Universidad de Montevideo – Introducción a la Ciencia de Datos

Examen 2020

02 de Diciembre, 2020

1. Cross Validation:
   a. ¿Cuál es la diferencia entre el Validation Set y el Test Set?
   b. ¿Cuándo se debería hacer el pre-procesamiento de los datos, antes o después del split? Explique su respuesta.

2. En el contexto de modelos de clasificación y en base al siguiente cuadro:

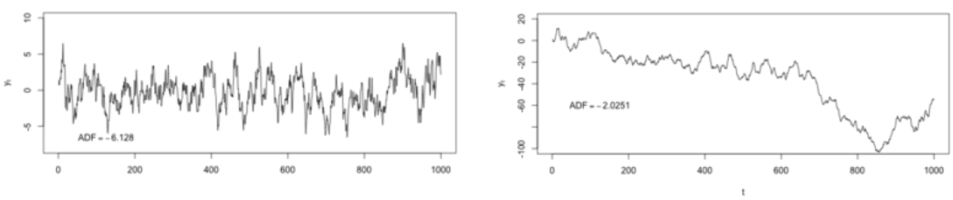
	Predicted	
Actual Class	1	0
1	f_{11}	f_{10}
0	f_{01}	f_{00}

   a. Identifique el error Tipo I y Tipo II.
   b. Si se estuviera tratando de predecir si una persona es inocente o no, con la condena de pena de muerte como consecuencia, ¿qué error considera más importante de minimizar?
   c. ¿Cómo se calcula el Accuracy, Precision, Specificity y Sensitivity? Explique qué significa cada una de ellas.

3. Enumere y explique los distintos componentes de una serie de tiempo. Puede apoyarse en la ecuación vista en clase: xt = ft + st + ct + et.

4. Estacionariedad:

   a. ¿Qué significa que una serie de tiempo sea estacionaria?
   b. Analizando gráficamente, ¿son los dos series de tiempo a continuación estacionarias?



5. ¿Qué es clustering? Explique un algoritmo visto en clase.

6. Explique 3 aplicaciones prácticas de procesamiento de lenguaje natural.

7. Explique 3 tipos de pre-procesamientos que se realizan en problemas de procesamiento de lenguaje natural.

8. ¿Qué es la maldición de la dimensionalidad (curse of dimensionality) y cómo puede resolverse?

9. Explique la diferencia entre One Hot Encoding y Ordinal Encoding. ¿Cuándo suele usarse uno y otro?

10. Explique los objetivos perseguidos al utilizar herramientas de Storytelling para comunicar resultados.