Linear Discriminant Analysis: Titanic

Kaggle

Maria Aquino
Alexander Montenegro
Cris Navarijo
Arin Parsanian

NSF Data Science Summer Institute B, June 2019

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

Table of Contents

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

Table of Contents

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

What is LDA?

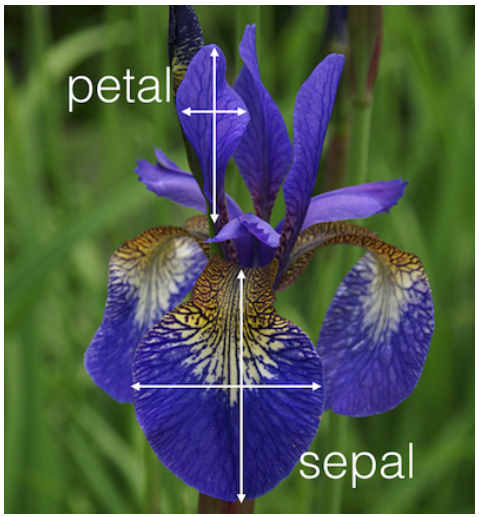
LDA can be used as a dimensional reduction technique similar to PCA or for classification tasks by making three crucial assumptions:

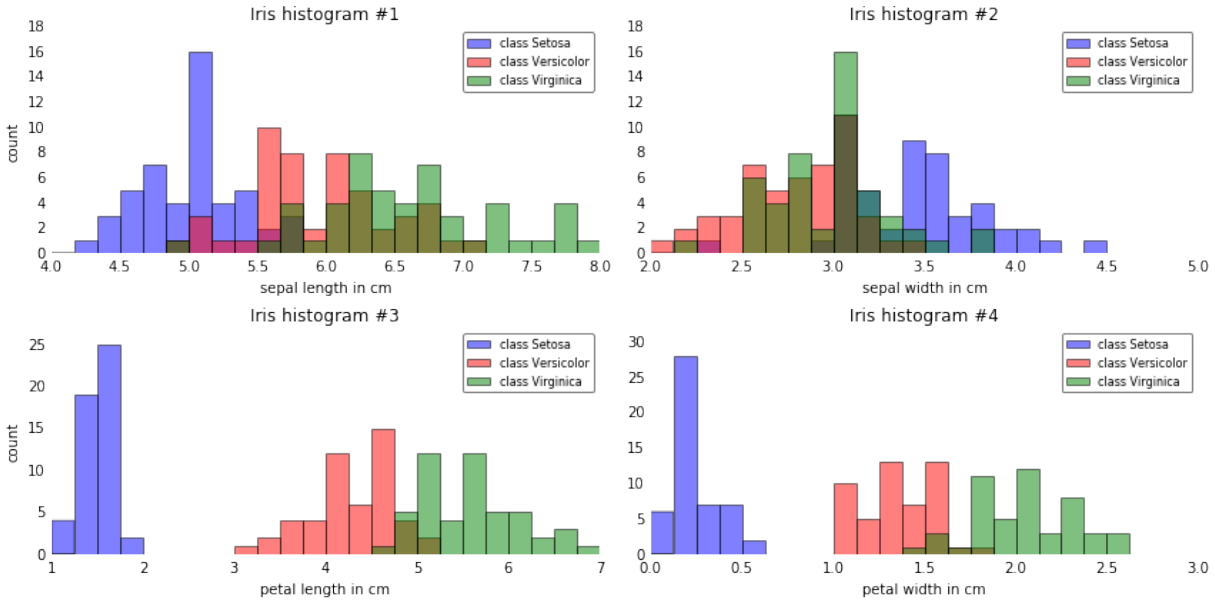
- Data is normally distributed
- Features are statistically independent
- Has identical covariance matrices for each class

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

What does LDA do for classification?

Example of ideal/non-ideal distributions for LDA:





What does that mean?

