Εξόρυξη και προετοιμασία δεδομένων

1η Εργασία PCA & LDA

Πάνος Καμπάσης
21/10/2023

Ανάλυση Κύριων Συνιστωσών

Η ανάλυση κύριων συνιστωσών αποτελεί την απλούστερη *unsupervised* πολυμεταβλητή ανάλυση και στοχεύει στην ανεύρεση από ένα πλήθος π μεταβλητών ορισμένων νέων ολιγάριθμων μεταβλητών οι οποίες έχουν την ιδιότητα να είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών και παράλληλα να μη συσχετίζονται μεταξύ τους. Το μεγάλο πλεονέκτημά τους έγκειται στην ιδιαιτερότητα που διαθέτουν, λόγω της ανάλυσης, να εξηγούν πολύ μεγάλο ποσοστό της ολικής μεταβλητότητας που αναπτύσσεται μεταξύ των π μεταβλητών, το οποίο τελικά κατανέμεται σε μερικές μόνο νέες μεταβλητές. Έτσι, το μέγιστο μέρος της πληροφόρησης που θα αντλούνταν αν λαμβάνονταν υπόψη οι π μεταβλητές συγκρατείται με τη δημιουργία αυτών των νέων μεταβλητών.
Έστω ένα δείγμα $X_1, X_2, ..., X_n$ όπου το κάθε X είναι μια τ.μ. και έστω ότι ακολουθεί κανονική κατανομή και του δίνουνε μια διάσταση m.Υπολογίζουμε τον πίνακα συνδιακύμανσης:

$$C_{m,n} = \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,n} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m,1} & c_{m,2} & \cdots & c_{m,n} \end{pmatrix}$$

Η διακύμανστη της κάθε μεταβλητής εξηγεί την μεταβλητότητα τους. Σε αυτό το σημείο θα κατασκευάσουμε νέες τ.μ. $Z_1, Z_2, ..., Z_k$, με $k < n$ για τις οποίες θα ισχύει:

- $Z_k = a_{k1}X_1 + a_{k2}X_2 + ... + a_{kn}X_n$. κ = 1,2,...,χ
  Όπου a_{kn} ειδικοί συντελεστές της ν μεταβλητής στην κ συντιστώσα και ισχύει πως: $a_{11}^2 + a_{12}^2 + ... + a_{1n}^2 = 1$
- Οι Z_k, Z_l , με $k \neq l$, είναι ασυσχέτιστες.

Επίσης:
Οι διακυμάνσεις (μεταβλητότητα) που αναπτύσσονται μεταξύ των μεταβλητών Z_i, διαβαθμίζονται με τέτοιο τρόπο ώστε η πρώτη μεταβλητή Z_1 επιλέγεται να εξηγεί ένα όσο το δυνατόν μέγιστο ποσοστό της ολικής μεταβλητότητας, η Z_2 ένα δεύτερο μέγιστο ποσοστό αυτής κοκ., υπακούοντας στη σχέση:
$\lambda_1 > \lambda_2 > ... > \lambda_p$, όπου λ_i η ι ποσότητα της διακύμανσης. Οι νέες μεταβλητές Z_i καλούνται κύριες συνιστώσες και με τον τρόπο αυτόν δημιουργούνται

ολιγάριθμες Z συνιστώσες, οι οποίες, ωστόσο, εξηγούν μεγάλο ποσοστό της συνολικής διακύμανσης $\sum_{i=0}^{n} \lambda_i$. Ταυτόχρονα, πολυάριθμες δευτερεύουσες συνιστώσες εξηγούν μικρό έως ελάχιστο ποσοστό και συνεπώς το στατιστικό τους αποτέλεσμα μπορεί να αγνοηθεί χωρίς την απώλεια ουσιαστικής πληροφορίας.
Ξεκινάμε από την εύρεση της πρώτης κύριας συνιστώσας. Έστω ότι $Z = \alpha_1 X_1^\star + \alpha_2 X_2^\star$. Αν υ είναι το μοναδιαίο διάνυσμα στη διεύθυνση του α, δηλαδή: $u = \frac{\alpha}{||\alpha||}$, τότε $Z_i = ||\alpha|| u' X_i^\star$, εδώ $\alpha = (\alpha_1, \alpha_2)' \in \mathbb{R}$. Χ.β.τ.γ. υποθέτους ότι $||\alpha|| = 1$ και έχω ισοδύναμα $Z = u_1 X_1^\star + u_2 X_2^\star, u_1^2 + u_2^2 = 1$.
Έτσι με την πρώτη κύρια συνιστώσα, έχοντας μειώσει τη διάσταση των αρχικών δεδομένων από ν σε 1, χάνουμε ένα μέρος της αρχικής πληροφορίας. Το ζητούμενο είναι η πληροφορία της Z_1 να αποτελεί όσο το δυνατόν μεγαλύτερο μέρος της αρχικής μεταβλητότητας των $X_1, X_2, ..., X_n$.
Έστω οι κ ιδιοτιμές $\lambda_1, \lambda_2, ..., \lambda_k$ του πίνακα συνδιακύμανσης S και

