**Respuestas examen 02-2021**

**2. Por qué es necesaria la ingeniería de atributos?**

- Si la clase es una función muy compleja de los atributos seguramente nuestro modelo no será capaz de aprender. Frecuentemente, los datos de entrada sin procesar no son interpretables por nuestros modelos. En consecuencia, es imprescindible generar atributos que sí lo sean.

- Crear atributos informativos y no redundantes (independientes o no correlacionados), facilitando el subsecuente aprendizaje y generalización.

- Crear atributos que conllevan a una mejor interpretación humana.

- Simplificar nuestra representación de la realidad.

- Reducir la dimensión de nuestros datos.

- Taclear el problema de la “Maldición de la dimensionalidad” sin una pérdida de información significativa.

- Hacer la representación de los datos más adaptables o más fáciles de procesar por un algoritmo.

- Resaltar las características que diferencian a nuestras instancias o registros.

- Los datos no son adecuados para su análisis o para ser ingresados a un modelo predictivo, por lo que nuevos atributos deben ser creados.

**3. Luego de generar predicciones con un modelo predictivo usted nota una gran diferencia de performance entre su set de entrenamiento y su set de validación. ¿Qué fenómeno es este? ¿Cómo puede evitarlo?**

- El fenómeno es overfitting.

Metódos de evitar overfitting:

- Cross-validation (principal)

- Entrenar con más datos de ser posible.

- Remover features irrelevantes.

- Añadir regularización.

- Ensembling de modelos.

**5. Una empresa "A" fue contratada por una empresa inmobiliaria "B" para realizar un proyecto de Data Science. La misma al finalizar el proyecto, entre otras cosas, entrega un modelo predictivo que tiene una precisión del 90%. Mucho tiempo ha pasado desde esto y la empresa "B" ha crecido significativamente aumentando su número de clientes. La misma ha experimentado una disminución progresiva de performance del modelo que le fue entregado. ¿Por qué usted cree que esto ha ocurrido? ¿Qué cree usted que debe hacerse?**

- Mucho tiempo ha pasado, por lo tanto los datos se encuentran desactualizados. El algoritmo al haber entrenado con datos actualmente desactualizados predice con un error mayor que cuando fue entrenado.

- La empresa al haber crecido aumenta su número de clientes y consecuentemente su número de registros. Esto usualmente se traduce en mayor varianza, la cual puede no haber sido capturada en el entrenamiento con menor número de casos.

- Se debe re-entrenar el modelo con los nuevos datos y hacer re-entrenamientos de manera peródica.

**9. Explique los problemas asociados al clustering de datos poniendo ejemplos que ilustren los problemas.**

- En un problema de unsupervised learning no tenemos un atributo objetivo que predecir, quedando a merced del algoritmo encontrar la estructura de los datos.

- El número correcto de clusters es frecuentemente ambigüo dependiendo de la interpretabilidad de la forma de la distribución y el nivel de resolución que se quiere alcanzar.

- Lidiar con volúmenes grandes de datos y un gran número de dimensiones puede ser problemático por la complejidad del tiempo de cómputo.

- La efectividad del método depende de la definición de distancia y la métrica a evaluar.

- El resultado de un análisis de clustering puede ser interpretado de distintas maneras pudiendo llegar a ser tanto ambigüo como arbitrario.

**10. Explique cómo determinaría la similitud entre los libros de una biblioteca incluyendo el pre-procesamiento necesario.**

- Hacer el pre-procesamiento visto en clase:

1.Tokenization.

2.Case folding.

3.Removal of short and stop words.

4.Stemming.

5.Lemmatization.

- Generar una BOW (bag of words)

- Normalizar la BOW o usar TF-IDF

- Computar la similitud usando alguna metrica de similitud (coseno, euclídea, pearson).