The model is limited in 2 major ways

- Data is assumed normally distributed
- Data must contain at least 1 continuous random variables to fit the model (IE, cannot use categorical data exclusively)

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

Table of Contents

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:
```
DF = pd.read_csv("train.csv")
print(DF.shape)
DF[:10]
```

```
(891, 12)
```

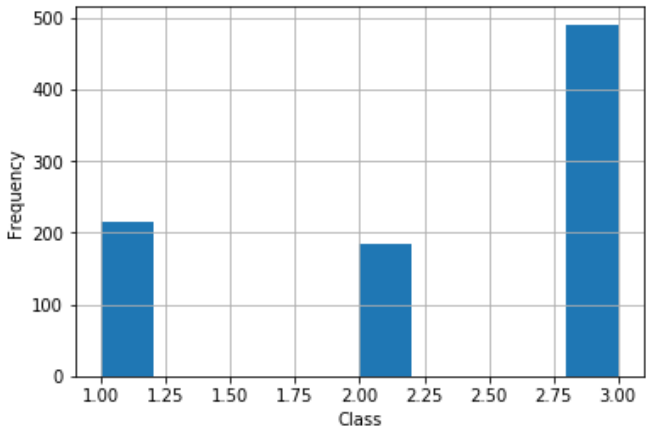
Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

In [3]:
```
Y = np.array(DF["Survived"])
X = pd.DataFrame([DF["Sex"], DF["Age"], DF["Fare"]]).T
```

In [4]:
```
plt.hist(DF["Pclass"])
plt.xlabel("Class")
plt.ylabel("Frequency")
plt.grid();
```