$$u_1, u_2, ..., u_k$$

τα αντίστοιχα ιδιοδιανύσματα. Ο S είναι θετικά ημιορισμένος, άρα $\lambda_j \geq 0, \forall j = 1, 2, ..., k$. Υποθέτουμε $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k \geq 0$. Επειδή ο S είναι συμμετρικός, τα ιδιοδιανύσματα είναι ορθογώνια: $\forall j \neq k, u_j u_k = 0$. Παίρνουμα τα μοναδιαία u_j:

$$u_j' u_j = 1 \forall j = 1, 2, ..., n$$

. Έστω $\mathrm{T} = (u_1, u_2, ..., u_k)$ ο πίνακας ιδιοδιανυσμάτων. Ισχύει ότι $S = \mathrm{T}\Lambda\mathrm{T}'$ ($Su_j = \lambda_j u_j \Rightarrow ST = TL$. Όμως, Τ ορθογώνιος $\Rightarrow TT' = I_k$, άρα $S = STT' = T\Lambda T'$. Γνωστό ως θεώρημα διαγωνοποίησης συμμετρικού πίνακα).
Το διάνυσμα που μεγιστοποιεί την ολική μεταβλητότητα είναι το u_1 που αντιστοιχεί στην μέγιστη ιδιοτιμή του S, την λ_1.

- Απόδειξη:
  Έχουμε $\Sigma = \mathrm{T}\Lambda\mathrm{T}' = \sum_{j=1}^{k} \lambda_j u_j u_j'$ (Θεώρημα φασματικής ανάπτυξης).
  Άρα για $u \in \mathbb{R}^k$ μοναδιαίο είναι:

  $$u' S u = u' (\sum_{j=1}^{k} \lambda_j u_j u_j') u$$

  , αφού $\lambda_1 max_{1 \leq i \leq k} \lambda_i$.
  Εφόσον $\mathrm{TT}' = \sum_{j=1}^{k} u_j u_j' = I_k$, έχουμε $u' S u \leq \lambda_1 u' u = \lambda_1, \forall u \in \mathbb{R}^k$.
  Για u=u_1: $S u_1 = \lambda_1 u_1 \Rightarrow u_1' S u_1 = \lambda_1 u_1' u_1 = \lambda_1$, άρα το $u = u_1$ μεγιστοποιεί την $u' S u$.

Το ποσοστό της ολικής μεταβλητότητας του πίνακα $X^\star$ που διατηρεί η πρώτη κύρια συνιστώσα είναι: $\frac{\cancel{(n-1)}\lambda_1}{\cancel{(n-1)} u' S u} = \frac{\lambda_1}{tr(S)} = \frac{\lambda_1}{\sum_{j=1}^{k} \lambda_j} = \frac{\lambda_1}{\lambda_1 + \lambda_2 + ... + \lambda_n}$.

Δίνουμε ένα παράδειγμα από δεδομένα της $SKlearn$: Έχουμε 150 μετρήσεις για 4 κατηγορίες χαρακτηριστικών από 3 είδη ενός συγκεκριμένου λουλουδιού. Δηλαδή έστω $X_{i,1}, X_{i,2}, ..., X_{i,150}$ και $i = 1, 2, 3, 4$..

Higher Dimension PCA.py - C:/Users/panos/AppData/Local/Programs/Python/Python310/Hi...

File Edit Format Run Options Window Help

```
from sklearn import datasets
import pandas as pd

iris = datasets.load_iris()

#Make a Dataframe, show its first 5 rows.
df = pd.DataFrame(iris.data, columns=iris.feature_names)
df['Target'] = iris.get('target')
df.head()

# Create features and target datasets
features = ['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal
X = df[features].values
y = df['Target'].values
```

Σχήμα 1: Δεδομένα απ' την $iris\ database$.

