Introduction to Statistical learning

DS413 - Assignment 2

Introduction

This assignment focuses on exploring and applying various machine learning algorithms to real-world and synthetic datasets. You will work with algorithms such as Principal Component Analysis (PCA) for dimensionality reduction, Linear and Polynomial Regression for predictive modeling, the Bayesian Classifier for probabilistic classification, and Fisher's Discriminant Analysis (FDA) for supervised dimensionality reduction.
The objective is to analyze datasets using appropriate techniques, visualize results effectively, and draw meaningful conclusions.

Regression

Simple Regression

Fit a regression model to the `CarSeat` dataset containing a single predictor and a quantitative response(say Sales and Population), as well as a separate polynomial regression.

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the polynomial regression(for $n = 1, 2, 3, 4, 5$). Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(b) Answer (a) using test rather than training RSS.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the polynomial regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(d) Answer (c) using test rather than training RSS.

(e) Plot n vs RSS.

Multiple Linear Regression

The following questions involve the use of the `CarSeat` dataset.

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

(b) Provide an interpretation of each coefficient in the model. Be careful - some of the variables in the model are qualitative!

(c) Write out the model in equation form, being careful to handle the qualitative variables.

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

(f) How well do the models in (a) and (e) fit the data?

(g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

(h) Is there evidence of outliers or high leverage observations in the model from (e)?

(i) Apply **Ridge regularization** to reduce overfitting

(j) Tune the regularization parameter using cross-validation.

(k) Comment on effect of regularization on coefficients.

(l) Repeat the above tasks for numerical predictor and quantitative response.

Note

On a dataset of your choice, fit a regression model for $exp(x)$ kind of feature.

Dimentionality reduction techniques

Dataset Description

The dataset of wildfire dataset from two regions of `Algeria|Bejaia` and `Sidi Bel-abbes` consists of wildfire-related meteorological factors recorded over a specific time period.

(a) Number of Instances: 122 (for each region)

(b) Time Period: June 2012 to September 2012

(c) Number of Variables (Factors): 10

(d) Missing Entries: None

(e) Classes: Binary classification (Fire or No Fire)

The dataset includes the following 10 variables:

(a) Temperature (°C): Noon temperature

(b) RH (%): Relative Humidity

(c) WS (km/h): Wind Speed

(d) Rain (mm): Precipitation

(e) FWI: Fire Weather Index

(f) FFMC: Fine Fuel Moisture Code

(g) DMC: Duff Moisture Code

(h) DC: Drought Code

(i) ISI: Initial Spread Index

(j) BUI: Build-up Index

Singular Value Decomposition (SVD)

(a) Calculate the SVD of a matrix $A = U\Sigma V^T$.

(b) Explain the significance of singular values and orthogonal matrices.

(c) Perform dimensionality reduction using SVD.

Principal Component Analysis (PCA)

The PCA procedure involves the following steps:

(a) Standardize the dataset: Since variables have different scales, standardization is necessary.

(b) Compute covariance matrix: To understand feature relationships.

(c) Find eigenvalues and eigenvectors: To determine principal components.

(d) Sort eigenvalues: Higher eigenvalues correspond to more important principal components.

(e) Project data: Transform data into new dimensions based on top principal components.

Visualization and presentation of results

Generate the following plots to analyze PCA results:

- Scree Plot: To show the variance explained by each principal component.
- Biplot: To visualize data in the reduced dimensional space.
- Cumulative variance plot: To decide the number of components to retain.

In conclusion, write the key findings of PCA:

- The percentage of variance explained by the top components.
- Observations regarding which factors contribute most to variance.
- Interpretation of PCA results in terms of wildfire prediction.

Linear Discriminant Analysis (LDA)

(a) Compute class means.

(b) Calculate within-class and between-class scatter matrices.

(c) Solve the generalized eigenvalue problem.

(d) Project data onto lower-dimensional space.

(e) Visualize results using appropriate plots.

Conclusion

All the results need to be included in a **report**, and the report should contain the results of all the models mentioned above, along with their explanations.

Bonus Question

Summarize your work methodology, and challenges faced during the assignment.

References

List all references and sources used for the assignment.