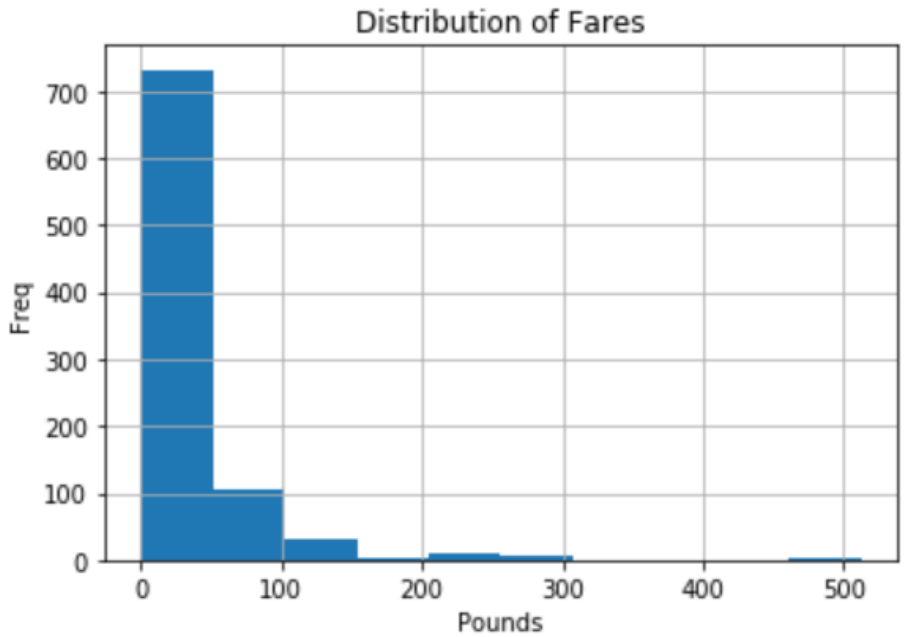


In [5]:
```
plt.scatter(DF["Pclass"], DF["Fare"])
plt.grid()
```

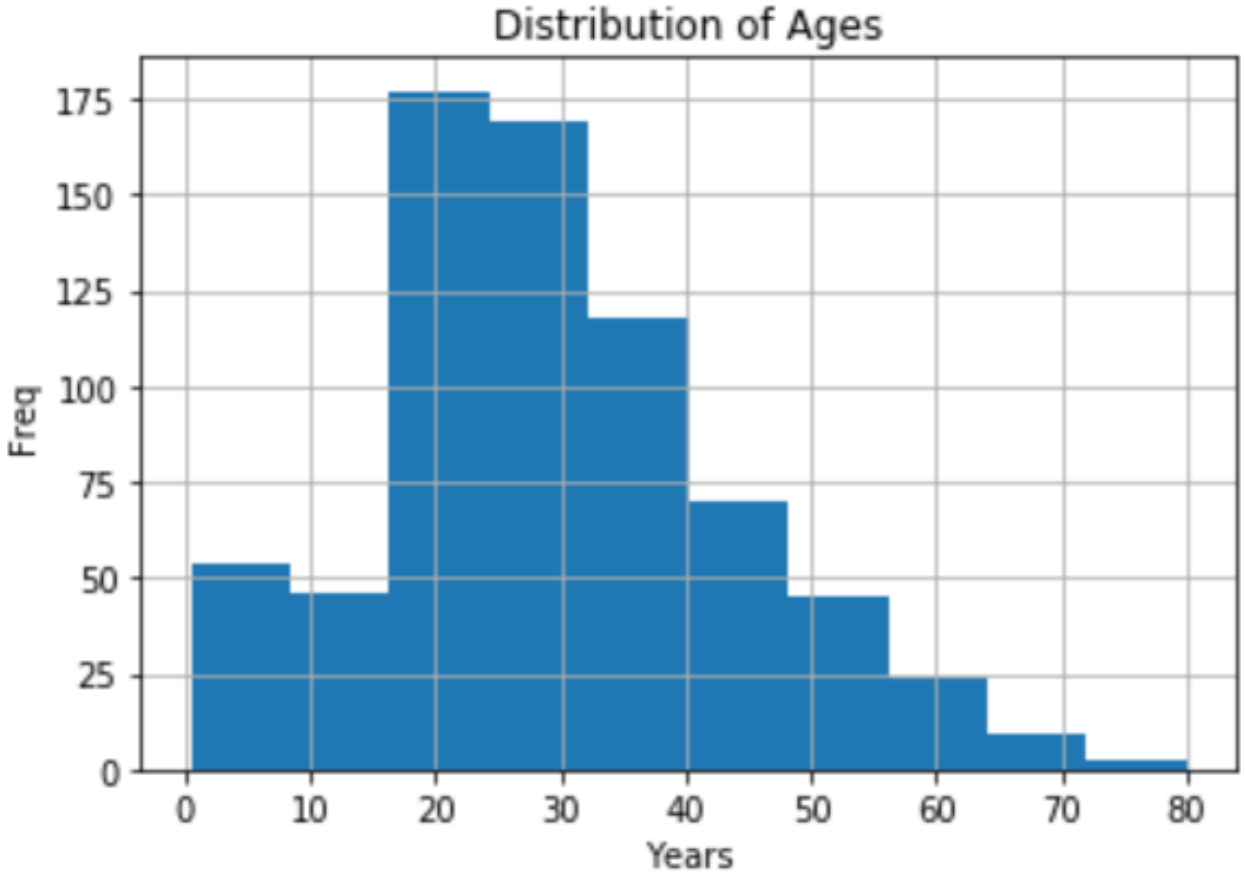
CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

In [24]:

```
plt.hist(DF['Fare'])
plt.title("Distribution of Fares")
plt.xlabel('Pounds')
plt.ylabel('Freq')
plt.grid();
```



CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE



CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

In [7]:
```
namedata
```

Out[7]:

	Surname/title	First/Middle
0	Braund, Mr	Owen Harris
1	Cumings, Mrs	John Bradley (Florence Briggs Thayer)
2	Heikkinen, Miss	Laina
3	Futrelle, Mrs	Jacques Heath (Lily May Peel)
4	Allen, Mr	William Henry
5	Moran, Mr	James
6	McCarthy, Mr	Timothy J
7	Palsson, Master	Gosta Leonard
8	Johnson, Mrs	Oscar W (Elisabeth Vilhelmina Berg)
9	Nasser, Mrs	Nicholas (Adele Achem)
10	Sandstrom, Miss	Marguerite Rut
11	Bonnell, Miss	Elizabeth

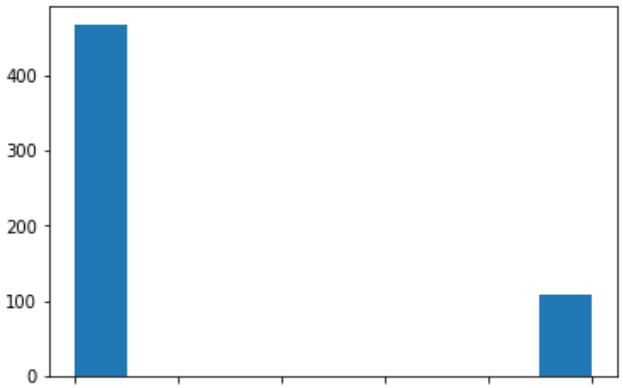
CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

