# Discriminant Analysis

* Concerned with separating distinct sets of objects (observations) and allocating new objects (observations) to previously defined groups.

# Classification for two population

## Separating two classes of objects

* Label the two classes Π1 and Π2

## Objects are classified on the basis of measurements on *p* associated variables

**X** = [*X*1, *X*2, … *X*p]

|  |  |
| --- | --- |
| Populations 1 and 2 | Measured variables, **X** |
| Solvent and Insolvent insurance company  Federalist papers written by James Madison and those written by Alexander Hamilton  Purchasers of new products and  laggards  Successful and unsuccessful students  Good and poor credit risks | Total assets, cost of stocks and bonds, market value of stocks and bonds, loss expenses, surplus, amount of premium.  Frequencies of different words and lengths of sentences  Education, income, family size, amount  of previous brand switching  Entrance examination scores, grade point average in school examination, number of school activities.  Income, age, number of credit cards, family size, occupation. |

25



20

LOTSIZE

15

10

30 40 50 60 70 80 90 100 110 120

INCOME

OWNERSHIP

1

2

Classification principles

1. A good classification procedure should result in few misclassifications
   * Probabilities of misclassification should be small
   * For unequal population size, one has a greater likelihood of occurrence for larger populations; Include the concept of *prior probability*: let *p*1 and *p*2 be the prior probabilities of Π1 and Π2 respectively.

*p*1 + *p*2 = 1

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Classify as | |
| 1 | 2 |
| True  population | 1 |  | P(21) |
| 2 | P(12) |  |

P(misclassified as 1) = P( Observation comes from 2 and is misclassified as 1)

= P(12) *p*2

P(misclassified as 2) = P( Observation comes from 1 and is misclassified as 2)

= P(21) *p*1

1. Another aspect of classification is cost
   * Classifying a Π1 object as belonging to Π2 represents a more serious

error than classifying a Π2 object as belonging to Π1

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Classify as | |
| 1 | 2 |
| True  population | 1 |  | C(21) |
| 2 | C(12) |  |

**A reasonable classification rule should have an Expected Cost of Misclassification (ECM) as small as possible.**

**ECM = C(2|1).P(2|1).*p*1 + C(1|2).P(1|2).*p*2**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Respond ent Number | Resort  visit | Annual family income (000s) | Attitude towads travel | Importance attached to family skiing holiday | Househol  d size | Age of head of household | Amount spent on family skiing |
| 1 | 1 | 50.2 | 5 | 8 | 3 | 43 | 2 |
| 2 | 1 | 70.3 | 6 | 7 | 4 | 61 | 3 |
| 3 | 1 | 62.9 | 7 | 5 | 6 | 52 | 3 |
| 4 | 1 | 48.5 | 7 | 5 | 5 | 36 | 1 |
| 5 | 1 | 52.7 | 6 | 6 | 4 | 55 | 3 |
| 6 | 1 | 75 | 8 | 7 | 5 | 68 | 3 |
| 7 | 1 | 46.2 | 5 | 3 | 3 | 62 | 2 |
| 8 | 1 | 57 | 2 | 4 | 6 | 51 | 2 |
| 9 | 1 | 64.1 | 7 | 5 | 4 | 57 | 3 |
| 10 | 1 | 68.1 | 7 | 6 | 5 | 45 | 3 |
| 11 | 1 | 73.4 | 6 | 7 | 5 | 44 | 3 |
| 12 | 1 | 71.9 | 5 | 8 | 4 | 64 | 3 |
| 13 | 1 | 56.2 | 1 | 8 | 6 | 54 | 2 |
| 14 | 1 | 49.3 | 4 | 2 | 3 | 56 | 3 |
| 15 | 1 | 62 | 5 | 6 | 2 | 58 | 3 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Respond ent Number | Resort  visit | Annual family income (000s) | Attitude towads travel | Importance attached to family skiing holiday | Househol  d size | Age of head of household | Amount spent on family skiing |
| 16 | 2 | 32.1 | 5 | 4 | 3 | 58 | 1 |
| 17 | 2 | 36.2 | 4 | 3 | 2 | 55 | 1 |
| 18 | 2 | 43.2 | 2 | 5 | 2 | 57 | 2 |
| 19 | 2 | 50.4 | 5 | 2 | 4 | 37 | 2 |
| 20 | 2 | 44.1 | 6 | 6 | 3 | 42 | 2 |
| 21 | 2 | 38.3 | 6 | 6 | 2 | 45 | 1 |
| 22 | 2 | 55 | 1 | 2 | 2 | 57 | 2 |
| 23 | 2 | 46.1 | 3 | 5 | 3 | 51 | 1 |
| 24 | 2 | 35 | 6 | 4 | 5 | 64 | 1 |
| 25 | 2 | 37.3 | 2 | 7 | 4 | 54 | 1 |
| 26 | 2 | 41.8 | 5 | 1 | 3 | 56 | 2 |
| 27 | 2 | 57 | 8 | 3 | 2 | 36 | 2 |
| 28 | 2 | 33.4 | 6 | 8 | 2 | 50 | 1 |
| 29 | 2 | 37.5 | 3 | 2 | 3 | 48 | 1 |
| 30 | 2 | 41.3 | 3 | 3 | 2 | 42 | 1 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Responde nt Number | Annual family income (000s) | Attitude towads travel | Importance attached to family skiing holiday | Household size | Age of head of household | Amount spent on family skiing |
| 31 | 50.8 | 4 | 7 | 3 | 45 | 2 |
| 32 | 49.6 | 5 | 3 | 5 | 39 | 1 |
| 33 | 54.5 | 7 | 3 | 3 | 37 | 2 |
| 34 | 45 | 5 | 4 | 3 | 60 | 2 |
| 35 | 68 | 6 | 6 | 6 | 46 | 3 |
| 36 | 62.1 | 5 | 6 | 3 | 56 | 3 |
| 37 | 35 | 4 | 3 | 4 | 54 | 1 |
| 38 | 54 | 6 | 7 | 4 | 58 | 2 |
| 39 | 39.4 | 6 | 5 | 3 | 44 | 3 |
| 40 | 37 | 2 | 6 | 5 | 51 | 1 |

Devise a Discriminant Rule and based on the rule find whether the respondent Number 31-40 will visit the resort second time or not?

import pandas as pd

df = pd.read\_csv("E:/MY DOCUMENTS/Desktop/Python/DAdata.csv") # Dropping unnecessary columns

df.drop(['RespNo'], axis = 1, inplace=True) # Dropping missing values rows df.dropna(inplace=True)

from sklearn.discriminant\_analysis import LinearDiscriminantAnalysis clf = LinearDiscriminantAnalysis()

X = df.iloc[:,1:].copy() visit = df['visit'].copy()

result = clf.fit(X, visit)

print("\n Classes: ", result.classes\_)

print("\n GROUP MEANS: \n", result.means\_.round(4)) print("\n PRIOR PROBABILITY", result.priors\_.round(4)) print("\n PREDICTIONS: ", result.predict(X))

print("\n CANONICAL DISCRIMINANT FUNCTION \n",

result.scalings\_.round(3))

from sklearn.metrics import confusion\_matrix results=confusion\_matrix(visit, result.predict(X)) print("\nThe Confusion Matrix is: \n", results)

#determining the accuracy of the model score=result.score(X, visit)

print("\n The accuracy is: ", score.round(4))