Δίνουμε τις 5 πρώτες γραμμές των δεδομένων μας χρησιμοποιώντας τον παραπάτω κώδικα:

```
...
   sepal length (cm)  sepal width (cm)  ...  petal width (cm)  Target
0                5.1               3.5  ...               0.2       0
1                4.9               3.0  ...               0.2       0
2                4.7               3.2  ...               0.2       0
3                4.6               3.1  ...               0.2       0
4                5.0               3.6  ...               0.2       0

[5 rows x 5 columns]
>>>
```

Ln: 109 Col: 0

Σε αυτό το σημείο είναι σημαντικό να κανονικοποιήσουμε τα δεδομένα. Ο λόγος είναι ότι η PCA είναι ευαίσθητη σε μεγάλο εύρος διακυμάνσεων. Δηλαδή θα δοθεί πολύ μεγαλύτερο βάρος σε τ.μ. με μεγάλο εύρος διακύμανσης και θα χαθεί η αμεροληψία.

```
from sklearn.preprocessing import StandardScaler

# Standardize the features
X = StandardScaler().fit_transform(X)

# Preview X
pd.DataFrame(data=X, columns=features).head()
```

Σχήμα 2: Κανονικοποιημένα δεδομένα

```
>>> from sklearn.preprocessing import StandardScaler;X = StandardScaler().fit_transform(X);pd.DataFrame(data=X,
columns=features).head()
   sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)
0          -0.900681          1.019004          -1.340227         -1.315444
1          -1.143017         -0.131979          -1.340227         -1.315444
2          -1.385353          0.328414          -1.397064         -1.315444
3          -1.506521          0.098217          -1.283389         -1.315444
4          -1.021849          1.249201          -1.340227         -1.315444
```

Σχήμα 3: Κανονικοποιημένα δεδομένα

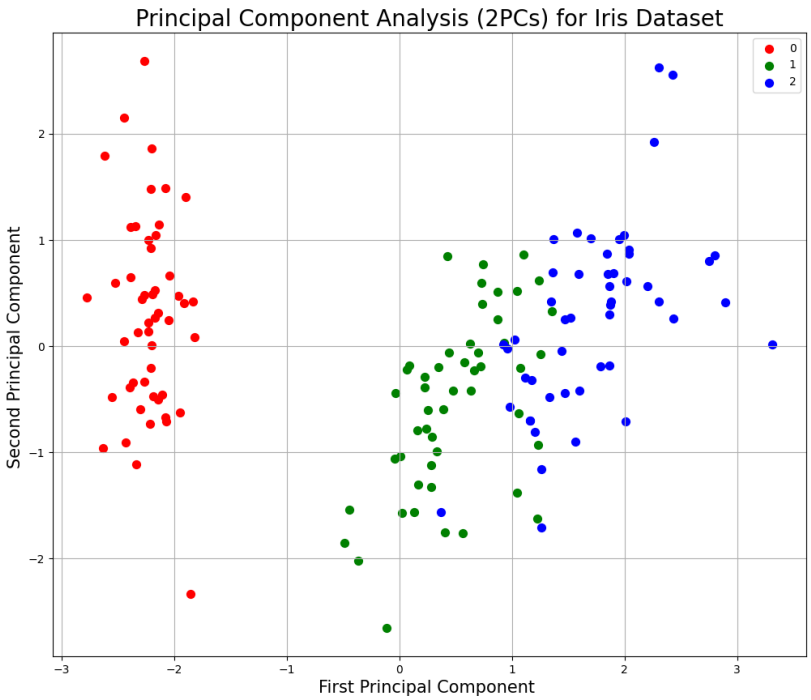
Εδώ είναι το σημείο που υπολογίζονται οι ιδιοτιμές και τα ιδιοδιανύσματα και επιλέγονται οι 2 πρώτες ιδιοτιμές καθώς αυτές επιτυγχάνουν το επίπεδο σημαντικότητας που ζητάμε.
Βλέπουμε τις νέες συνιστώσες. Τα δεδομένα μας έχουν 4 διαστάσεις και τα

```
# Import PCA from sklearn
from sklearn.decomposition import PCA

# Instantiate PCA
pca = PCA(n_components=2)

# Fit PCA to features
principalComponents = pca.fit_transform(X)
```

```
          PC1       PC2  target
0   -2.264703  0.480027       0
1   -2.080961 -0.674134       0
2   -2.364229 -0.341908       0
3   -2.299384 -0.597395       0
4   -2.389842  0.646835       0
```

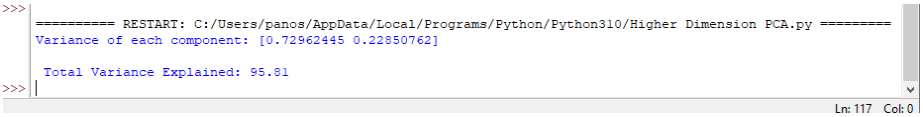
Σχήμα 4: Οι 2 νέες συνιστώσες Z.

