Taller Evaluado N° 2
Herramientas Estadísticas y Forecast (HEF)

Facultad de Matemáticas - Departamento de Estadística
Pontificia Universidad Católica de Chile

2024

Contenido I

Regresión lineal

Introducción

En los sistemas de bicicletas compartidas, el proceso de arriendo y devolución es automatizado. Normalmente, las bicicletas se pueden arrendar en un lugar y regresar en otro sin tener que depender de humanos.

Actualmente, existen varios programas para compartir bicicletas en diferentes ciudades.

El data set **bikes.xlsx**, contiene 731 registros diarios entre 2011-01-01 y 2012-12-31, en Washington D.C., EE.UU., de las siguientes variables:

Regresión lineal

Introducción

- `date`: Fecha en formato año-mes-día.
- `season`: Estación del año [1 (invierno), 2 (primavera), 3 (verano) y 4 (otoño)].
- `year`: Año [2011 y 2012].
- `month`: Mes [1 a 12].
- `holiday`: 1 (día festivo) y 0 (día no festivo).
- `weekday`: Día de la semana [1 (lunes), 2 (martes), . . . , 7 (domingo)].
- `workingday`: 1 (día laboral), 0 (día no laboral).
- `weather`: Estado del clima [1: soleado o nubosidad parcial, 2: nublado, 3: lluvioso, 4: tormentoso]

Regresión lineal

Introducción

- `temp`: Temperatura normalizada en grados Celsius.

  Los valores se obtienen mediante $\frac{t-tmin}{tmax-tmin}$, con $tmin = -8$ y $tmax = +39$.
- `atemp`: Temperatura de sensación normalizada en grados Celsius.

  Los valores se obtienen mediante $\frac{t-tmin}{tmax-tmin}$, con $tmin = -16$ y $tmax = +50$.
- `humidity`: Humedad normalizada (% de humedad dividido por 100).
- `windspeed`: Velocidad normalizada del viento (velocidad del viento en millas por hora dividido por 67).
- `registered`: Número de alquileres de bicicletas ese día por usuarios registrados.
- `target`: Cantidad total de arriendos de bicicletas ese día, incluidos los casuales y los usuarios registrados.

Regresión lineal

Pregunta 1

Realice los siguientes procesos al data set:

- Transforme a factor las variables que son categóricas y asigne etiqueta. Ej: 1 $\rightarrow$ lunes.
- Transforme las variables `temp` y `atemp` a grados celsius, ya que se encuentran en escala normalizada.
- Transforme la variable humedad a porcentaje.
- Transforme la velocidad del viento a millas por hora.

Regresión lineal

Pregunta 2

- Realice un gráfico de dispersión entre el número de arriendos de bicicletas (`target`) vs temperatura (`temp`).
- Agregue la recta de regresión lineal.
- Comente brevemente.

Regresión lineal

Pregunta 3

¿Es la relación entre la temperatura y el número de bicicletas arrendadas igual en los dos años?, para responder, compare en un mismo gráfico la relación entre el número de arriendo vs temperatura para los dos años, añada ambas rectas de regresión lineal.

Regresión lineal

Pregunta 4

Por selección forward, construya un modelo sin considerar las variables

- `date`
- `month`
- `registered`

Interpreta el factor asociado a la Temperatura y días feriados.

Regresión lineal

Pregunta 5

Utilice los residuos del modelo elegido para estudiar la validez de los supuestos:

- Normalidad.
- Homocedasticidad.

Comente.

Regresión lineal

Pregunta 6

Realice una predicción de arriendos de bicicletas para un día con las siguientes cualidades:

- season: 2.
- year: 2011.
- holiday: día festivo.
- weekday: sábado.
- workingday: no laboral
- weather: nublado.
- temp: 12.
- atemp: 11.
- humidity: 66.3.
- windspeed: 12.5.

DATA SCIENCE UC
FACULTAD DE MATEMÁTICAS

Series de Tiempo

Pregunta 7

Cuando las mediciones presentan un orden temporal, como es el caso de este data set, los residuos del modelo de regresión usualmente presentan estructura de auto-correlación que implica que el supuesto de independencia de las observaciones no se cumpla.

- A partir de el ACF verifique que el supuesto de independencia dno se cumple y junto al PACF proponga los ordenes p y q de un potencial modelo ARMA para los residuos del modelo de regresión ajustado en la pregunta 5.
- Utilizando auto.arima() de forecast de R o su equivalente en Python obtenga un modelo ARMA a partir de los ordenes propuestos en el item anterior.
- Realice el test de Box-Ljung y chequee si la hipótesis de blancura se cumple.

DATA SCIENCE UC
FACULTAD DE MATEMÁTICAS

Regresión Logística

Introducción

En un centro meteorológico, se contratan sus servicios como Data Scientist para construir un modelo que prediga si lloverá o no en las próximas 24 horas, utilizando información de las 24 horas previas.

El data set **lluvia.xlsx** contiene un conjunto de 19 variables meteorológicas.

Regresión Logística

Introducción

La descripción de las variables es la siguiente:

- `MinTemp`: Temperatura mínima registrada.
- `MaxTemp`: Temperatura máxima registrada.
- `Lluvia`: Cantidad de lluvia registrada ese día en mm.
- `Evaporacion`: Evaporación (mm) en 24 horas.
- `Sol`: Número de horas de sol brillante en el día.
- `VelRafaga`: La velocidad (km/h) de la ráfaga de viento más fuerte en las 24 horas hasta la medianoche.
- `Vel9am`: La velocidad (km/h) de la ráfaga de viento a las 9am.
- `Vel3pm`: La velocidad (km/h) de la ráfaga de viento a las 9am.
- `Hum9am`: Porcentaje de humedad a las 9am.
- `Hum3pm`: Porcentaje de humedad a las 3pm.

Regresión Logística

Introducción

- `Pres9am`: Presión atmosférica (hpa) a nivel del mar a las 9am.
- `Pre3pm`: Presión atmosférica (hpa) a nivel del mar a las 3pm.
- `Nub9am`: Fracción del cielo cubierto por nubes a las 9am. Se mide en "octavos", de manera que un valor 0 indica cielo. totalmente despejado y 8, cielo totalmente cubierto.
- `Nub3pm`: Fracción del cielo cubierto por nubes a las 3pm. Se mide en "octavos", de manera que un valor 0 indica cielo. totalmente despejado y 8, cielo totalmente cubierto.
- `Temp9am`: Temperatura en grados celsius a las 9am.
- `Temp3pm`: Temperatura en grados celsius a las 3pm.
- `LluviaHoy`: Variable indicadora que toma el valor 1 si la precipitación en mm. en las últimas 24 horas. excede 1 mm. y 0, si no.
- `Koppen`: Clasificación Koppen de la zona de medición (Temperate, Subtropical, Grassland, Tropical, Desert). Estación Estación del Año.
- `LluviaMan`: Indicador de lluvia al día siguiente de la medición (No y Yes).
- `Estacion`: Estación del año.

Regresión Logística

Pregunta 1

, codifique la variable `LluviaMan` como 0 y 1 para los días sin lluvia y con lluvia respectivamente.

Además, realice una separación de la base de datos en un set de entrenamiento y set de validación. Utilice una proporción de 80:20 respectivamente.

Para poder replicar sus resultados, fije una semilla antes de obtener los indices. Para ello utilice la función `random.seed(2024)` o `set.seed(2024)` dependiendo si lo hace en Python o R.

Regresión Logística

Pregunta 2

Realice un modelo de regresión logística para predecir si lloverá mañana utilizando la variable Evaporación, ¿es este Un factor significativo? Intérprete el odd ratio de la evaporación.

Regresión Logística

Pregunta 3

Utilizando un método automatizado, ajuste un modelo de regresión logística, utilizando la metodología de dirección both (forward y backward a la vez).

Regresión Logística

Pregunta 4

Considerando la base de entrenamiento, ajuste la curva ROC y KS asociada al modelo, ¿Qué puede concluir sobre la discriminación del modelo?.

Con la información obtenida encuentre un punto de corte que tenga una sensibilidad mínima del 80% y la máxima especificidad.

Regresión Logística

Pregunta 5

Considerando la base de test, obtenga nuevamente la curva ROC y KS asociada al modelo, ¿cómo han variado los indicadores?, además, utilizando el punto de corte obtenido, obtenga la precisión.

Regresión Logística

Pregunta 6

Utilizando el punto de corte encontrado, determine si el día de mañana lloverá.

- `MinTemp`: 7
- `MaxTemp`: 18
- `Lluvia`: 0
- `Evaporacion`: 7
- `Sol`: 12
- `VelRafaga`: 72
- `Vel9am`: 10
- `Vel3pm`: 54
- `Hum9am`: 65
- `Hum3pm`: 77
- `Pres9am`: 1001
- `Pre3pm`: 1025
- `Nub9am`: 3
- `Nub3pm`: 2
- `Temp9am`: 11.4
- `Temp3pm`: 16.2
- `LluviaHoy`: No
- `Koppen`: Subtropical
- `Estacion`: Primavera