Prediccion de *lpsa* con Regresion Lineal sin Regularizacion y PCA en un Caso Biomedico de Prostata

Daniel Leyton
Maestria en Inteligencia Artificial Aplicada
Universidad de la Salle
Bogotá, Colombia
yleyton96@unisalle.edu.co

***Abstract*—Este trabajo desarrolla un problema de aprendizaje supervisado de regresion para predecir la variable continua *lpsa* a partir de variables clinicas del estudio de prostata de Stamey et al. Se implementaron dos alternativas: (i) regresion lineal ordinaria sin regularizacion y (ii) PCA seguido de regresion lineal. Se utilizo la particion oficial del dataset (67 observaciones de entrenamiento y 30 de prueba), con estandarizacion basada solo en entrenamiento. Los resultados muestran que PCA+OLS con 7 componentes (95% de varianza explicada acumulada) mejora en prueba frente a OLS directo en RMSE y R^2, lo que evidencia mejor generalizacion.**

***Index Terms*—aprendizaje supervisado, regresion lineal, PCA, prediccion biomedica, PSA**

I. Introduccion

Este informe aborda la prediccion del logaritmo del antigeno prostatico especifico (*lpsa*) a partir de 8 variables clinicas: *lcavol, lweight, age, lbph, svi, lcp, gleason, pgg45*. El problema es de regresion porque la variable objetivo es cuantitativa continua. La interpretacion del caso consiste en construir un modelo que explique y prediga *lpsa* con bajo error en datos no vistos, evitando sobreajuste.

Se comparan dos enfoques:

- Modelo base: regresion lineal sin regularizacion (OLS).
- Modelo alternativo: reduccion de dimensionalidad con PCA y posterior OLS.

II. Metodologia

A. Dataset y particion

Se empleo el archivo `prostate_data.txt` con 97 observaciones y particion oficial:

- Entrenamiento: 67 casos (`train = T`).
- Prueba: 30 casos (`train = F`).

B. Preprocesamiento

Las variables predictoras se estandarizaron utilizando solo estadisticos del conjunto de entrenamiento:

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}. \tag{1}$$

Esta decision evita fuga de informacion desde prueba hacia entrenamiento.

C. Formulacion matematica

La regresion lineal sin regularizacion minimiza:

$$\min_{\boldsymbol{\beta}} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \tag{2}$$

con solucion cerrada:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3}$$

En PCA, la matriz estandarizada se descompone como:

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T, \tag{4}$$

y se seleccionan componentes que explican al menos 95% de varianza acumulada.

D. Metricas de evaluacion

Se evaluo con:

- Error cuadratico medio (MSE).
- Raiz del error cuadratico medio (RMSE).
- Error absoluto medio (MAE).
- Coeficiente de determinacion (R^2).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{5}$$

E. Implementacion computacional

El desarrollo se realizo en Python con `numpy`, `pandas`, `scikit-learn`, `matplotlib` y `seaborn`. Notebook reproducible:

https://github.com/ingeleyton/Regresion_Unisalle_ML/

III. Desarrollo de modelos

A. Modelo 1: OLS sin regularizacion

Se implemento un `Pipeline` con:

- `StandardScaler`
- `LinearRegression`

B. Modelo 2: PCA + OLS

Se implemento un `Pipeline` con:

- `StandardScaler`
- `PCA(n_components=0.95)`
- `LinearRegression`

El PCA seleccionó $k = 7$ componentes principales.

C. Resultados

TABLE I
COMPARACION CUANTITATIVA DE DESEMPENO

Modelo	Conjunto	MSE	RMSE	MAE	R^2
OLS sin regularizacion	Train	0.4392	0.6627	0.4986	0.6944
OLS sin regularizacion	Test	0.5213	0.7220	0.5234	0.5034
PCA(95%) + OLS ($k = 7$)	Train	0.4835	0.6953	0.5141	0.6636
PCA(95%) + OLS ($k = 7$)	Test	0.4483	0.6696	0.5259	0.5729