προβάλουμε σε 2.
Κάνοντας ένα *scatter plot* βλέπουμε το εξής αποτέλεσμα:

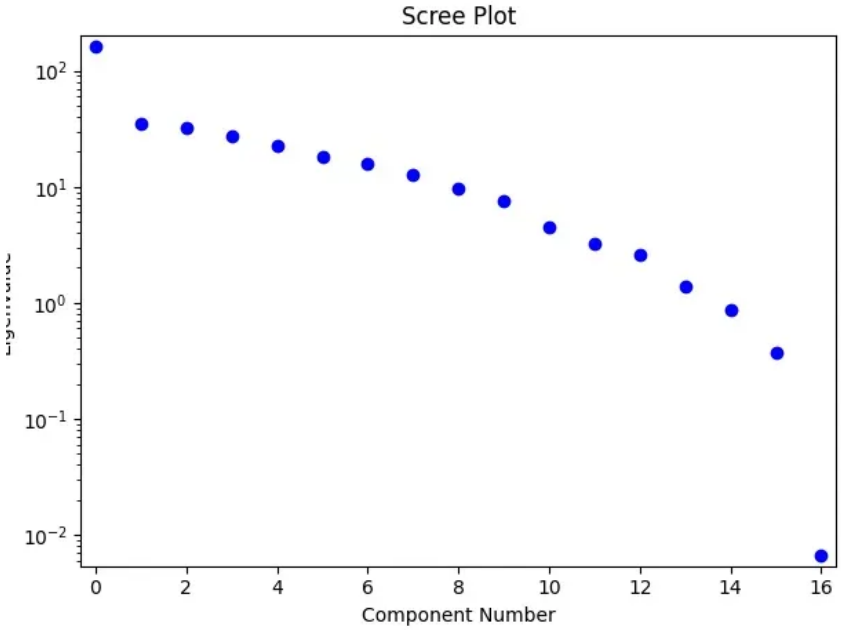


Σχήμα 5: Οι 4 διαστάσεις έγιναν 2.

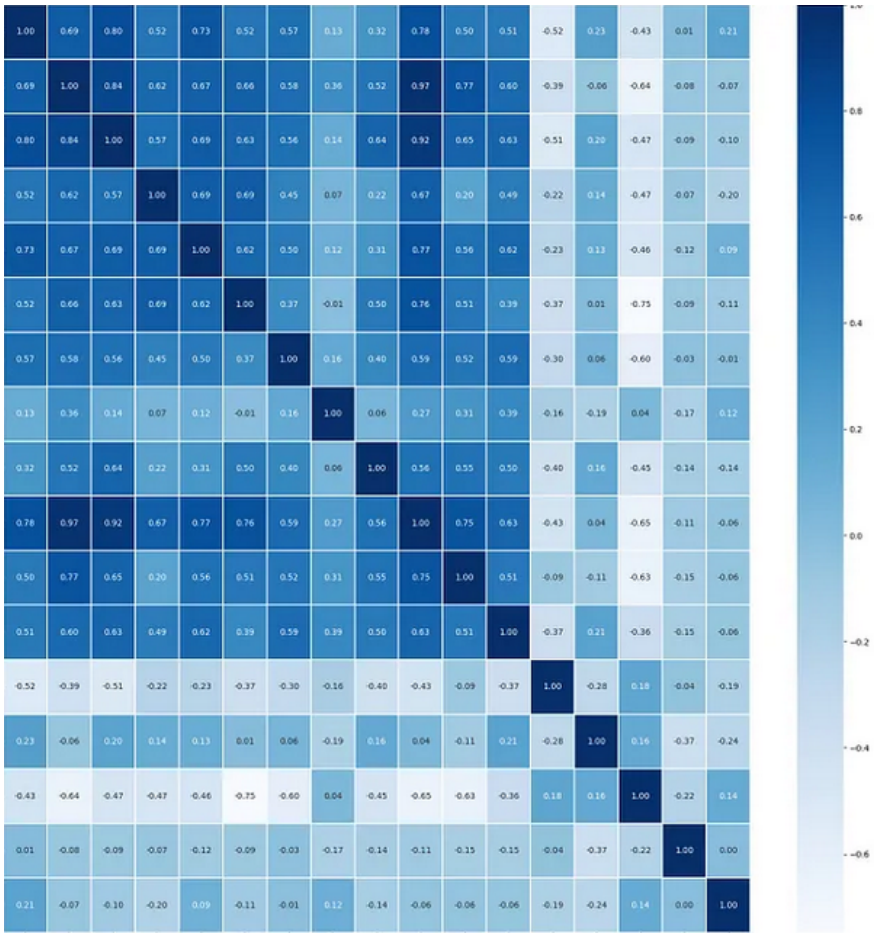
Βλέπουμε πως οι 3 κλάσεις είναι αρκετά καλά διαχωρισμένες. Μπορούμε να πούμε ότι τα δεδομένα μας επιδέχονται ακόμα ένα $classification\ model$. Μπορούμε να συγκρίνουμε την διακύμανση των αρχικών δεδομένων με αυτήν που έδωσαν οι κύριες συνιστώσες ώστε να δούμε τι ποσοστό της μεταβλητότητας διατηρήσαμε.