Some broad data visualization

In [10]:
```
ismale = DF["Sex"]=="male"
males = DF[ismale]
females = DF[ismale == False]
```

In [11]:
```
plt.hist(males["Survived"]);
print("Male survival rate: ", (len(males[males["Survived"] == 1])/len(males))*100, "%")
```

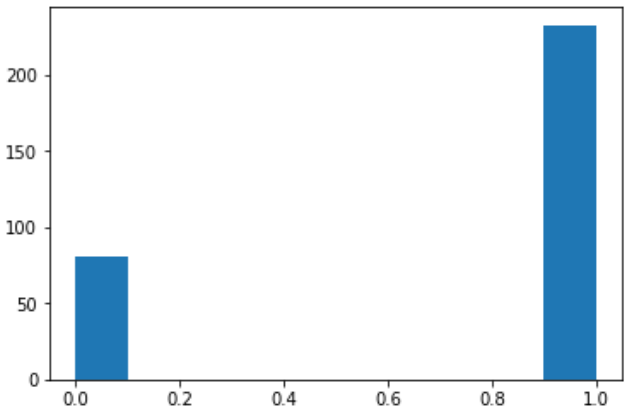
```
Male survival rate:  18.890814558058924 %
```



CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

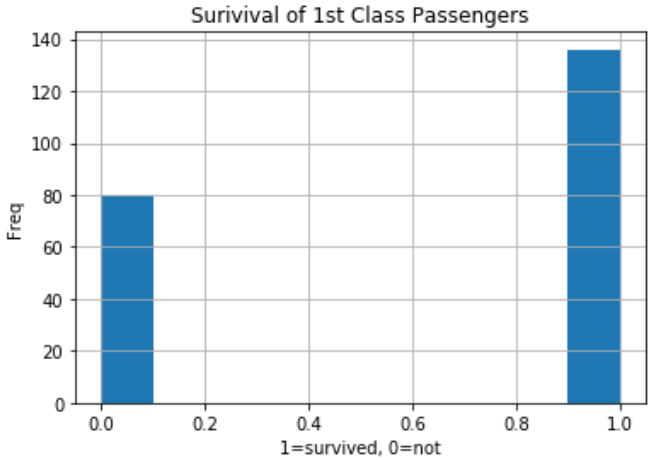
In [12]:
```
plt.hist(females["Survived"]);
print("Female Survival rate: ",len(females[females["Survived"] == 1])/len(DF)*100, "%" )
```

```
Female Survival rate:  26.15039281705948 %
```



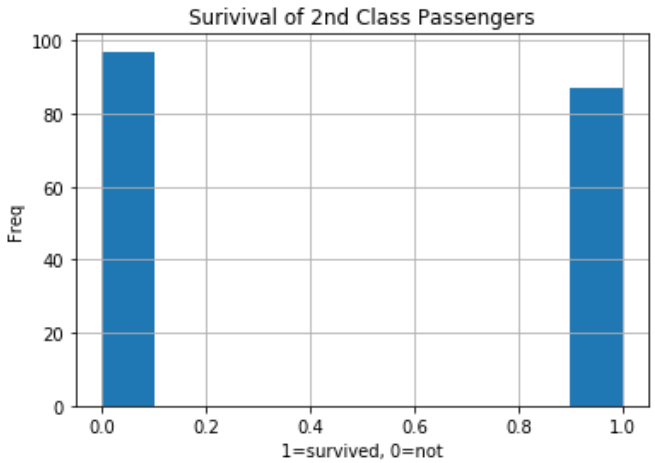
In [13]:
```
plt.hist(DF[DF["Title"]=="Master"]["Survived"])
plt.title("Survival of Young Men (< 12yo)")
plt.xlabel("Survived")
plt.ylabel("Frequency")
plt.grid();
```

