CC58 Tópicos en Ciencia de la Computación
Carrera de Ciencias de la Computación
Tiempo Límite: 110 Minutos

Examen Final
Julio, 2021
MSc. Pedro Shiguihara

- **Objetivo**. Aplicar las técnicas de estimación de modelos de grafos probabilísticos para PLN.
- **Lenguaje de programación**. El lenguaje a utilizar es Python en Notebook. Adjuntar: videos y descripciones gráficas o textuales para documentar y reforzar tu respuesta.
- **Entrega de la evidencia**. Cada una de sus respuestas debe estar respaldada por el código fuente respectivo. Utilizar el lenguaje *markdown* dentro del Notebook para documentar correctamente la respuesta a cada pregunta. Trata de que **tu programa sea genérico y útil para cualquier dataset**. El alumno debe enviar como evidencia el archivo en extensión **ipynb** dentro de un archvo comprimido (ZIP) y además el link compartido con permisos de edición comentado en la primera línea del archivo **ipynb** a fin de poder visualizarlo también *online*.
- **Video**. Debe haber un video demo para explicar cómo funciona su notebook desde la carga de datos hasta la predicción de datos final. El video debe durar entre 10-20 minutos. Se puede adjuntar el link. Usar dataset completo tanto para entrenamiento como validación.
- **Informe**. Todo este proceso describiendo el paso a paso, debe estar documentado en un informe usando la plantilla IEEE con un máximo de 3 páginas:
  https://www.ieee.org/conferences/publishing/templates.html.

Problema	Puntos	Resultado
1	5	
2	5	
3	5	
4	5	
Total:	20	

1. Descripción

Los alumnos deben leer todas las preguntas y, en caso de duda, deben enviar un email durante los primeros 20 minutos al coordinador asignado a su examen, quien absolverá la duda vía email.

2. PLN y Modelos de Grafos Probabilísticos

Usar dataset completo tanto para entrenamiento como validación.

A partir de un dataset en CSV en la sección Examen Final de Blackboard, tu programa debe estimar:

1. (5 puntos) [Aprendizaje Supervisado y PLN:] se debe entrenar redes bayesianas generadas usando (1) hill-climbing/greedy, (2) búsqueda K2 con la métrica entropía y métrica K2 con vectorización bag-of-words. Aplicar las métricas de desempeño vistas en clase: F1, precisión, recall, accuracy, gráficos ROC y plasmar conclusiones al respecto.

2. (5 puntos) [Ensemble por Moda de Clases Predecidas y PLN:] se debe vectorizar el texto con bag-of-words y crear un Ensemble basado en votación usando internamente 4 redes bayesianas:

    1. Una red bayesiana generada con búsqueda greedy y métrica de calidad entropía
    2. Una red bayesiana generada con búsqueda K2 y métrica de calidad entropía
    3. Una red bayesiana generada con búsqueda greedy y métrica de calidad K2
    4. Una red bayesiana generada con búsqueda K2 y métrica de calidad K2

    Aplicar las métricas de desempeño vistas en clase: F1, precisión, recall, accuracy, gráficos ROC y plasmar conclusiones al respecto.

3. (5 puntos) [Ensemble por Máximo Promedio de Probabilidades y PLN:] se debe vectorizar el texto con bag-of-words y crear un Ensemble basado en votación usando internamente 4 redes bayesianas:

    1. Una red bayesiana generada con búsqueda greedy y métrica de calidad entropía
    2. Una red bayesiana generada con búsqueda K2 y métrica de calidad entropía
    3. Una red bayesiana generada con búsqueda greedy y métrica de calidad K2
    4. Una red bayesiana generada con búsqueda K2 y métrica de calidad K2

    Aplicar las métricas de desempeño vistas en clase: F1, precisión, recall, accuracy, gráficos ROC y plasmar conclusiones al respecto.

4. (5 puntos) [Informe PLN:] todos los aspectos solicitados deben reflejarse en el Notebook y en un informe en MS Word. Es importante mantener una documentación organizada y con un nivel de detalle apropiado.

3. Entregables

Informe IEEE (relacionado al trabajo principal): informe grupal que describe el proceso y basado en el template de la IEEE con las secciones: Experimentos, Resultados y Conclusiones (entre 1-3 páginas). **De no entregar completo, se descuenta hasta 5 puntos.**

Notebook: código correctamente organizado y documentado del grupo adjuntado con formato **ipynb**.

Video: explicación del demo que debe durar entre 10-20 minutos. Los estudiantes deben mostrarse en el video y compartir pantalla en la demo. El video es el único recurso que puede ser adjuntado como link.

Los archivos deben ser adjuntados en un ZIP. En caso no estén adjuntos, el estudiante tendrá nota CERO.

Lima, julio del 2021.