Σχήμα 6: Η πληροφορία (δηλαδή η μεταβλητότητα που εξηγείται απτην διακύμανση διατηρήθηκε στο 95.81%.



Σχήμα 7: Σε ένα αλλό παράδειγμα βρίσκουμε τις ιδιοτιμές ενός πίνακα διακύμανσης και της προβάλουμε αύξουσα σειρά. Παρατηρούμε πως στους ¨πρόποδες' του γραφήματος η ερμηνεία είναι ότι οι κύριες συνιστώσες που χαρακτηρίζονται από τις $\lambda_2, \lambda_3, ..., \lambda_{15}$ προσθέτουν μικρή ακόμα αξία στην διατήρηση της ολικής μεταβλητότητας. Η διαφορά όμως της λ_1, λ_2 είναι σημαντική. Αυτό μας προτείνει ότι δύο συνιστώσες εξηγούν ένα μεγάλο κομμάτι της αρχικής μεταβλητότητας(Εδώ π.χ. η κύρια συνιστώσα που χαρακτηρίζει η λ_1 μπορεί να εξηγεί 81% της αρχικής μεταβλητότητας. Η λ_2 88.3%, ενώ από λ_3 και έπειτα θα είναι σαν 89.4%, 89.9%, 90%, 90.01% κλπ). Ξέρουμε πως η λ_{16} αποδίδει πολύ λίγο και προτείνει ότι υπάρχει θόρυβος/διπλότυπα και δεδομένα που συμβάλουν λίγο στην διακύμανση.)



Σχήμα 8: Εδώ έχουμε έναν πίνακα συνδιακύμανσης που προέρχεται από δεδομένα με πολλές διαστάσεις. Είναι χρήσιμο να βάλουμε έναν απλό χρωματισμό για να καταλάβουμε την σχέση που έχουν οι διάφορες διαστάσεις των δεδομένων. Εδώ το πιο βαθύ μπλε χαρακτηρίζει ισχυρή συσχέτιση και ασθενέστερη στο ανοικτό μπλε. Αυτό μπορεί να σημαίνει το εξής: Έστω δυο διαστάσεις έχουν μετρήσεις για το ύψος και το βάρος ενός πληθυσμού. Εκεί θα παρατηρούσαμε ισχυρή συσχέτιση (σκούρο μπλε) καθώς το ύψος και το βάρος είναι άμεσα συνδεδεμένα, ή αλλιώς η μεταβολή της διακύμανσης του ενός σημαίνει την μεταβολή της διακύμανσης του άλλου. Σε άλλη περίπτωση για διαστάσεις με μετρήσεις για την διάρκεια ζωής μιας λάμπας, ο παράγοντας ''πιθανότητα εργοστασίακου σφάλματος'' έχει πολύ μικρότερη συνμεταβλητότητα από τον ''ώρες λειτουργίας'' σε σχέση με την ''Διάρκεια ζωής λάμπας''.

