Práctica Calificada 2 CC0C2

Fecha de entrega: 12 de octubre

Puntaje máximo: 20 puntos

Entrega del proyecto: 8 puntos

Exposición del proyecto: 12 puntos

Debes presentar un repositorio donde se encuentre todos tus resultados.

Instrucciones generales:

1. Cada grupo debe estar compuesto por1 o 2 estudiantes.

2. Los proyectos serán asignados por orden de elección, asegurando que cada grupo trabaje en un proyecto diferente.

3. La fecha límite para la entrega del proyecto es el 12 de octubre.

4. Las exposiciones tendrán lugar en la misma fecha de entrega. Se asignarán 15 minutos a cada grupo para exponer su proyecto, seguidos de 5 minutos para preguntas.

5. Referencia: Naive Bayes, Text Classification, and Sentiment y Logistic Regression.

Proyectos disponibles:

Proyecto 1: Integración de modelos Naive Bayes y regresión logística multinomial para clasificación multiclase con evaluación

Descripción: Desarrolla un sistema de clasificación de texto que combine modelos generativos (Naive Bayes) y discriminativos (Regresión Logística Multinomial) para tareas de clasificación multiclase, como la categorización de noticias. Implementa técnicas de descenso de gradiente estocástico con mini-lotes y regularización para optimizar los modelos. Evalúa el rendimiento utilizando métricas como precisión, recall, medida F y realiza pruebas de significancia estadística, incluyendo la prueba bootstrap pareada.

Resultados esperados:

- Implementación funcional de Naive Bayes y Regresión Logística Multinomial.
- Sistema de clasificación que integra ambos modelos para mejorar la precisión.
- Optimización mediante descenso de gradiente estocástico con mini-lotes y regularización.
- Evaluación exhaustiva utilizando precisión, recall, medida F, y pruebas de significancia.
- Análisis comparativo entre modelos generativos y discriminativos.
- Documentación detallada de la implementación y los resultados.

Entradas:

- Conjunto de datos etiquetado para clasificación multiclase (e.g., Reuters News Dataset).
- Parámetros de optimización y regularización.
- Configuraciones para pruebas estadísticas.

Salidas:

- Modelos entrenados de Naive Bayes y Regresión Logística Multinomial.
- Reporte de métricas de evaluación (precisión, recall, F1) para cada modelo.
- Resultados de pruebas de significancia estadística.
- Visualizaciones de la convergencia del descenso de gradiente.
- Análisis interpretativo de los modelos.

Proyecto 2: Desarrollo de un modelo de lenguaje basado en Naive Bayes con evaluación de perplejidad y entropía

Descripción: Construye un modelo de lenguaje utilizando Naive Bayes y explora su capacidad como modelo generativo. Implementa métricas de evaluación avanzadas, como la perplejidad y la entropía, para medir la calidad del modelo. Analiza cómo la perplejidad se relaciona con la entropía y utiliza esta relación para optimizar el modelo. Realiza evaluaciones en conjuntos de entrenamiento y prueba para analizar la generalización y evitar el sobreajuste.

Resultados esperados:

- Modelo de lenguaje Naive Bayes funcional.
- Cálculo y análisis de perplejidad y entropía del modelo.
- Evaluación detallada de la capacidad de generalización del modelo.
- Estrategias implementadas para evitar el sobreajuste.
- Reporte sobre la relación entre perplejidad y entropía.
- Documentación completa de la metodología y los resultados.

Entradas:

- Corpus de texto grande (e.g., Wikipedia).
- Configuraciones de suavización y regularización.

Salidas:

- Perplejidad y entropía calculadas para el modelo.
- Reporte de métricas de evaluación en conjuntos de entrenamiento y prueba.
- Análisis de resultados y estrategias de optimización.
- Visualizaciones que muestren la relación entre perplejidad y entropía.