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE



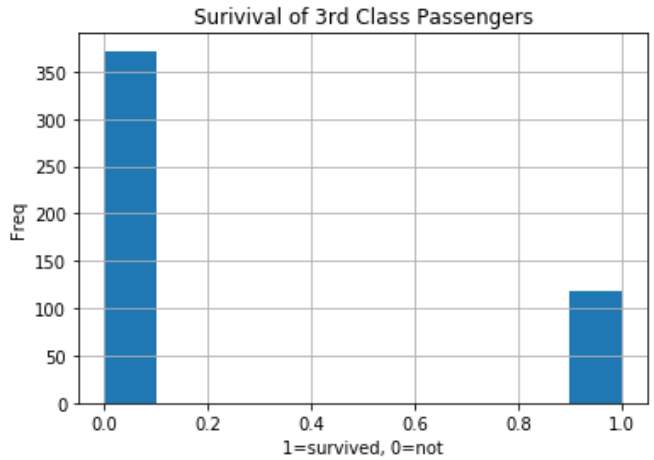
In [22]:
```
plt.hist(secondclass['Survived'])
plt.title("Surivival of 2nd Class Passengers")
plt.xlabel('1=survived, 0=not')
plt.ylabel('Freq')
plt.grid();
```

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE



In [23]:
```
plt.hist(thirdclass['Survived'])
plt.title("Surivival of 3rd Class Passengers")
plt.xlabel('1=survived, 0=not')
plt.ylabel('Freq')
plt.grid();
```

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE



In [24]:
```
for x in np.unique(DF["Title"]):
    tokenDF = pd.DataFrame(DF[DF["Title"]==x])
    averageage = np.mean(tokenDF["Age"])
    empties = tokenDF[tokenDF["Age"].isnull()]
    DF[(DF["Title"]==x) & (DF["Age"].isnull())] = DF[(DF["Title"]==x) & (DF["Age"].isnull())].fillna(value = averageage)
```

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

Training the model

In [25]:
```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score,confusion_matrix
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
import warnings
warnings.filterwarnings("error")
```

In [26]:
```
errs=[]
nsplits=100
skipped=0
X = pd.DataFrame([DF["Sex"], DF["Age"], DF["Pclass"]]).T

# Casting categorical data to numerical categories
X['Sex']= X['Sex'].astype('category')
X['Sex'] = X['Sex'].cat.codes


Y= np.array(DF["Survived"])
for j in range(nsplits):
  try:
    XTRAIN, XTEST, YTRAIN, YTEST=train_test_split(X,Y, test_size = .2)
    LDA = LinearDiscriminantAnalysis(solver = "lsqr", shrinkage = "auto")
    LDA.fit(XTRAIN,YTRAIN)
    YP=LDA.predict(XTEST)
    errs.append(1-accuracy_score(YTEST,YP))
  except UserWarning:
    skipped+=1
```

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

```
print("LDA mean error=%7.5f std=%7.5f" %(np.mean(errs),np.std(errs)))
print(skipped,"train/test splits had to be skipped because of Normalization Errors")
```

```
LDA mean error=0.21318 std=0.03139
0 train/test splits had to be skipped because of Normalization Errors
```

In [27]:
```
X
```

Out[27]:

	Sex	Age	Pclass
0	1	22	3
1	0	38	1
2	0	26	3
3	0	35	1
4	1	35	3
5	1	32.3681	3
6	1	54	1
7	1	2	3
8	0	27	3
9	0	14	2
10	0	4	3
11	0	58	1

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

Prediction to upload

In [28]:
```
testDF = pd.read_csv("test.csv")
```

Cleaning test data

In [29]:
```
testnames = []
for x in np.array(testDF["Name"]):
    tokens = x.split('. ', maxsplit = 1)
    testnames.append(tokens)
testnamedata = pd.DataFrame(testnames, columns = ["Surname/title", "First/Middle"])

testsurnamesandtitles = pd.DataFrame(testnamedata["Surname/title"])
testnames2 = []
for x in np.array(testsurnamesandtitles["Surname/title"]):
    tokens = x.split(', ', maxsplit = 1)
    testnames2.append(tokens)
testnames2 = pd.DataFrame(testnames2, columns = ["Surname", "Title"])
np.unique(testnames2["Title"])
testDF = pd.concat([testDF, testnames2], axis = 1)
```