Γραμμική Διαχωριστική ανάλυση

Η Γραμμική Διαχωριστική ανάλυση αποτελεί μια *supervised* ανάλυση μείωσης διαστάσεων που είναι η γενίκευση του διαχωριστικού κανόνα του Φισηερ και είναι η μέθοδος εκείνη που χρησιμοποιείται στη στατιστική για την εύρεση ενός γραμμικού συνδυασμού των χαρακτηριστικών που διαχωρίζουν δύο ή περισσότερες ομάδες αντικειμένων. Ο προκύπτων συνδυασμός μπορεί να χρησιμοποιηθεί ως ένας γραμμικός ταξινομητής ή, πιο συχνά, για τη μείωση των διαστάσεων πριν από την μεταγενέστερη ταξινόμηση. Στο πλαίσιο αυτό, η Γραμμική Διαχωριστική Ανάλυση όχι μόνο κατατάσσει τις παρατηρήσεις σε ομάδες, αλλά προσδιορίζει και την πιθανότητα με την οποία γίνεται η κατάταξη αυτή, υπό την προϋπόθεση ότι τα δεδομένα ακολουθούν πολυμεταβλητή κανονική κατανομή.
Η Γραμμική Διαχωριστική Ανάλυση, ως στατιστική τεχνική ταξινόμησης παρατηρήσεων, η οποία όμως προϋποθέτει τον εκ των προτέρων διαχωρισμό των δεδομένων σε δύο ή περισσότερες ομάδες, μας επιτρέπει να μελετήσουμε τις διαφορές μεταξύ των ομάδων αυτών λαμβάνοντας υπόψη ένα σύνολο μεταβλητών που περιγράφουν τα χαρακτηριστικά των ομάδων. Με βάση το παραπάνω, προκύπτει ότι η Γραμμική Διαχωριστική Ανάλυση σχετίζεται με την $ANOVA$ και την Γραμμική Παλινδρόμηση, οι οποίες επίσης προσπαθούν να εκφράσουν την εξαρτημένη μεταβλητή ως γραμμικό συνδυασμό των άλλων χαρακτηριστικών ή μετρήσεων. Όμως, η $ANOVA$ έχει ως εξαρτημένη μεταβλητή, ποσοτική μεταβλητή, και ως ανεξάρτητες, κατηγορικές, ενώ η Γραμμική Διαχωριστική Ανάλυση έχει ποσοτικές ανεξάρτητες μεταβλητές και μια κατηγορική μεταβλητή που δηλώνεται η ομάδα ($class\,label$). Επίσης, η Γραμμική Διαχωριστική Ανάλυση σχετίζεται με την Ανάλυση Κύριων Συνιστωσών, καθώς και οι δύο τεχνικές αναζητούν γραμμικούς συνδυασμούς των δεδομένων που εξηγούν με τον καλύτερο τρόπο τα δεδομένα. Στο πλαίσιο αυτό, η Γραμμική Διαχωριστική Ανάλυση προσπαθεί να μοντελοποιήσει με τον καλύτερο τρόπο τις διαφορές μεταξύ των ομάδων, ενώ η Ανάλυση Κύριων Συνιστωσών δε λαμβάνει υπόψη καμία διαφοροποίηση μεταξύ ομάδων.
Γενικά, η μέθοδος εφαρμόζεται χωρίς ιδιαίτερους ελέγχους αρκεί να μην υπάρχουν ακραίες παρατηρήσεις. Επιπρόσθετα υποθέτουμε:

- Οι μεταβλητές να ακολουθούν την πολυμεταβλητή κανονική κατανομή. (ή τουλάχιστον προτιμούμε ένα αρκετά μεγάλο δείγμα).
- Να μην υπάρχουν μεγάλες διαφορές στην διακύμανση.
- Να αποφεύγεται η πολυσυγγραμμικότητα.
- Καμία μεταβλητή δεν πρέπει να είναι γραμμικός συνδυασμός των υπολοίπων διακριτών μεταβλητών.

Όμοια με παραπάνω κατασκευάζουμε νέες μεταβλητές
$Z_k = a_{k1}X_1 + a_{k2}X_2 + ... + a_{kn}X_n$ από το αρχικό δείγμα $X_1, X_2, ..., X_n$ οι οποίοι διαχωρίζουν τα δεδομένα στις ομάδες και μπορούμε να αξιολογήσουμε τη σημασία κάθε μεταβλητής στον διαχωρισμό των ομάδων.

Οι γραμμικοί συνδυασμοί που προκύπτουν ονομάζονται κανονικές διαχωριστικές συναρτήσεις ($canonical\, discriminant\, functions$) ή κανονικές μεταβλητές ($canonical\, variables$) και ο μέγιστος αριθμός τους ισούται με τον αριθμό των ομάδων μείον ένα ή με τον αριθμό των μεταβλητών, στην περίπτωση που αυτός είναι μικρότερος από τον αριθμό των ομάδων.
Ακόμη, ισχύει ότι η πρώτη κανονική μεταβλητή Z_1, δίνει το μέγιστο $F - ratio$ $(between - classvariance/within - classvariance)$, καθώς και ότι οι προκύπτουσες κανονικές μεταβλητές $(Z_1, Z_2, .., Z_{n-1})$ είναι ασυσχέτιστες μεταξύ των ομάδων. Τέλος, κρίνεται αναγκαίο να αναφερθεί ότι οι διακριτές τιμές $(discriminant/, scores)$ που προκύπτουν από τις διαχωριστικές συναρτήσεις συνήθως υπολογίζονται σε τυπικές τιμές, οπότε η διακριτική τιμή μιας παρατήρησης αντιπροσωπεύει τον αριθμό των τυπικών αποκλίσεων κατά τον οποίο η παρατήρηση απέχει από το κέντρο της ομάδας.
Ας υποθέσουμε την $two - class\, LDA$. Έχουμε λοιπόν το σετ $X_1, X_2, ..., X_n$ (εδώ για ευκολία μπορούμε να υποθέσουμε ότι οι μισές τ.μ. $X_1, X_2, ..., X_{\frac{n}{2}} \sim N_2(2, 1)$ και $X_{\frac{n}{2}+1}, X_{\frac{n}{2}+2}, ..., X_n \sim N_2(1, 1)$. Αυτό το κάναμε για να υπάρχει ένας διαχωρισμός των δεδομένων. Πιο ρεαλιστικά μπορούμε να εφαρμόσουμε στα αρχικά δεδομένα άλλες τεχνικές ομαδοποίησης όπως $k - means$.)
Στόχος λοιπόν είναι να προβάλουμε τις τιμές σε μια ευθεία ώστε να είναι εμφανής ο διαχωρισμός τους αλλά και το διάστημα στο οποίο έχουμε επικάλυψη. Βρίσκουμε λοιπόν το διάνυσμα v που περιγράφει 'καλύτερα' αυτόν τον διαχωρισμό.
Χ.β.τ.γ. υποθέτω ένα μοναδιαίο διάνυσμα v από την αρχή των αξόνων και παίρνω τις προβολές των σημείων σε αυτό. Καταλήγω σε διανύσματα $\alpha_i = v'X_i, i = 1, 2, ..., n..$ Ένας απλός τρόπος να τα διαχωρίσω είναι να μετρήσω την απόσταση των μέσων $|\mu_1 - \mu_2|$ όπου:

