Entrevista

Fecha: 17/06/2024

Problemática: Optimización Bayesiana en modelos de clasificación de Minería de Datos para determinar patrones en los asesinatos de la Zona 8 del Ecuador

Presentación (5 minutos):

La presente entrevista tiene como objetivo recolectar información y profundizar sobre las técnicas de optimización, en este caso la optimización Bayesiana, que se usará para mejorar el rendimiento de los modelos, al momento de evaluar con la métrica de precisión los modelos de Minería de Datos: Árbol de Decisión y SVM, para determinar los patrones en los asesinatos de la Zona 8 del Ecuador. Las respuestas de la Ingeniera en Computación, ayudarán a solventar cualquier duda, ya que estarán profundizando sobre el tema en cuestión, mejorando el entendimiento y análisis por parte del estudiante. Agradeciendo de antemano su colaboración y predisposición para brindar sus conocimientos en este campo de la Minería de Datos.

Actores:

- **Datos Personales del Entrevistador:**

Nombres: Cecilia Fernanda Trueba Reyes

Cargo: Estudiante de la Carrera de Ingeniería en Ciencias de la Computación

Correo electrónico: cecilia.trueba@unl.edu.ec

- **Datos Personales del Entrevistado:**

Nombres: Ing. Genoveva Suing Albito

Cargo: Docente de la materia de Machine Learning

Objetivos:

- Obtener conocimientos sobre la optimización Bayesiana que ayudaran a mejorar el rendimiento de los modelos de minería de datos.
- Obtener información valiosa que profundicen sobre la importancia de las métricas de rendimiento al momento de evaluar los modelos de Minería de Datos.
- Conocer las herramientas, métodos y técnicas de preferencia a usar, para realizar el análisis, entrenamiento, implementación y evaluación de los modelos de minería de datos, aplicadas a los datos de los homicidios de la zona 8 del Ecuador para determinar patrones.

Preguntas y Respuestas de la Entrevista:

1. **¿Por qué es importante la métrica de rendimiento específicamente la precisión para evaluar los modelos de minería de datos?**

   Al poder considerar algoritmos relacionados dentro de la rama a nivel general Machine Learning, existen algunas métricas, una de ellas parte importante y decisiva para poder saber en cuanto al rendimiento, es la métrica de precisión, pero siempre es aconsejable no únicamente considerar una métrica, sino hacer una comparativa, poder ver algunas métricas adicionales, para que en base a ello poder considerar la evaluación del rendimiento del modelo, que usted quiere involucrar en este caso relacionados con la minería de datos. Sí, considero que la precisión es muy importante, pero adicional podríamos complementarlas con evaluación de otras métricas también.

2. **¿Por qué es importante aplicar la optimización Bayesiana en el ajuste de parámetros de los modelos de minería de datos: Árbol de Decisión y SVM?**

   La optimización en todos los modelos minería de datos es un factor muy importante, ya que conlleva a la búsqueda de parámetros que mejor se ajusten a la exploración y explotación de datos, con la finalidad de llegar a obtener un modelo eficiente y efectivo, que logre el mejor rendimiento posible minimizando el tiempo y recursos computacionales necesarios.

3. **¿Qué ventajas presenta el uso de optimización Bayesiana en comparación con otras técnicas de optimización?"**

   Generalmente cuando se involucra la optimización Bayesiana, presenta ventajas que permiten tener ese criterio de optimización, ya que nos permite tener técnicas de optimización más ajustables al proyecto que usted está involucrado, es un algoritmo que por lo general sí está vinculado para hacer esta minería de datos.

4. **¿Qué beneficios trae la técnica de optimización en los modelos clasificadores de minería de datos?**

   Bueno, una de las consideraciones que como ya lo hemos visto, igualmente relacionados con modelos de Machine Learning, involucra el poder tener argumentos para poder hacer esa clasificación, hoy en día, los porcentajes a los cuales se puede llegar con algoritmos de Machine Learning, pueden estar superando un 90%, en el mismo escenario, se puede considerar la optimización Bayesiana, no únicamente quedarse con un solo algoritmo de configuración normal, siempre tratando de evidenciar o trabajar con esos hiperparámetros que hacen referencia para poder mejorar cada vez el rendimiento, por ende poder mejorar lo que involucra tener un algoritmo relacionado con optimización, existen algunas técnicas que hemos visto, relacionadas con descenso de gradiente, que nos permite de igual manera poder disminuir el error, siempre tratar de buscar esa representación del error mínimo, lo más aproximada posible a una solución analítica pero con porcentajes de error que sean lo menos posibles.

