220731M

Withange W.I.N.

# Answers to Initial Part of the Exercise

1.What is stored in `.coef_` and `. intercept_`? Why are there so many of them?

coef_: Stores the coefficients (weights) for each feature in X_train for every class. The number of coefficients is the number of features multiplied by the number of classes , resulting in coefficients total. This reflects the "one-vs-rest" (ovr) multi-class approach, where each class has its own set of weights.

intercept_: Stores the intercept (bias) term for each class, with one intercept per class.

2.cross-validation didn't seem to offer improved results. Is this correct? Is it possible for cross-validation to not yield better results than non-cross-validation? If so, how and why?

Yes. cross-validation didn't seem to offer improved results.

Yes, it’s possible. Cross-validation (CV) doesn’t always guarantee better performance.

Cross-validation didn’t enhance the logistic regression model’s performance, as the CV with Lasso model showed slightly lower test accuracy compared to the non-CV model, despite CV’s goal of improving generalization. This can happen if the dataset is small or simple, or if the non-CV model already generalizes well, as seen in its higher train accuracy. The non-CV model’s edge might result from a favorable train-test split, while CV’s Lasso regularization could over-penalize, leading to underfitting. CV’s averaging across folds ensures robustness, but it may not surpass a "lucky" non-CV split in less complex problems, explaining the outcome.

3. Why didn't we scale the y-values (class labels) or transform them with PCA? Is this a mistake?

> The y-values (class labels) weren’t scaled or transformed with PCA because they’re categorical targets in a classification task, and such transformations are neither necessary nor appropriate for logistic regression. Scaling applies to numerical features (X), not labels, and PCA is for reducing feature dimensionality, not transforming targets. This is not a mistake—applying these techniques to y would be incorrect and disrupt the model. However, ensuring the X-values are scaled (if not already done) would be a good practice to improve model performance, especially with regularization.

4.Our data only has 2 dimensions/features now. What do these features represent?

The two features in x_train_2d and x_test_2d represent the first two principal components (PC1 and PC2) from PCA. They are linear combinations of the original scaled features, designed to capture the maximum variance in the data, with PC1 capturing the most and PC2 the second most, while being orthogonal. They don't directly correspond to the original features but are abstract dimensions for a simplified 2D representation.

5.

I. What critique can you make against the plot above? Why does this plot not prove that the different wines are hopelessly similar?

**Critique**: The plot only shows the first two principal components (PCA Dimension 1 and PCA Dimension 2), which capture the majority but not all of the variance in the data. This 2D projection may oversimplify the data, potentially masking differences that exist in higher dimensions. The overlap of bad, average, and great wines suggests difficulty in separation, but the choice of two components might not fully represent the data's structure.
**hopelessness**: The plot doesn't account for variance in other dimensions (beyond the top two PCs), where classes might be more separable. Additionally, the scaling or number of components might not optimally distinguish classes, and a different model or more PCs could reveal clearer boundaries. The overlap could also reflect PCA's focus on variance rather than class separation, not necessarily indicating inherent similarity.

II. The wine data we've used so far consist entirely of continuous predictors. Would PCA work with categorical data?

PCA is designed for continuous, numerical data and assumes variables are on a similar scale, making it unsuitable for raw categorical data. Categorical variables (e.g., "red" or "white" wine) lack the linear relationships PCA relies on, and their variance isn't meaningful in the same way. However, PCA can work with categorical data if it's encoded numerically followed by standardization.

6.What could cause this? What does this mean?

The two disjoint clusters with significant overlap in wine qualities on the PCA plot could stem from the 2D projection capturing only partial variance, leaving higher-order components unrepresented, or from latent subgroups (e.g., regions) that PCA doesn't fully resolve due to its variance focus. Inconsistent scaling, outliers, or genuine quality similarity might also contribute. This suggests the clusters differ in non-quality aspects, but wine qualities are not clearly separated in this space, indicating PCA's limitation for quality classification and the potential need for supervised methods or additional components.

7.Does this graph help you answer our previous question? Does it change your thoughts?

Yes, this PCA plot helps address the earlier question about disjoint clusters and wine quality overlap. The clear separation between red and white wines along PCA Dimension 1 suggests that color (red vs. white) is a strong differentiating factor in the data, captured well by the first two principal components. This contrasts with the previous quality-based plot, where overlap indicated poor class separation, highlighting that PCA effectively separates categorical variables like wine color but may not for quality traits.

This reinforces that PCA's effectiveness depends on the target variable. For color prediction, the plot suggests good separability, potentially improving model performance if predicting red vs. white were the goal. However, it doesn't alter the earlier conclusion that quality separation (e.g., bad, average, great wines) is challenging, as color and quality likely involve different data structures.

8.Use Logistic Regression (with and without cross-validation) on the PCA-transformed data. Do you expect this to outperform our original 75% accuracy? What are your results? Does this seem reasonable?

I didn't expect logistic regression with PCA to outperform the original accuracy, as reducing to two dimensions likely discards critical information needed for quality prediction, especially given the overlap in classes seen earlier. The results confirm this, showing significantly lower train and test accuracies compared to the original model, CV with Lasso, and even the baseline. This is reasonable because PCA's dimensionality reduction, while useful for visualization or tasks like red vs. white wine separation, oversimplifies the data for quality classification, losing variance essential for distinguishing classes

| | Train Accuracy | Test Accuracy |
|---|---|---|
| MLE | 0.602655 | 0.602308 |
| Logistic Regression | 0.718299 | 0.710769 |
| Logistic Regression w/ CV + Lasso | 0.720031 | 0.716923 |
| Logistic Regression w/ PCA | 0.616317 | 0.599231 |