Ejercicio 3

Profesores: Sebastián Espinosa, Marcos Orchard
Auxiliar: Cristóbal Allendes

Ayudantes: Fernando Palma, José Pablo Araya, Sebastian Guzman, Simón Vidal

Instrucciones: Pueden realizar el ejercicio en grupos de 5 personas, individualizando correctamente a cada integrante en su entrega. No se permite el uso de procesadores de texto como MS Word o LaTeX, sino que deben realizar su desarrollo a mano alzada o digitalizado en un dispositivo que permita el uso de lápiz electrónico. Los gráficos pueden ser generados en computador y anexados a su entrega.

P1 - Pregunta teórica

1. Considere una secuencia de variables aleatorias $\{X_n\}_{n\geq 1}$ con segundo momento finito, las cuales son tales que

$$\lim_{n\to\infty} \mathbb{E}\{X_n\} = \mu \quad \lim_{n\to\infty} \mathrm{Var}\{X_n\} = 0., \tag{1}$$

con $\mu \in \mathbb{R}$. Demuestre que

$$\mathbb{E}\Big\{(X_n - \mu)^2\Big\} = \mathrm{Var}\{X_n\} + \left(\mathbb{E}\{X_n\} - \mu\right)^2, \tag{2}$$

y utilice ese resultado para demostrar que $X_n \xrightarrow{P} \mu$.

2. Considere la región $A \subset \mathbb{R}^2$ visible en la siguiente figura.

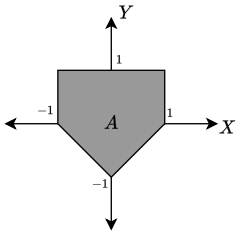

Figura 1: Diagrama de $A \subset \mathbb{R}^2$ a considerar.

Considere un vector aleatorio (X, Y), el cual distribuye uniformemente sobre la región A.

a) Obtenga la distribución conjunta $f_{X,Y}(x, y)$.

b) Obtenga las distribuciones marginales $f_X(x)$, $f_Y(y)$, y verifique si X e Y son independientes. Calcule $\mathbb{E}\{X\}$ y $\mathbb{E}\{Y\}$.

c) Obtenga la matriz de covarianza del vector (X, Y).

d) Calcule $\mathbb{E}\{Y|X = x\}$.

P2 - Pregunta computacional

En el contexto de la electromovilidad, a diferencia de lo que ocurre con los vehículos a combustión interna, saber si un vehículo eléctrico será capaz o no de terminar correctamente una cierta ruta es un problema no trivial, debido a la diversa fenomenología que hay en torno a ellos: fenómenos como el frenado regenerativo, cambios en el rendimiento de las baterías en función al perfil de uso, el impacto de la temperatura en el rendimiento de un vehículo eléctrico, entre muchos más, hacen que generar estimados precisos sea un problema altamente complejo.

En este problema, usted tendrá un primer acercamiento a la resolución de este problema frente a datos sintéticos y condiciones altamente ideales. El objetivo de esta pregunta es que, a partir de una cierta ruta con consumos de energía conocidos, usted sea capaz de indicarle al usuario si su vehículo será capaz, o no, de terminar efectivamente la ruta, o si se quedará sin carga antes del fin de esta.

Para esto, se colocan a su disposición dos conjuntos de datos simulados del consumo energético, realizados con un medidor ruidoso, de un vehículo eléctrico ante ciertas rutas conocidas, junto al porcentaje de la carga total que ha sido consumido al final de esta ruta. Los conjuntos los puede encontrar en los archivos `train.txt` y `val.txt`. En ambos archivos, cada fila corresponde a una ruta distinta con el formato

`Estado de carga final`, `Energía viaje 1`, `Energía viaje 2`, ...

donde la energía consumida por cada viaje de la ruta se encuentra en Wh. En la siguiente figura, puede ver un ejemplo de una ruta que el vehículo puede seguir, donde Y_i corresponde a la energía consumida en cada viaje, y X corresponde al porcentaje de carga consumida al final de la ruta.

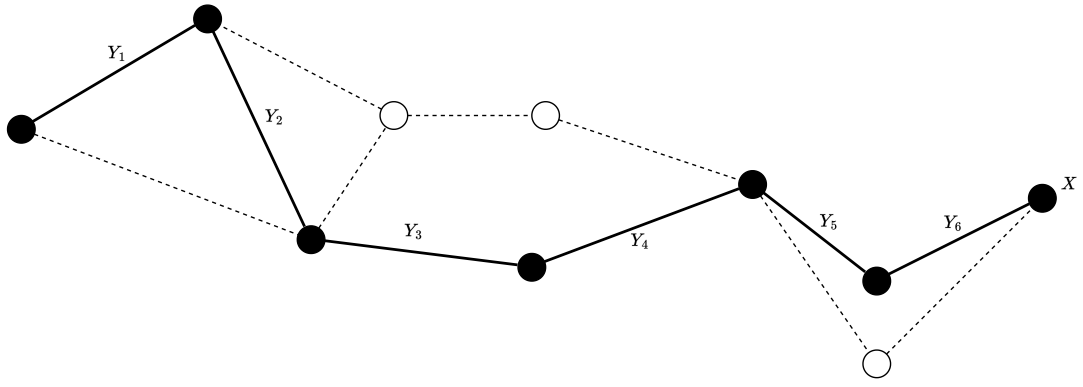

Figura 2: Ejemplo de ruta seguida por un vehículo eléctrico.

Para el desarrollo de este sistema, considere los siguientes pasos.

Análisis preliminar

1. Defínase X e Y de forma adecuada, considerando que X es la variable oculta que busca estimar, e Y debe ser una variable medible a partir de la cual busca estimar X.

2. Formule un estimador $\hat{x}(Y)$ el cual le permita estimar X a partir de Y, dejándolo únicamente en términos de la distribución conjunta $f_{X,Y}$.

3. Formule un estimador lineal $\hat{x}_{\text{linear}}(Y)$, distinto al anterior, que le permita estimar X a partir de Y. Este estimador debe tener la estructura $\hat{x}_{\text{linear}}(Y) = AY + b$, y debe indicar, explícitamente, las expresiones asociadas a A y b.

4. Cargue los archivos `train.txt` y `val.txt`, los cuales utilizará como conjuntos de entrenamiento y validación, respectivamente. Siempre que esté trabajando con datos, en cualquier contexto, es recomendable trabajar con conjuntos de entrenamiento y validación separados: el primero es utilizado para entrenar sus modelos, y el segundo es utilizado para verificar que estos modelos pueden generalizarse, correctamente, a situaciones no registradas con anterioridad. **Nunca debe incluir sus datos de validación en su entrenamiento**.

5. Obtenga las distribuciones empíricas de X e Y obtenidos desde la data de entrenamiento y de validación (debería obtener 4 figuras). El número de "bins" a utilizar debe elegirlo usted, procurando un adecuado balance entre número de datos por bin y caracterización visual de la incertidumbre.

Entrenamiento

6. Utilizando solamente los datos de entrenamiento, obtenga la distribución conjunta de forma empírica. Grafique esta distribución, colocando X en el eje X e Y en el eje Y.

7. Comente sobre la distribución empírica de esta conjunta. ¿Es capaz de identificar algún comportamiento que relacione las variables X e Y? ¿Puede hacer alguna afirmación, a priori, sobre la incertidumbre asociada a estimar X a partir de Y?

8. Implemente el estimador $\hat{x}$ obtenido en el punto 2, utilizando la distribución conjunta empírica obtenida anteriormente..

9. Implemente el estimador $\hat{x}_{\text{linear}}$ obtenido en el punto 3, entrenado a partir de los datos de entrenamiento.

Validación

10. Considerando los dos estimadores anteriores, entrenados a partir de los datos de entrenamiento, genere un estimado del porcentaje de carga utilizado para cada ruta del conjunto de validación.

11. Genere un gráfico del error de cada estimador para cada valor de y. ¿Cómo se relaciona esto con lo que fue capaz de observar a partir de la distribución conjunta empírica?

12. Considerando RMSE (Root Mean Square Error) y MAE (Mean Absolute Error) como métricas de error, definidos como

$$\text{RMSE} = \sqrt{\frac{1}{|\,\text{Val}\,|} \sum_{(x,y)\in \text{Val}} |\hat{x}(y) - x|^2}, \tag{3}$$

$$\text{MAE} = \frac{1}{|\,\text{Val}\,|} \sum_{(x,y)\in \text{Val}} |\hat{x}(y) - x|\,, \tag{4}$$

donde Val es el conjunto de validación, y en base a lo observado anteriormente, concluya sobre el rendimiento de los estimadores. ¿Cuáles son las ventajas y desventajas de cada uno? ¿Cuál elegiría usted?