In [30]:
```
testDF
```

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

Out[30]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Surname	Title
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q	Kelly	Mr
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S	Wilkes	Mrs
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q	Myles	Mr
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S	Wirz	Mr
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S	Hirvonen	Mrs
5	897	3	Svensson, Mr. Johan Cervin	male	14.0	0	0	7538	9.2250	NaN	S	Svensson	Mr
6	898	3	Connolly, Miss. Kate	female	30.0	0	0	330972	7.6292	NaN	Q	Connolly	Miss
7	899	2	Caldwell, Mr. Albert Francis	male	26.0	1	1	248738	29.0000	NaN	S	Caldwell	Mr
8	900	3	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	female	18.0	0	0	2657	7.2292	NaN	C	Abrahim	Mrs
9	901	3	Davies, Mr. John Samuel	male	21.0	2	0	A/4 48871	24.1500	NaN	S	Davies	Mr

In [31]:
```
for x in np.unique(testDF["Title"]):
    tokenDF = pd.DataFrame(testDF[testDF["Title"]==x])
    averageage = np.mean(tokenDF["Age"])
    empties = tokenDF[tokenDF["Age"].isnull()]
    testDF[(testDF["Title"]==x) & (testDF["Age"].isnull())] = testDF[(testDF["Title"]==x) & (testDF["Age"].isnull())].fillna(val
```

In [32]:
```
testDF[(testDF["Title"]=="Ms") & (testDF["Age"].isnull())] = testDF[(testDF["Title"]=="Ms") & (testDF["Age"].isnull())].fillna(v
testDF
```

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

Out[32]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Surname	Title
0	892	3	Kelly, Mr. James	male	34.500000	0	0	330911	7.8292	NaN	Q	Kelly	Mr
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.000000	1	0	363272	7.0000	NaN	S	Wilkes	Mrs
2	894	2	Myles, Mr. Thomas Francis	male	62.000000	0	0	240276	9.6875	NaN	Q	Myles	Mr
3	895	3	Wirz, Mr. Albert	male	27.000000	0	0	315154	8.6625	NaN	S	Wirz	Mr
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.000000	1	1	3101298	12.2875	NaN	S	Hirvonen	Mrs
5	897	3	Svensson, Mr. Johan Cervin	male	14.000000	0	0	7538	9.2250	NaN	S	Svensson	Mr
6	898	3	Connolly, Miss. Kate	female	30.000000	0	0	330972	7.6292	NaN	Q	Connolly	Miss
7	899	2	Caldwell, Mr. Albert Francis	male	26.000000	1	1	248738	29.0000	NaN	S	Caldwell	Mr
8	900	3	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	female	18.000000	0	0	2657	7.2292	NaN	C	Abrahim	Mrs

In [33]:
```
XT = pd.DataFrame([testDF["Sex"], testDF["Age"], testDF["Pclass"]]).T
XT['Sex']= XT['Sex'].astype('category')
XT['Sex'] = XT['Sex'].cat.codes

XT
```

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

Out[33]:

	Sex	Age	Pclass
0	1	34.5	3
1	0	47	3
2	1	62	2
3	1	27	3
4	0	22	3
5	1	14	3
6	0	30	3
7	1	26	2
8	0	18	3
9	1	21	3
10	1	32	3
11	1	46	1

Prediction

In [34]:
```
YT = LDA.predict(XT)
```

In [35]:
```
pass_id = testDF["PassengerId"]
output = pd.DataFrame({"PassengerId":np.array(pass_id),
                       "Survived": YT})
output.to_csv("result.csv", sep =",",index = False)
```

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

In [36]:
```
output
```

Out[36]:

	PassengerId	Survived
0	892	0
1	893	0
2	894	0
3	895	0
4	896	1
5	897	0
6	898	1
7	899	0
8	900	1
9	901	0
10	902	0
11	903	0

In [ ]:

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

```
In [36]: output
```

Out[36]:

	PassengerId	Survived
0	892	0
1	893	0
2	894	0
3	895	0
4	896	1
5	897	0
6	898	1
7	899	0
8	900	1
9	901	0
10	902	0
11	903	0

```
In [ ]:
```

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

Submission and Description	Public Score	Use for Final Score
result.csv 20 hours ago by Bungles Linear Determinants: Sex, Pclass, Age(normalized)	0.78947	☐

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE