Maximum Likelihood Classification

Vashu Agarwal

E21CSEU0054 ¶

Lab 5

In [14]:
```
import numpy as np
import pandas as pd
```

In [15]:
```
# Task 1: Extract the dataset using panda / read the dataset


df=pd.read_csv('/Users/vashuagarwal/Downloads/train.csv')

df
```

Out[15]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0

887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7

891 rows × 12 columns

In [16]:
```
# Task 2: Generate descriptive statistics of df


df.describe()
#print(description)


# O/P is shown below
```

Out[16]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Far
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.00000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.20420
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.69342
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.00000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.91040
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.45420
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.00000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.32920

Simplifying our data

In [17]:
```
def simplify(df: pd.DataFrame):
    del df['PassengerId']
    del df['Name']
    del df['Pclass']
    df['Sex'] = (df['Sex'].values == 'male').astype(int)
    mean_age = np.mean(df['Age'].values[~np.isnan(df['Age'].values)
    df['Age'] = [mean_age if np.isnan(age) else age for age in df['
    del df['Ticket']
    del df['Cabin']
    mean_fare = np.mean(df['Fare'].values[~np.isnan(df['Fare'].valu
    df['Fare'] = [mean_fare if np.isnan(fare) else fare for fare in
    df['S'] = (df['Embarked'].values == 'S').astype(int) + df['Emba
    del df['Embarked']
```

In [18]:
```
labels = df['Survived'].values
del df['Survived']

simplify(df)
```

In [19]:
```
df
```

Out[19]:

	Sex	Age	SibSp	Parch	Fare	S
0	1	22.000000	1	0	7.2500	1
1	0	38.000000	1	0	71.2833	0
2	0	26.000000	0	0	7.9250	1
3	0	35.000000	1	0	53.1000	1
4	1	35.000000	0	0	8.0500	1
...	...	...	...	...	...	...
886	1	27.000000	0	0	13.0000	1
887	0	19.000000	0	0	30.0000	1
888	0	29.699118	1	2	23.4500	1
889	1	26.000000	0	0	30.0000	0
890	1	32.000000	0	0	7.7500	0

891 rows × 6 columns

In [20]:
```
df.describe()
```

Out[20]:

	Sex	Age	SibSp	Parch	Fare	S
count	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000
mean	0.647587	29.699118	0.523008	0.381594	32.204208	0.725028
std	0.477990	13.002015	1.102743	0.806057	49.693429	0.446751
min	0.000000	0.420000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	22.000000	0.000000	0.000000	7.910400	0.000000
50%	1.000000	29.699118	0.000000	0.000000	14.454200	1.000000
75%	1.000000	35.000000	1.000000	0.000000	31.000000	1.000000
max	1.000000	80.000000	8.000000	6.000000	512.329200	1.000000

In [21]:
```
def train_test_split(x: np.ndarray, y: np.ndarray, train_ratio: flo
    '''
    Returns: tuple of form (x_train, y_train, x_test, y_test)
    '''
    n = x.shape[0]
    train_size = int(n * train_ratio)

    train_indices = np.random.choice(n, train_size)
    test_indices = [i for i in np.arange(n) if i not in train_indic

    x_train = np.array([x[i] for i in train_indices])
    y_train = np.array([y[i] for i in train_indices])
    x_test = np.array([x[i] for i in test_indices])
    y_test = np.array([y[i] for i in test_indices])

    return (x_train, y_train, x_test, y_test)
```

Implementing Maximum Likelihood Classification (MLClassifier)

I will use here a maximum likelihood classifier that assumes each observation is a random vector with a Multivariate Gaussian Distributuin:

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \cdot e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

where:

$$\begin{aligned} x &= \text{a column vector with data from one observation} \\ d &= \text{dimension of x (x is a } d \times 1 \text{ vector)} \\ \mu &= \text{mean of x (also } d \times 1) \\ \Sigma &= \text{covariance matrix of x (} d \times d \text{)} \end{aligned}$$

I will make the assumption that each class in our dataset (Survived / Not Survived) has different mean μ and variance Σ.

Training this model will consist mainly in the following:

- first split the dataset into Survived, Not Survived
- compute μ and Σ for each of these two classes

When making a prediction:

- plug input x and the computed μ and Σ into the Gaussian PDF (the formula above) for each class
- output y for the class with the highest value for PDF computed at previous step (y that maximizes the likelihood of our data vector x)

For this method to work the covariance matrix Σ should be **positive definite**. We will check in our code and show a warning if it is not.

In [22]:
```
class MLClassifier:
    def fit(self, x: np.ndarray, y: np.ndarray):
        self.d = x.shape[1] # no. of variables / dimensions
        self.nclasses = len(set(y))

        self.mu_list = []
        self.sigma_list = []

        n = x.shape[0] # no. of observations
        for i in range(self.nclasses):
            cls_x = np.array([x[j] for j in range(n) if y[j] == i])
            mu = np.mean(cls_x, axis=0)
            sigma = np.cov(cls_x, rowvar=False)
            self.mu_list.append(mu)
            self.sigma_list.append(sigma)

    def _class_likelihood(self, x: np.ndarray, cls: int) -> float:
        mu = self.mu_list[cls]
        sigma = self.sigma_list[cls]
        if np.sum(np.linalg.eigvals(sigma) <= 0) != 0:
            print(f'Warning! Covariance matrix for label {cls} is n
            print('The predicted likelihood will be 0.')
            return 0.0
        d = self.d

        #Task 3: Compute function f(x) given in description above

        exp = (-1/2)*np.dot(np.matmul(x-mu, sigma), x-mu)
        s_val = np.linalg.inv(sigma)
        c =  c = 1/np.sqrt(((2*np.pi)**self.d)*np.linalg.det(sigma)




        return c * (np.e**exp)

    def predict(self, x: np.ndarray) -> int:
        likelihoods = [self._class_likelihood(x, i) for i in range(
        return np.argmax(likelihoods)

    def score(self, x: np.ndarray, y: np.ndarray):
        n = x.shape[0]
        predicted_y = np.array([self.predict(x[i]) for i in range(n
        n_correct = np.sum(predicted_y == y)
        return n_correct/n
```

In [23]:
```
(x_train, y_train, x_test, y_test) = train_test_split(df.values, la
```

In [24]:
```
# Task 4: Call Maximum Likelihood classifier function

mlc = MLClassifier()



#Task 5: fit the Maximum Likelihood classifier for train test data

mlc.fit(x_train,y_train)
```

In [25]:
```
score = mlc.score(x_test,y_test)# pass first parameter, # pass seco

print(score)

# o/ p is shown below
```

```
0.6414634146341464
```