Data Reduction on the Iris Dataset

April 15, 2021

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import (datasets, decomposition, ensemble, discriminant_analysis)
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from mpl_toolkits.mplot3d import Axes3D # needed to modify the way figure␣
↪behaves
```

[2]:
```
# load data from scikit-learn
iris = datasets.load_iris()
X = iris.data
y = iris.target
target_names = iris.target_names
```

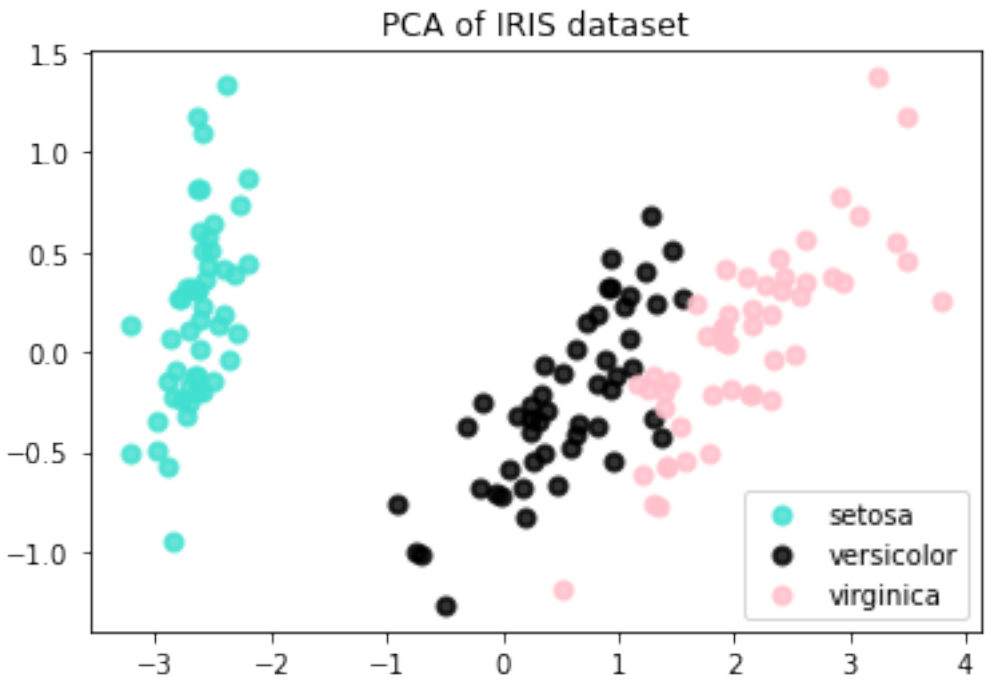
1 Part1 - PCA

1.1 1. Apply PCA projection of the features of IRIS dataset in 2 dimensions

[3]:
```
# print("Computing PCA projection..."),
pca = decomposition.PCA(n_components=2)
X_pca = pca.fit_transform(X)
# print("done.")

# Plot PCA result
colors = ['turquoise', 'black', 'pink']
lw = 2

for color, i, target_name in zip(colors, [0, 1, 2], target_names):
    plt.scatter(X_pca[y == i, 0], X_pca[y == i, 1], color=color, alpha=.8,␣
↪lw=lw,
                label=target_name)
plt.legend(loc='best', shadow=False, scatterpoints=1)
plt.title('PCA of IRIS dataset')
plt.show()
```



1.2 2. Show how much variance ratio is explained by the reduced Dimension

[4]:
```
print(pca.explained_variance_ratio_)
```

```
[0.92461872 0.05306648]
```

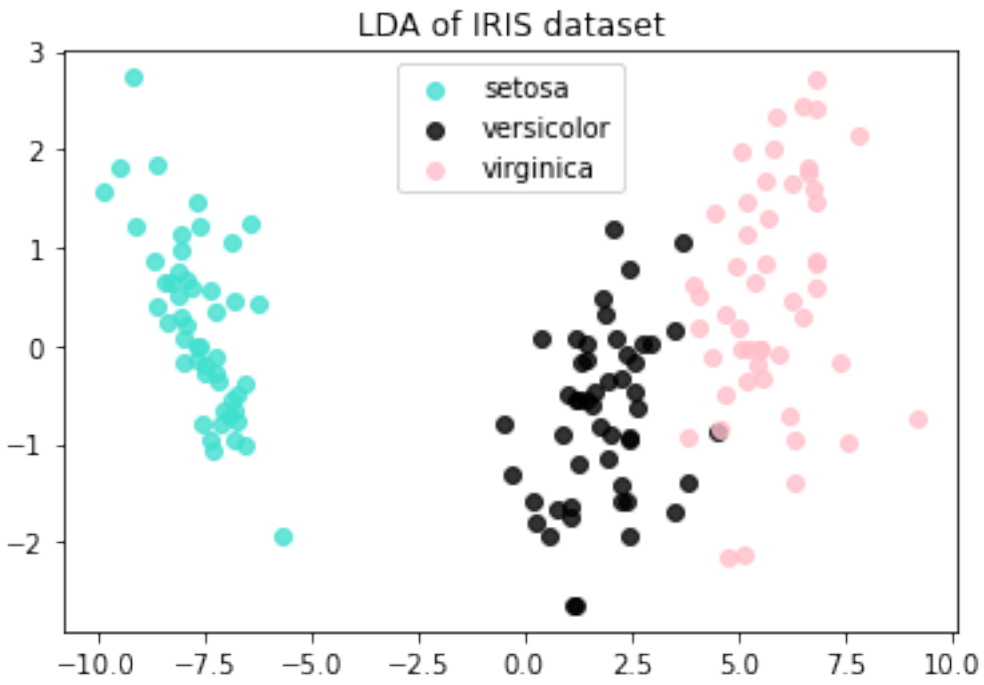
2 Part 2 - LDA

2.1 1. Apply LDA projection of the features of IRIS dataset in 2 dimensions

[5]:
```
# print("Computing LDA..."),
lda =  LinearDiscriminantAnalysis(n_components=2)
X_lda = lda.fit(X, y).transform(X)
# print("done.")

# Plot LDA result
for color, i, target_name in zip(colors, [0, 1, 2], target_names):
    plt.scatter(X_lda[y == i, 0], X_lda[y == i, 1], alpha=.8, color=color,
                label=target_name)
plt.legend(loc='best', shadow=False, scatterpoints=1)
plt.title('LDA of IRIS dataset')

plt.show()
```



2.2 2. Show how much variance ratio is explained by the reduced dimension

[6]:
```
print(lda.explained_variance_ratio_)
```

```
[0.9912126 0.0087874]
```

3 Part 3 Comparison of PCA and LDA

1. Compare the variance ratio explained by the 2-dimensions of the methods you have used. Which is better?

   LDA is better since it explains more variances of the first principle component.

2. Compare the scatter plot of the two methods after reduction. Which is a better methodfor separating the different classes of data?

   LDA is better since it gives a clearer boundary between classes (especially the boudary between versicolor and virginica).

3. What is the primary difference between the two methods? Which method works betterin this case and why?

   Principal Component Analysis (PCA) is unsupervised method, it "ignores" class labels and identifies the attributes maxaimize the vairiance in the dataset. Linear Discriminant Analysis (LDA) is supervised method, it identifies attributes account for the most variance between classes. In this case, LDA is better because LDA can separate the two normally distributed classes well, and it also works well as the data's class labels are known.