FACULTAD DE CIENCIAS SOCIALES

Docente: Josué Cox (a20062953@pucp.edu.pe)

Fecha de Entrega: Hasta el miércoles 24 de enero a las 11:59 p.m.

Primer Trabajo Práctico

Para este trabajo deben enviar un archivo pdf y un notebook de Jupyter (.ipynb) que muestre todos los pasos de sus respuestas.

EJERCICOS TEÓRICOS

Pregunta 1 (2 puntos): Para los siguientes enunciados, indique si se esperaría un mejor o peor desempeño (en términos del *trade-off* de varianza y sesgo) de un método de aprendizaje estadístico más flexible o menos flexible

a. El tamaño de la muestra es grande y el número de predictores es pequeño
b. El número de predictores es grande y el tamaño de la muestra es grande
c. La relación entre los predictores y la variable de respuesta es no-lineal
d. La varianza del término de error, $\sigma^2 = Var(\epsilon)$ es bastante alta

Pregunta 2 (2 puntos): Grafique en una sola figura (puede usar cualquier editor de gráficos, ppt, o cualquier otro programa que prefieran) cinco curvas para un problema en donde el MSE se minimiza con una flexibilidad media:

- Sesgo al cuadrado
- Varianza
- MSE en el conjunto de datos de entrenamiento
- MSE en el conjunto de datos de prueba
- El error irreducible

En el eje-**x** debe estar la flexibilidad y en eje-**y**, los valores de cada una de estas curvas

Pregunta 3 (3 puntos): En esta pregunta entenderemos de forma simple cómo funciona el método de aproximación de K-vecinos-cercanos (KNN) en un entorno de clasificación. Tenemos seis observaciones de una muestra donde hay cuatro predictores y una variable de respuesta cualitativa (si el individuo está desempleado o no)

Obs.	Edad	Educación	Proprietario Casa	Número de hijos	Estado Laboral
1	18	1	0	0	Empleado
2	25	2	0	1	Empleado
3	36	2	0	2	Desempleado
4	45	3	1	2	Empleado
5	50	2	0	0	Empleado
6	60	3	1	4	Desempleado

Donde educación toma valor 1 si el individuo solo fue a primaria, toma valor 2 si fue a secundaria y toma valor 3 si tiene educación superior. La variable propietario casa toma el valor 0 si la persona no posee una casa (alquila) o valor 1 si es propietaria.

Queremos saber si un individuo con 48 años, nivel educativo superior, que posee una casa y tiene 3 hijos tendrá más probabilidad de estar desempleado o empleado.

a. Encuentre la distancia euclidiana[1] de cada observación con respecto al individuo que queremos considerar
b. ¿Cuál es la predicción si usamos un KKN con K=1? ¿Por qué?
c. ¿Cuál es la predicción si usamos un KKN con K=3? ¿Por qué?

Pregunta 4 (3 puntos): Supongamos que tenemos una muestra con 5 predictores. X_1 es la nota promedio del estudiante, X_2 es el resultado de un test que mide el IQ del estudiante, X_3 es el nivel educativo (1 es universitario y 0 es secundario), X_4 es una interacción entre X_1 y X_2, X_5 es una interacción entre X_1 y X_3. La variable de interés es el salario después de terminar la universidad en miles de soles. Tenemos los resultados al estimar un modelo con MCO:

$$\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$$

a. ¿Qué respuesta es verdadera? ¿Y por qué?
    i. Para un nivel dado de IQ y de notas, los graduados de secundaria ganan más, en promedio, que los graduados universitarios
    ii. Para un nivel dado de IQ y de notas, los graduados universitarios ganan más, en promedio, que los graduados de secundaria

[1] La distancia euclidiana entre dos vectores $x, y \in R^4$ se define como $d(x,y) = (x_1 - y_1)^2 + \cdots + (x_4 - y_4)^2$

iii. Para un nivel dado de IQ y de notas, los graduados de secundaria ganan más, en promedio, que los graduados universitarios siempre que sus notas son mayores
iv. Para un nivel dado de IQ y de notas, los graduados universitarios ganan más, en promedio, que los graduados de secundaria siempre que sus notas son mayores

b. Predecir el salario de un graduado universitario con un IQ de 110 y un nivel de notas de 4.0
c. Diga si es cierto que como el coeficiente de la interacción entre notas e IQ es pequeño, existe poca evidencia de algún efecto de dicha interacción

EJERCICOS PRÁCTICOS

Pregunta 6 (5 puntos): Usando la API del Banco Central de Reserva del Perú, construya una base de datos (Pandas) con las siguientes variables para cada frecuencia de datos. Genere gráficos para cada una de ellas (Tip: pueden usar un *for loop* para generar los gráficos).

Frecuencia Diaria:

1. Tasa interbancaria (S/.)
2. Rendimiento del Bono del gobierno peruano a 10 años (en S/)
3. TC Interbancario (S/ por US$) – Compra
4. TC Interbancario (S/ por US$) – Venta
5. TC Euro (S/ por Euro) – Compra
6. TC Euro (S/ por Euro) – Venta
7. Índice General Bursátil BVL (índice)
8. Índice Selectivo Bursátil BVL (índice)
9. Monto negociado en acciones (millones S/) - Promedio diario
10. Cobre (Londres, cUS$ por libras)
11. Plata (H.Harman, US$ por onzas troy)
12. Zinc (Londres, cUS$ por libras)
13. Oro (Londres, US$ por onzas troy)
14. Petróleo (West Texas Intermediate, US$ por barriles)
15. Bonos del Tesoro EE.UU. - 5 años (%)

16. Bonos del Tesoro EE.UU. - 10 años (%)
17. Bonos del Tesoro EE.UU. - 30 años (%)
18. Spread - EMBIG Perú (pbs)
19. Spread - EMBIG America Latina (pbs)
20. Dow Jones (var%)

Frecuencia Mensual:

1. Exportaciones totales (Valores FOB en millones de US$)
2. Importaciones totales (Valores FOB en millones de US$)
3. Circulante (millones S/.)
4. Preferencia por Circulante (millones S/.)
5. Indicadores de las empresas bancarias - Utilidad acumulada - Empresas Bancarias (millones S/)
6. Liquidez de las empresas bancarias - Liquidez (millones S/)
7. Liquidez de las empresas bancarias - Coeficiente de Dolarización de la Liquidez (%)
8. Índice de Precios al Consumidor (var% mensual)
9. Índice de Precios al Consumidor Sin Alimentos y Energía (var% mensual)
10. Índice de Precios al Consumidor Alimentos y Energía (var% mensual)
11. Índice de Precios al Consumidor Subyacente (var% mensual)
12. Índice de Precios al Consumidor No Subyacente (var% mensual)
13. Índice de Precios al por Mayor (var% mensual)
14. Expectativa de Inflación a 12 meses
15. Expectativa de PBI a 12 meses
16. Expectativa de TC a 12 meses
17. Producto bruto interno (índice 2007=100)
18. Producto bruto interno (variaciones porcentuales anualizadas)

Frecuencia Trimestral:

1. *Índice de precios hedónicos de inmuebles*
2. *Términos de Intercambio (2007=100)*
3. *Producto bruto interno (índice 2007 = 100)*
4. *Demanda Interna - Consumo Privado (millones 2007)*
5. *Demanda Interna - Consumo Público (millones 2007)*

6. *Demanda Interna - Inversión Bruta Interna (millones 2007)*
7. *Demanda Interna - Exportaciones (millones 2007)*
8. *Demanda Interna - Importaciones (millones 2007)*
9. *Venta de energía eléctrica total (gwh)*
10. *Producción de energía eléctrica total (gwh)*

Pregunta 7 (5 puntos): Para esta pregunta usará la base de datos ("AsientoBebe"). Esta base de datos contiene datos de ventas de asientos para bebes para carros en 400 mercados. Entre los predictores que tenemos en la base de datos se enumeran: (i) el precio promedio de la competencia [CompPrice], (ii) el ingreso promedio en miles el mercado o localidad [Income], (iii) el gasto en publicidad en cientos [Advertising], (iv) el número de personas en el mercado en miles [Population], (v) el precio de venta en dólares [Price], (vi) si la localidad o mercado es urbana o no [Urban], (vii) si el asiento se produce o no en los EEUU [US], entre otros.

1. Estime un modelo de regresión múltiple para predecir las ventas [Sales] usando Price, Urban y US
2. De una interpretación de cada uno de los coeficientes del modelo teniendo en cuanta que hay variables cualitativas
3. ¿Para cuáles predictores se puede rechazar la hipótesis nula de que $H_0: \beta_j = 0$?
4. Estime un nuevo modelo usando solo las variables en la que los predictores tengan una asociación significativa con las ventas
5. ¿Qué tan bien los modelos en (1) y (4) ajustan los datos en términos de R^2_{adj}?
6. Divida la muestra en una de entrenamiento y una de prueba o control. Estime los modelos en (1) y (4) y presente el MSE en la muestra de entrenamiento y prueba. ¿Cuál modelo usted elegiría?