Introducción a la Ciencia de Datos (Optativa)
Licenciatura en Informática
2025
Trabajo práctico 3

Problema:

Para un trabajo de investigación sobre la ionosfera, se deben clasificar los retornos de radar de la ionosfera como retorno adecuado para un análisis posterior o retorno no adecuado para un análisis posterior. Esta tarea que requiere mucho tiempo generalmente ha requerido la intervención humana. Para este ejercicio, se dispone del archivo "ionosphere.data" con datos registrados por antenas de radares más su resultado de si es retorno adecuado o no, determinado manualmente por expertos.

Para este problema:

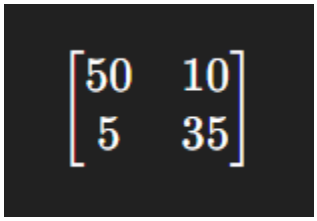
1. Se desea armar un clasificador que permita determinar, a partir de los retornos (features) disponibles de las antenas si se trata de un retorno adecuado o no (target).
2. Durante la etapa de ingeniería de datos:
    a. ¿Qué representan las features medidas por el radar? ¿Son todas numéricas? ¿Es necesario algún tipo de normalización o escalado previo al entrenamiento de la red?
    b. Determinar si los datos están balanceados. Si hay desbalance, ¿Qué técnica se podría elegir para corregirlo de acuerdo a las ventajas y desventajas que tienen cada una de las siguientes técnicas: undersampling, oversampling, SMOTE (Synthetic Minority Over-sampling Technique), data augmentation?
    c. ¿Qué impacto puede tener un dataset desbalanceado en métricas como accuracy, recall, precision, F1 Score?
    d. ¿Cuántas capas ocultas y cuántas neuronas por capa se pueden considerar adecuadas para este dataset? ¿Por qué?
    e. ¿Qué función de activación se puede elegir (sigmoid, tanh, ReLU) y cómo impacta en la propagación del gradiente?
    f. ¿Qué algoritmo de optimización usaría (SGD, Adam, RMSprop)? ¿Por qué?
3. Analizar los resultados a partir de varios modelos generados. Utilizar MLP.
4. Mostrar gráfica de la función de pérdida de cada modelo obtenido.

5. **Verdadero o falso (justificar en caso de falso):**

- El desbalance de clases puede generar un modelo con alta accuracy pero bajo recall para la clase minoritaria.
- La función de activación softmax se suele usar en la capa de salida de problemas de clasificación multiclase.
- SGD con un learning rate muy alto puede hacer que la red nunca converja.
- El F1 Score combina la precisión y el recall en una sola métrica.
- La curva de pérdida durante el entrenamiento debería ser siempre decreciente sin oscilaciones.
- Los pesos de una red neuronal se inicializan generalmente en cero para facilitar la simetría.

Introducción a la Ciencia de Datos (Optativa)
Licenciatura en Informática
2025
Trabajo práctico 3

- La técnica de early stopping ayuda a evitar sobreajuste deteniendo el entrenamiento cuando no mejora la validación.
- El F1 Score combina la precisión y el recall en una sola métrica.
- Dada la siguiente matriz de confusión (el eje vertical es la clase real y el horizontal la clase predicha):



¿El accuracy es 85%, el recall 87,5%, precision 77,8%, y el F1 Score 0,875?

- En un dataset altamente desbalanceado (990 negativos y 10 positivos), un modelo que siempre predice "negativo" alcanza un accuracy del 99% y un recall para la clase positiva de 0%.
- Batch size grande reduce ruido estocástico y puede llevar a convergencia más lenta en generalización; batches pequeños añaden ruido que a veces ayuda a generalizar.
- Si durante el entrenamiento se observa que el loss en entrenamiento = 0.1 y el loss en validación = 0.5, es posible que el modelo esté sobreajustando.
- Una red neuronal bien entrenada puede alcanzar 100% de accuracy en datos de test sin riesgo de sobreajuste.
- Una red neuronal con 1 capa oculta de 5 neuronas y función de activación sigmoide puede aproximar funciones no lineales.
- Una red con 10 capas ocultas y 2 neuronas por capa tiene mayor capacidad de representación que una red con 2 capas ocultas y 100 neuronas cada una.