. El conjunto de datos `Auto` en la librería `ISLR2`, utilizado en clase, contiene la información del rendimiento y otras variables para un total de 392 vehículos. Como nos dimos cuenta, la relación entre dos de sus variables (`horsepower` y `mpg`) es resumida de manera parsimoniosa mediante un polinomio global de grado 2, sin embargo un spline suavizado (*smoothing spline*) parece dar un menor error de predicción. Por otra parte, determinar la ubicación y cantidad de knots en el spline de regresión (*regression spline*) fue un problema que desincentivó su uso. El método de validación externa utilizado para comprar los modelos fue **validación regular**.

(1) Separe aleatoriamente (pero guarde la semilla) su conjunto de datos en dos partes:

- Entrenamiento: 90 % de los autos.
- Prueba: 10 % de los autos.

(2) **Usando los datos de entrenamiento** Mediante validación cruzada en 10 folds, determine el número óptimo de knots para el problema de regresión spline. Considere como número de posible de knots 1,...,10, igualmente espaciados en el rango de la variable `horsepower`. ¿Qué modelo (es decir, cual valor de knot con $k = 1, ..., 10$) resulta en un menor ECM de predición?

(3) **Usando los datos de entrenamiento, determine el mejor modelo basado en base de funciones** Compare el poder de predicción de los modelos: polinomio grado 2 global, spline suavizado y del modelo de regresión spline óptimo (encontrado en el punto anterior) utilizando validación cruzada en 10 folds. ¿Cuál de los tres modelos seleccionaría basado en el ECM de predición?.

(4) **Usando los datos de entrenamiento, determine el mejor modelo basado en regresión local** Determine la regresión polinomial local con kernel gaussiano que resulte en menor error de predicción: regresión de grado 1 o 2. Use el ancho de banda óptimo dado por defecto por la función `loess()`.

(5) **Usando los datos de entrenamiento y de prueba, determine el mejor de los tres paradigmas de modelamiento** Ajuste el mejor modelo basado en base de funciones, el mejor modelo basado en regresión local y un polinomio global de grado dos con los datos de entrenamiento y calcule el ECM de prueba para cada modelo.

(6) Repita (1) - (5) un total de 10 veces de manera que en el paso (1) conforme una nueva muestra de validación cruzada, esto le permitirá obtener 10 ECM de prueba para cada paradigma de modelamiento. Grafique las tres distribuciones del ECM de prueba y responda ¿Cuál acercmiento seleccionaría basado en el ECM de predición: basado en base de funciones, basado en regresión local o polinomial global?.

Ayudas opcionales: los ejercicios pueden ser solucionados de múltiples formas. A continuación se presentan ayudas completatente opcionales. Algunas de ellas están orientadas hacia el lenguaje R, pero esto no implica que el trabajo deba hacerse en dicho lenguaje:

(1) Puede ser conveniente definir los folds de manera aleatoria como `sample(1:10,n,replace=TRUE)` para facilitar la validación cruzada.

(2a) Tiene que ajustar en cada fold los 10 posibles knots, y probar su error en el conjunto de prueba. Por lo tal, tendría un total de 100 errores externos al hacer validación cruzada: 10 por cada uno de los 10 posibles knots.

(2b) `splines::bs` permite encontrar la base b-spline (aquella base 'multiproposito' mencionada en clase) para usar en el problema de regresión.

(2c) `splines::bs` en su argumento `Boundary.knots` permite definir los knots de frontera. Puede dejarlos unas unidades alejadas del rango para asegurar un ajuste apropiado, por ejemplo `Boundary.knots = range(horsepower) + c(-5,+5)`.

(3) Tiene que ajustar en cada fold los 3 modelos, y probar su error en el conjunto de prueba. Por lo tal, tendría un total de 30 errores externos al hacer validación cruzada: 10 por cada uno de los 3 modelos.

(4) Es altamente recomendado utilizar `ksmooth` en lugar de `loess` para el estimador de Nadarya–Watson, sin embargo esto dificulta el proceso de predicción de manera poco interesante. Considere únicamente los grados 1 y 2 en su decisión (note que esto ha sido cambiado en el cuerpo de la pregunta).

Problema 2 - 20 pts. En el contexto de análisis de datos funcionales se tiene una colección finita de observaciones ruidosas, donde para cada individuo, estas se asumen provenientes de una curva de dimensión infinita la cual es evaluada en puntos de un intervalo determinado. Para la i-ésima unidad estadística se tiene un conjunto de n_i observaciones discretizadas $x_{i1}, ..., x_{ij}, ..., x_{in_i}$ de la función x_i en los puntos t_{i1}, ..., t_{ij} , ..., t_{in_i} con $x_{ij} \in R$, $t_{ij} \in T$ y T un intervalo que representa el dominio sobre los reales donde se definen los datos funcionales.

(7) Escriba el estimador de Nadarya–Watson para la i-ésima unidad estadística en t, es decir, $x(t)$.

La centralidad de los datos funcionales se resume en la **función media** μ, la cual puede interpretarse en cada valor $t \in T$ como el valor promedio de la función aleatoria subyacente en t, $\mu(t)$. Fíjese que el estimador de Nadarya–Watson puede extenderse a más de una unidad estadística, resultando en t como un promedio ponderado de las observaciones cercanas para todas las observaciones x_{ij}:

(8) Escriba el estimador de Nadarya–Watson para la función media en t, es decir, $\hat{\mu}(t)$. Note que todos los datos discretizados son utilizados en la estimación de la función media.