5. **¿Recomienda usar Google Colab en la nube o Python de manera local para realizar el análisis, procesamiento e implementación de los modelos?**

   El uso de herramientas para la implementación de modelos de minería de datos para su trabajo de titulación, se puede establecerse en base a un análisis comparativo de las ventajas y desventajas de varias herramientas existentes actualmente. Por otro lado, también se debe considerar el conocimiento que tenga en el uso de estas aplicaciones.
   Si es para estudio académico, tanto Google Colab en la nube o Python si son adecuadas, ya que proporcionan las librerías necesarias para la implementación de modelos de minería de datos.

6. **¿Qué aplicación o herramienta según su experiencia se podría usar para la limpieza de los datos?**

   Actualmente existen gran cantidad de herramientas para la limpieza de datos como OpenRefine, Trifacta, DataCleaner, pandas, R y Matlab, esta selección debe ser considera de acuerdo a un análisis comparativo entre ellas, tomando en cuenta la información disponible.

7. **¿Es más factible entrenar y validar los modelos programándolos en Python o simplemente usando la herramienta Weka?**

   La implementación mediante lenguaje de programación le facilita el manejo y control de arquitectura del modelo, para poder hacer los ajusten de hiperparámetros con la finalidad de obtener un modelo optimizado, mientras que las herramientas como Weka nos limitan en algunos cambios en la arquitectura.

8. **¿Por qué el porcentaje de las métricas de rendimiento de los modelos pueden ser muy bajos en algunos casos?**

   Generalmente podemos tener esos porcentajes bajos por algunos escenarios, los más comunes podrían ser, por falta de datos, es muy poca la información que se puede tener para poder generar una proyección, otra de las características pueden ser las configuraciones internas que podemos estar dando ya que no son adecuadas, si yo estoy modelando una regresión polinomial, a lo mejor yo visualizo que el escenario lo esté relacionando con alguna regresión cuadrática y realmente mi escenario está vinculado con una regresión polinomial de un mayor grado, por configuraciones internas, a lo mejor una particularidad gaussiana, entonces todas estas configuraciones permiten que nosotros mejoremos esos porcentajes, y uno de los factores es la configuración y lo que involucra los datos, las dos condiciones, que se pueden dar.

9. **¿Por qué es importante evaluar y comparar los modelos de Minería de Datos antes y después de aplicar la optimización Bayesiana?**

   Bueno a nivel general , no únicamente dentro de minería de datos, sino a nivel de todo algoritmo relacionado con el aprendizaje de máquina, podríamos considerar el poder crear un modelo, poderlo generar y tenerlo establecido, está correcto, pero siempre tiene que pasar por la validación, si está relacionado con el ámbito de la salud, tendría

que ir con el experto, poder probar, poder tener esos resultados, y ver si efectivamente tengo alguien quien lo valide, porque yo puedo como técnico en el área de Ingeniería puedo configurar a mi criterio pero generalmente trabajamos con requerimientos de usuario, así sea un modelo pequeño, para yo poderlo interpretar, para poder sacar y ver información, qué hiperparámetros puedo cambiar, cual es la predicción adecuada, entonces yo siempre tengo que hacer esa etapa de validación, el decir sí, esta correcto o no, de esta manera no únicamente quedarme con mi criterio técnico, porque recordemos nosotros creamos y diseñamos para clientes, finalmente validarlo con el experto o con el más cercano, para tener una validez de lo que he podido configurar.

10. **¿Por qué es importante determinar patrones en los asesinatos de la Zona 8 del Ecuador con modelos clasificadores de minería de datos?**

    Realmente poder identificar este tipo de información, es algo que les permite a quienes estén vinculados, como en este caso la policía, el poder tener esa comparativa, poder visualizar cuales serían los factores, ya que esto es parte esencial de un análisis, el poder ver cuáles son esas características, esa información relevante, que les permita a ellos aplicar algunas correctivas y tener algunas alertas relacionadas con este tipo de investigaciones que usted está desarrollando y la optimización Bayesiana es uno de los algoritmos que se los ha podido vincular, ya que tiene sus características importantes para poder hacer este tipo de análisis.

Firmado electrónicamente por:
GENOVEVA JACKELINNE SUING ALBITO

…….. ……

Ing. Genoveva Suing Albito — **Entrevistado**

Cecilia Trueba Reyes — **Entrevistador**