Proyecto 3: Implementación de clasificadores generativos y discriminativos para análisis de sentimientos con evaluación

Descripción: Desarrolla e implementa clasificadores generativos (Naive Bayes) y discriminativos (Regresión Logística Multinomial) para la tarea de análisis de sentimientos. Optimiza ambos modelos utilizando descenso de gradiente estocástico con mini-lotes y técnicas de regularización. Realiza una evaluación exhaustiva utilizando precisión, recall, medida F, y aplica pruebas de significancia estadística, incluyendo la prueba bootstrap pareada, para comparar el rendimiento de ambos enfoques.

Resultados esperados:

- Implementación funcional de ambos clasificadores.
- Modelos optimizados para la tarea de análisis de sentimientos.
- Evaluaciones detalladas y comparativas utilizando múltiples métricas.
- Resultados de pruebas de significancia que demuestren diferencias de rendimiento
- Análisis interpretativo sobre las fortalezas y debilidades de cada clasificador.
- Documentación completa y estructurada.

Entradas:

- Conjunto de datos etiquetado para análisis de sentimientos (e.g., IMDB Reviews).
- Parámetros de optimización y regularización.

Salidas:

- Modelos entrenados de Naive Bayes y Regresión Logística Multinomial.
- Reporte de métricas de evaluación para ambos modelos.
- Resultados de pruebas de significancia estadística.
- Visualizaciones comparativas del rendimiento de los clasificadores.
- Análisis interpretativo de los resultados.

Proyecto 4: Análisis semántico de vectores y clasificación de Ddocumentos usando TF-IDF y Embeddings

Descripción: Crea un sistema de análisis semántico que utilice representaciones vectoriales avanzadas, incluyendo TF-IDF, Word2Vec (CBOW y Skip-Gram), GloVe y técnicas de factorización de matrices como PPMI y Shifted PPMI. Implementa clasificadores como Naive Bayes y Regresión Logística Multinomial para la clasificación de documentos. Evalúa el impacto de diferentes representaciones vectoriales en la precisión, recall y medida F, y realiza pruebas de significancia estadística para comparar los resultados.

Resultados esperados:

- Implementación funcional de múltiples técnicas de representación vectorial.
- Clasificadores entrenados utilizando diferentes embeddings.
- Evaluación comparativa de la efectividad de cada representación vectorial en la clasificación de documentos.
- Análisis de la influencia de las dimensiones y métodos de embedding en las métricas de rendimiento.
- Resultados de pruebas de significancia que validen las diferencias observadas.
- Documentación detallada y estructurada.

Entradas:

- Corpus de texto grande (e.g., Wikipedia, Reuters).
- Parámetros para la construcción de vectores (dimensiones, métodos de ponderación).

Salidas:

- Vectores de palabras y documentos generados por diferentes técnicas.
- Métricas de clasificación (precisión, recall, F1) para cada método de representación.
- Resultados de pruebas de significancia estadística.
- Visualizaciones comparativas de rendimiento.
- Reporte de análisis sobre la efectividad de cada técnica.

Proyecto 5: Desarrollo y evaluación de un sistema de clasificación multiclase usando embeddings y regresión logística multinomial optimizada

Descripción: Implementa un sistema de clasificación multiclase que utilice embeddings avanzados (Word2Vec, GloVe) como representaciones vectoriales de palabras y documentos. Utiliza Regresión Logística Multinomial optimizada con descenso de gradiente estocástico, mini-lotes y regularización para entrenar el modelo. Evalúa el rendimiento utilizando métricas como precisión, recall, medida F, y aplica pruebas de significancia estadística para comparar diferentes configuraciones de embeddings y técnicas de optimización.

Resultados esperados:

- Sistema de clasificación multiclase funcional utilizando embeddings avanzados.
- Implementación optimizada de Regresión Logística Multinomial con técnicas de regularización.
- Evaluación comparativa del rendimiento con diferentes tipos de embeddings y configuraciones de optimización.
- Resultados de pruebas de significancia que demuestren diferencias significativas entre configuraciones.
- Análisis interpretativo sobre el impacto de los embeddings y las técnicas de optimización.
- Documentación completa de la implementación y los resultados.

Entradas:

- Conjunto de datos etiquetado para clasificación multiclase (e.g., 20 Newsgroups).
- Parámetros para la construcción de embeddings y optimización del modelo.

Salidas:

- Modelos entrenados con diferentes embeddings y configuraciones de optimización.
- Reporte de métricas de evaluación para cada configuración.
- Resultados de pruebas de significancia estadística.
- Visualizaciones comparativas del rendimiento de los modelos.
- Análisis interpretativo de los resultados obtenidos.

Proyecto 6: Implementación de un sistema de similaridad semántica basado en coseno y embeddings con evaluación de perplejidad y entropía

Descripción: Desarrolla un sistema que mida la similaridad semántica entre palabras y documentos utilizando el cálculo del coseno aplicado a diferentes representaciones vectoriales (TF-IDF, Word2Vec, GloVe). Implementa un modelo de lenguaje basado en n-gramas para calcular perplejidad y entropía, y analiza cómo estas métricas afectan la calidad de las representaciones vectoriales. Realiza una evaluación exhaustiva utilizando pruebas de significancia estadística para validar las relaciones encontradas.

Resultados esperados:

- Sistema funcional de medición de similaridad semántica utilizando diferentes técnicas vectoriales.
- Implementación de modelos de lenguaje n-grama para calcular perplejidad y entropía.

- Análisis detallado de cómo perplejidad y entropía influyen en la calidad de las representaciones vectoriales.
- Resultados de pruebas de significancia estadística que validen las relaciones observadas.
- Visualizaciones que muestren las similitudes semánticas y las métricas de evaluación.
- Documentación completa de la metodología y los resultados.

Entradas:

- Corpus de texto grande (e.g., Wikipedia).
- Parámetros para la construcción de vectores y modelos de lenguaje.

Salidas:

- Vectores de palabras y documentos generados por diferentes técnicas.
- Métricas de similaridad semántica calculadas.
- Perplejidad y entropía del modelo de lenguaje.
- Reporte de análisis y visualizaciones de los resultados obtenidos.
- Resultados de pruebas de significancia estadística.

Proyecto 7: Optimización de regresión logística multinomial para clasificación de texto con embeddings y regularización

Descripción: Implementa una Regresión Logística Multinomial para clasificación de texto, optimizada mediante descenso de gradiente estocástico con mini-lotes y técnicas avanzadas de regularización (L1, L2). Utiliza diferentes representaciones vectoriales (TF-IDF, Word2Vec, GloVe) como características de entrada. Evalúa el impacto de estas representaciones y las técnicas de regularización en la precisión, recall, y medida F, realizando también pruebas de significancia estadística para validar los resultados.

Resultados esperados:

- Implementación optimizada de Regresión Logística Multinomial.
- Clasificador entrenado utilizando diferentes representaciones vectoriales.
- Evaluación comparativa del rendimiento con distintas técnicas de regularización y embeddings.
- Resultados de pruebas de significancia que demuestren diferencias significativas entre configuraciones.
- Análisis interpretativo sobre la influencia de las representaciones y regularización en el rendimiento del modelo.
- Documentación completa de la implementación y los resultados obtenidos.

Entradas:

- Conjunto de datos etiquetado para clasificación de texto (e.g., 20 Newsgroups)
- Parámetros para la construcción de vectores y optimización del modelo.

Salidas:

- Modelos entrenados con diferentes embeddings y configuraciones de regularización.
- Reporte de métricas de evaluación para cada configuración.
- Resultados de pruebas de significancia estadística.
- Visualizaciones comparativas del rendimiento de los modelos.
- Análisis interpretativo de los resultados obtenidos.

Proyecto 8: Desarrollo de un sistema de desambiguación de sentidos basado en similaridad de vectores y clasificación multiclase

Descripción: Crea un sistema de desambiguación de sentidos de palabras que combine técnicas de similitud semántica basadas en vectores (Word2Vec, GloVe, TF-IDF) con clasificadores multiclase como Naive Bayes y Regresión Logística Multinomial. Implementa métodos avanzados de evaluación, incluyendo precisión, recall, medida F, y pruebas de significancia estadística, para medir la efectividad del sistema en diferentes contextos y corpus.

Resultados esperados:

- Sistema funcional de desambiguación de sentidos utilizando diferentes técnicas de vectorización.
- Implementación de clasificadores multiclase optimizados para la tarea.
- Evaluación comparativa de la efectividad de diferentes representaciones vectoriales y clasificadores.
- Resultados de pruebas de significancia estadística que validen las diferencias observadas.
- Análisis interpretativo sobre las fortalezas y debilidades del sistema.
- Documentación completa y estructurada de la metodología y los resultados.

Entradas:

- Conjunto de datos etiquetado para desambiguación de sentidos (e.g., WordNet Sense Inventory).
- Parámetros para la construcción de vectores y optimización de clasificadores.

Salidas:

- Vectores de palabras generados por diferentes técnicas.
- Modelos de clasificación entrenados para desambiguación de sentidos.

- Reporte de métricas de evaluación para cada configuración.
- Resultados de pruebas de significancia estadística.
- Visualizaciones comparativas del rendimiento del sistema.
- Análisis interpretativo de los resultados obtenidos.

Consideraciones adicionales para todos los proyectos

1. Documentación:

- Cada proyecto debe incluir una documentación detallada que explique la lógica detrás de las implementaciones, las decisiones técnicas tomadas, y cómo ejecutar el proyecto correctamente.

- Incluir comentarios claros en el código y una estructura modular que facilite la comprensión y el mantenimiento.

2. Pruebas y Validación:

- Implementar pruebas unitarias y de integración para asegurar la funcionalidad correcta de cada componente.

- Utilizar conjuntos de datos de validación y prueba adecuados para evaluar el rendimiento de los modelos.

3. Optimización:

- Considerar la eficiencia del código, especialmente en proyectos que manejan grandes volúmenes de datos.

- Utilizar técnicas de optimización como el procesamiento paralelo o el uso de bibliotecas optimizadas (e.g., NumPy, pandas).

4. Resultados y análisis:

- Presentar los resultados de manera clara, utilizando gráficos y tablas para facilitar la interpretación.

- Incluir un análisis crítico sobre el rendimiento de los modelos y las posibles mejoras futuras.

Rúbricas de evaluación:

1. Entrega del proyecto (8 puntos):

La entrega debe incluir el código fuente, la documentación del proyecto, y los resultados de las pruebas realizadas con el corpus asignado.

Criterio	Puntos	Descripción
Funcionalidad del código	3	El código debe implementar correctamente el proyecto propuesto y ser completamente funcional, sin errores que afecten su desempeño.
Eficiencia del algoritmo	2	El código debe demostrar eficiencia en el procesamiento, especialmente en proyectos que tratan con grandes volúmenes de datos.
Claridad y estructura del código	1.5	El código debe estar bien estructurado, con comentarios claros y buenas prácticas de programación (modularización, nombres descriptivos, etc.).
Documentación del proyecto	1.5	La documentación debe explicar la implementación, las decisiones técnicas y cómo ejecutar el proyecto correctamente.

2. Exposición del proyecto (12 puntos):

Cada grupo tendrá 15 minutos para exponer su proyecto, seguidos de 5 minutos de preguntas y respuestas.

Criterio	Puntos	Descripción

Claridad en la explicación	4	El grupo debe explicar el proyecto de manera clara, estructurada y coherente, destacando los aspectos clave de su implementación.
Entendimiento técnico	3	El grupo debe demostrar un entendimiento profundo de los conceptos aplicados en el proyecto (ej: tokenización, modelos n-grama, suavizado).
Resultados y análisis	3	El grupo debe presentar los resultados obtenidos de manera clara, con análisis crítico sobre el rendimiento del modelo o algoritmo implementado.
Manejo de preguntas	2	El grupo debe ser capaz de responder a las preguntas de los compañeros o del profesor de manera adecuada y demostrando comprensión del tema.