$$\mu_1 = \frac{1}{n_1} \sum_{x_i \in C_1} \alpha_i = \frac{1}{n_1} \sum_{x_i \in C_1} v'X_i = v' \frac{1}{n_1} \sum_{X_i \in C_1} X_i = v'm_1$$

(εδώ $C_1 = X_i : i = 1, 2, ..., \frac{n}{2}$. ή γενικότερα η μία κλάση. Όμοια $C_2 = X_i : i = \frac{n}{2} + 1, \frac{n}{2} + 2, ..., n.$)
Όμοια:

$$\mu_2 = v'm_2,\ m_2 = \frac{1}{n_2} \sum_{X_i \in C_2}$$

Έτσι λύνουμε το πρόβλημα:

$$\max_{||v||=1} |\mu_1 = \mu_2|$$

.
Επειδή η παραπάνω μέθοδος δεν είναι πάντα αποδοτική (έχουμε διαφορετικά διαστήματα ανάλογα με τη θέση του διανύσματος v), υπολογίζουμε και τις διακυμάνσεις.
Έχουμε:

$$s_1^2 = \sum_{X_i \in C_1} (\alpha_1 - \mu_1)^2, \sigma_2 = \sum_{X_i \in C_2} (\alpha_2 - \mu_2)^2$$

(Ιδανικά θα έχουμε απομακρυσμένους μέσους και μικρές διακυμάνσεις).
Έτσι λύνουμε την:

$$\max_{||v||=1} \frac{|\mu_1 = \mu_2|}{\sigma_1^2 - \sigma_2^2}$$

Με άλλα λόγια μεγιστοποιούμε την απόσταση του μέσου και ελαχιστοποιούμε τις διακυμάνσεις.
Εδώ εμφανίζονται οι $between - class\ scatter\ matrix$,
$within - class\ scatter\ matrix$ και $total\ within - class\ scatter\ matrix$:
Έχω:

$$(\mu_1 = \mu_2)^2 = (v'\mu_1 - v'\mu_2)^2 = (v'(\mu_1 - \mu_2))^2 = v'(\mu_1 - \mu_2)(\mu_1 - \mu_2)'v = v'S_b v$$

Όπου $S_b = (\mu_1 - \mu_2)(\mu_1 - \mu_2)'$ ο $between - class\ scatter\ matrix$
Και:

$$s_j^2 = \sum_{X_i \in C_j} (\alpha_i - \mu_j)^2 = \sum_{X_i \in C_j} (v'X_i - v'\mu_j)^2 = \sum_{X_i \in C_j} v'(X_i - \mu_j)(X_i - \mu_j)'v$$

$$= v' \sum_{X_i \in C_j} (X_i - \mu_j)(X_i - \mu_j)'v = v'S_j v$$

Όπου S_j ο $within - class\ scatter\ matrix$.
Επίσης:

$$\sigma_1^2 - \sigma_2^2 = v'S_1 v + v'S_2 v = v'(S_1 + S_2)v = v'S_w v$$

. Όπου:

$$S_w = S_1 + S_2 = \sum_{X_i \in C_1} (X_i - \mu_1)(X_i - \mu_1)' + \sum_{X_i \in C_2} (X_i - \mu_2)(X_i - \mu_2)'$$

ο $total\ within - class\ scatter\ matrix$.
Τέλος το πρόβλημα γίνεται:

$$\max_{||v||=1} \frac{v'S_b v}{v'S_w v}$$

Δίνουμε ένα παράδειγμα από τα ίδια δεδομένα της $SKlearn$: Έχουμε 150 μετρήσεις για 4 κατηγορίες χαρακτηριστικών από 3 είδη ενός συγκεκριμένου λουλουδιού (Εδώ δίνουμε και τα είδη αυτά:$setosa, versicolor, virginica$). Δηλαδή έστω $X_{i,1}, X_{i,2}, ..., X_{i,150}$ και $i = 1, 2, 3, 4$.

```
#load iris dataset
iris = datasets.load_iris()

#convert dataset to pandas DataFrame
df = pd.DataFrame(data = np.c_[iris['data'], iris['target']],
                  columns = iris['feature_names'] + ['target'])
df['species'] = pd.Categorical.from_codes(iris.target, iris.target_names)
df.columns = ['s_length', 's_width', 'p_length', 'p_width', 'target', 'species']

#view first six rows of DataFrame
df.head()

#define predictor and response variables
X = df[['s_length', 's_width', 'p_length', 'p_width']]
y = df['species']

#Fit the LDA model
model = LinearDiscriminantAnalysis()
model.fit(X, y)
```

Σχήμα 9: Εφαρμόζουμε LDA

Αφού έχουμε πραγματοποιούμε LDA στον πίνακα με τα δεδομένα $iris$, θα κάνουμε ένα $k - fold\ cross\ validation$ για να δούμε τι ακρίβεια έχει το μοντέλο μας.

```
#Define method to evaluate model
cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)

#evaluate model
scores = cross_val_score(model, X, y, scoring='accuracy', cv=cv, n_jobs=-1)
print(np.mean(scores))
```

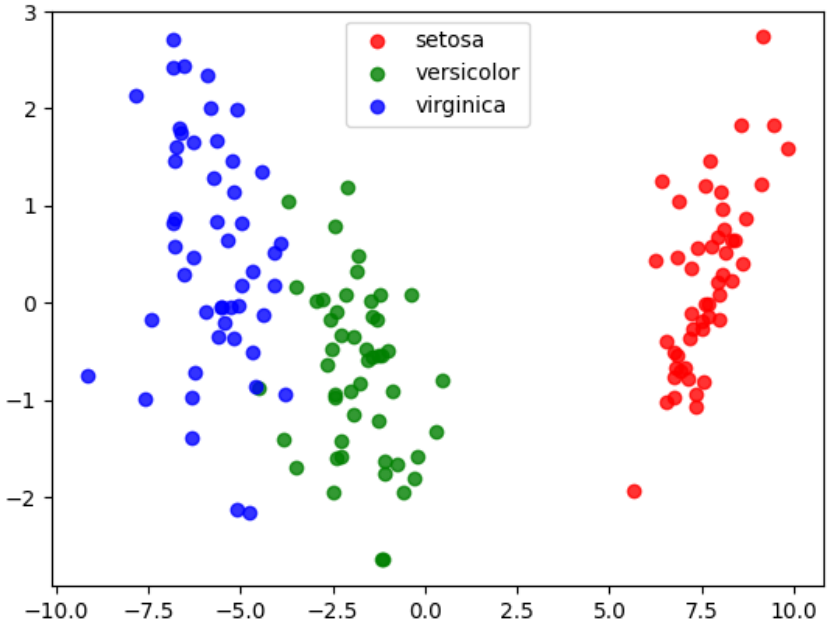
Σχήμα 10: Με την $k-fold\ cross\ validation$ βλέπουμε πόσο % ακρίβεια παίρνουμε.

Σε αυτή τη μέθοδο θα δούμε ότι μπορούμε να κάνουμε και προβλέψεις για νέα δεδομένα. Δηλαδή για μια νέα παρατήρηση θα δείξουμε σε ποια ομάδα ανήκει με ε.σ.σ. $\sim 98\%$.

```
#define new observation
new = [5, 3, 1, 4]

#predict which class the new observation belongs to
model.predict([new])
```

Σχήμα 11: Εισάγουμε τα χαρακτηριστικά ενός νέου λουλουδιού new. Η μέθοδος μας απαντά σε ποια κατηγορία ανήκει με ε.σ.σ $\sim 98\%$.

Οπτικοποιούμε το αποτέλεσμα και παρατηρούμε το εξής:



Σχήμα 12: Κανονικοποιημένα δεδομένα

Βλέπουμε και πάλι ότι δεδομένα που προηγουμένως είχαν 4 διαστάσεις πλέον περιγράφονται από 2 χωρίς σημαντική απώλεια πληροφορίας.
Η εικόνα είναι αρκετά όμοια με εκείνη της PCA αφού άλλωστε οι δύο μέθοδοι έχουν κοινό στόχο μείωσης διαστάσεων.