Ejercicio Grupo E8FP

Jesús Gomez, Maxim Polyak

Resumen

En este trabajo se analiza un conjunto de datos de señales electromiográficas (EMG) correspondientes a cuatro gestos de mano diferentes. Se detallan los pasos de carga, preprocesado y particionado de datos; un completo Análisis Exploratorio de Datos (EDA) con estadísticas descriptivas y visualizaciones (boxplots, histogramas, distribución de clases, proyección PCA y boxplots por gesto); y la evaluación comparativa de tres métodos de clasificación (Random Forest, LDA y QDA) sobre una partición 80 % entrenamiento – 20 % test. Finalmente, se interpretan los resultados y se proponen recomendaciones para optimizar la discriminación de gestos.

1. Introducción

Las señales EMG registran la actividad eléctrica generada por la contracción de fibras musculares y se emplean en campos como prótesis controladas por músculo, interfaces cerebro-máquina y sistemas de reconocimiento de gestos. La variabilidad de estas señales hace necesario el uso de técnicas estadísticas y de aprendizaje automático para distinguir gestos de forma fiable.

2. Objetivos

1. Preprocesar los datos brutos de cuatro archivos CSV con muestras EMG.

2. Realizar un EDA que incluya:
    a. Estadísticas descriptivas (media, desviación, cuartiles, rango).
    b. Gráficos: boxplot por canal, histograma de canal de ejemplo, distribución de clases, PCA 2D, boxplots por gesto.

3. Entrenar y evaluar tres clasificadores:
    a. Random Forest
    b. Linear Discriminant Analysis (LDA)
    c. Quadratic Discriminant Analysis (QDA)

4. Comparar accuracy, precision, recall y F1-score sobre el conjunto de test (20 %).

5. Extraer conclusiones, discutir fortalezas y debilidades de cada modelo y proponer mejoras para futuras iteraciones.

3. Descripción de los datos

Origen: cuatro ficheros HandGesture0.csv … HandGesture3.csv en la carpeta /data.
Formato: cada fila contiene las amplitudes (μV) de 64 canales EMG seguidas de la etiqueta de gesto (0–3). No hay cabeceras.
Volumen: 11 678 muestras totales (≈ 2 920 por gesto). Después de concatenar, separamos X (64 columnas) y y (etiqueta).

4. Metodología

1. Carga y concatenación
    a. Lectura de todos los CSV con pandas.read_csv(header=None) y unión en un solo DataFrame.
    b. Separación de X e y, comprobando dimensiones: 11 678 × 64 para X, 11 678 etiquetas.

2. Particionado
    a. División en 80 % entrenamiento (9 342 muestras) y 20 % testeo (2 336 muestras), usando train_test_split(..., stratify=y) para preservar el balance de clases.

3. Análisis Exploratorio de Datos (EDA)
    a. Estadísticas descriptivas: X.describe().T → media, std, min, Q1, mediana, Q3, max y rango por canal.
    b. Boxplot por canal: visualización de mediana, cuartiles, mínimos, máximos y media (rombo rojo) para cada uno de los 64 canales.
    c. Histograma de un canal de ejemplo (canal 0) para detectar concentración central y picos extremos.
    d. Distribución de clases: gráfico de barras confirmando ≈ 2 919 muestras por gesto.
    e. PCA 2D: proyección sobre dos componentes principales para inspeccionar agrupamientos y superposiciones entre gestos.
    f. Boxplots por gesto: para cada una de las cuatro etiquetas, se muestran las 64 distribuciones de amplitud canal a canal.

4. Modelado y evaluación
    a. Se configuran tres clasificadores con valores por defecto (fix random_state donde aplica).
    b. Cada modelo se entrena con X_train, y_train y se evalúa con X_test, y_test.
    c. Métricas:
        i. Accuracy global
        ii. Classification report: precision, recall y F1-score por clase
        iii. Matriz de confusión para el mejor modelo (QDA).

5. Análisis Exploratorio de Datos

1. Estadísticas descriptivas
    a. Medias cercanas a 0 µV en todos los canales.
    b. Desviaciones estándar entre 5 µV y 25 µV.
    c. Rangos de hasta ±120 µV, garantizando variabilidad suficiente.

2. Boxplot por canal
    a. Medianas centradas, ligera asimetría negativa en algunos canales.
    b. Pocos outliers extremos, datos relativamente limpios.

3. Histograma canal 0
    a. Distribución leptocúrtica: foco en [–5, +5] µV, colas finas que reflejan picos musculares.

4. Distribución de gestos
    a. Balance casi perfecto: entre 2 900 y 2 940 muestras por clase.

5. PCA 2D
    a. Gestos 0 y 1 forman clusters relativamente separados.
    b. Gestos 2 y 3 se superponen moderadamente, indicando necesidad de modelos con flexibilidad.

6. Boxplots por gesto
    a. Tendencia central similar para todos (mediana ≈ 0).
    b. Gesto 3 tiene rango algo menor en varios canales, correlacionado con su menor recall.

6. Resultados de Modelado

Modelo	Accuracy	Precision medio	Recall medio	F1-score medio
Random Forest	0.9225	0.922	0.922	0.922
LDA	0.3390	0.350	0.340	0.340
QDA	0.9358	0.935	0.935	0.935

7. Discusión

- Random Forest: accuracy ~ 92.3 %, precision/recall uniformes (~ 0.92).
- LDA: accuracy ~ 34 %, descartado por suponer covarianza común.
- QDA: mejor accuracy (~ 93.6 %), explota diferencias de covarianza por gesto.

Matriz de confusión (QDA)

- Gestos 0, 1 y 2 con > 94 % de aciertos.
- Gesto 3: recall ~ 0.89, con la mayoría de confusiones dirigidas al gesto 2.

La PCA anticipó la dificultad de separar gestos 2 y 3, confirmada por el menor recall en ambos.

QDA se beneficia de modelar covarianza específica, mientras que LDA se ve penalizado por su restricción lineal.

Random Forest ofrece un compromiso excelente entre precisión y robustez ante ruido u outliers.

8. Conclusiones

1. Los datos EMG tienen distribuciones centradas y equilibradas, idóneas para entrenamiento.

2. Para prototipos con baja latencia, QDA es la opción principal.

3. En producción, Random Forest brinda robustez y fácil ajuste.

4. Mejorar la discriminación del gesto 3:

5. Aumentar muestras o grabaciones.

6. Extraer features en dominio tiempo-frecuencia (p.ej. RMS, energía en bandas).