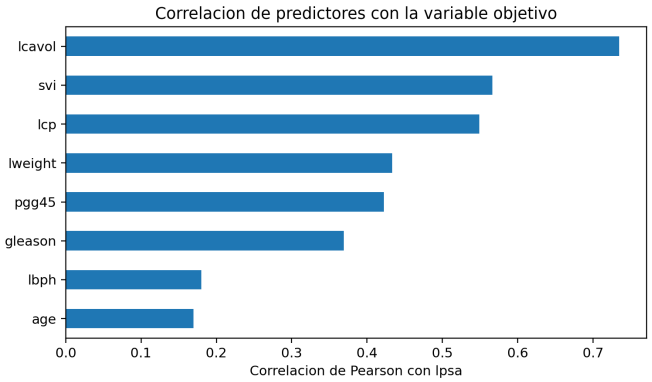

Fig. 1. Correlacion de predictores con *lpsa*.

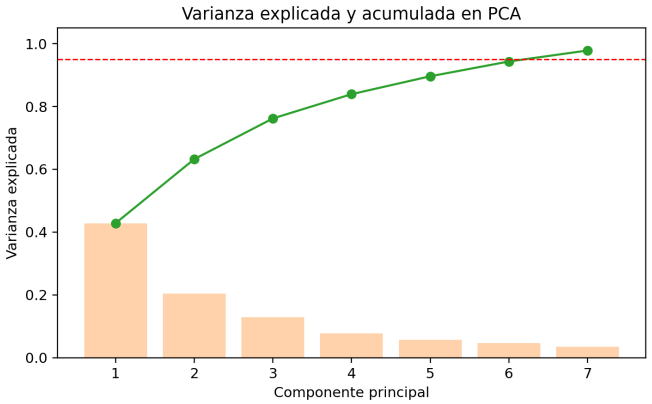

Fig. 2. Varianza explicada por PCA y varianza acumulada.

D. Analisis tecnico

El modelo OLS logra mejor ajuste en entrenamiento, pero PCA+OLS reduce el error cuadratico en prueba y mejora R^2. Esto sugiere menor sensibilidad a colinealidad entre predictores y mejor capacidad de generalizacion.

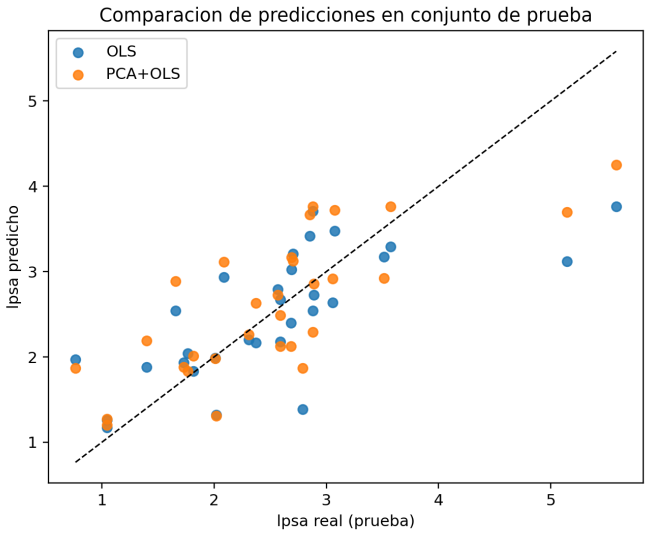

Fig. 3. Comparacion real vs predicho en conjunto de prueba.

IV. CONCLUSIONES

- Se interpreto correctamente el problema como regresion supervisada sobre una variable continua.
- Se establecio y aplico un procedimiento matematico correcto para OLS y PCA.
- Se implemento el algoritmo en Python en un notebook reproducible.
- El enfoque PCA+OLS ($k = 7$) presento mejor desempeno en prueba que OLS directo (RMSE de 0.6696 frente a 0.7220 y R^2 de 0.5729 frente a 0.5034).
- OLS mantiene ventaja de interpretabilidad sobre variables originales; PCA+OLS fue superior para prediccion en este caso.

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.

[2] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002.

[3] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[4] T. A. Stamey *et al.*, "Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate," *New England Journal of Medicine*, vol. 317, no. 15, pp. 909–916, 1987.