

The data set (and description) can be downloaded here:

<http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data>

Description:

1. Title: Iris Plants Database

Updated Sept 21 by C.Blake - Added discrepancy information

2. Sources:

(a) Creator: R.A. Fisher

(b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)

(c) Date: July, 1988

3. Past Usage:

- Publications: too many to mention!!! Here are a few.

1. Fisher, R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950).

2. Duda, R.O., & Hart, P.E. (1973) Pattern Classification and Scene Analysis. (Q327.D83) John Wiley & Sons. ISBN 0-471-22361-1. See page 218.

3. Dasarthy, B.V. (1980) "Nosing Around the Neighborhood: A New System Structure and Classification Rule for Recognition in Partially Exposed Environments". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No. 1, 67-71.

-- Results:

-- very low misclassification rates (0% for the setosa class)

4. Gates, G.W. (1972) "The Reduced Nearest Neighbor Rule". IEEE Transactions on Information Theory, May 1972, 431-433.

-- Results:

-- very low misclassification rates again

5. See also: 1988 MLC Proceedings, 54-64. Cheeseman et al's AUTOCLASS II conceptual clustering system finds 3 classes in the data.

4. Relevant Information:

--- This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

--- Predicted attribute: class of iris plant.

--- This is an exceedingly simple domain.

--- This data differs from the data presented in Fishers article (identified by Steve Chadwick, spchadwick@speedaz.net)

The 35th sample should be: 4.9,3.1,1.5,0.2,"Iris-setosa" where the error is in the fourth feature.

The 38th sample: 4.9,3.6,1.4,0.1,"Iris-setosa"

where the errors are in the second and third features.

5. Number of Instances: 150 (50 in each of three classes)

6. Number of Attributes: 4 numeric, predictive attributes and the class

7. Attribute Information:

1. sepal length in cm

2. sepal width in cm

3. petal length in cm

4. petal width in cm

5. class:

-- Iris Setosa

-- Iris Versicolour

-- Iris Virginica

8. Missing Attribute Values: None

Summary Statistics:

	Min	Max	Mean	SD	Class Correlation
sepal length:	4.3	7.9	5.84	0.83	0.7826
sepal width:	2.0	4.4	3.05	0.43	-0.4194
petal length:	1.0	6.9	3.76	1.76	0.9490 (high!)
petal width:	0.1	2.5	1.20	0.76	0.9565 (high!)

9. Class Distribution: 33.3% for each of 3 classes.

Citation Request:

Please refer to the repository <http://archive.ics.uci.edu/ml> (see citation policy).

See also Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>].

Irvine, CA: University of California, School of Information and Computer Science.

Descriptive statistics:

Dataset= iris_setosavsversicolor : n= 100 , d= 4

Class1: n= 50

Covariance matrix:

	[,1]	[,2]	[,3]	[,4]
[1,]	0.1242	0.1003	0.0161	0.0105
[2,]	0.1003	0.1452	0.0117	0.0114
[3,]	0.0161	0.0117	0.0301	0.0057
[4,]	0.0105	0.0114	0.0057	0.0115

Correlation matrix:

	[,1]	[,2]	[,3]	[,4]
[1,]	1.0000	0.7468	0.2639	0.2791
[2,]	0.7468	1.0000	0.1767	0.2800
[3,]	0.2639	0.1767	1.0000	0.3063
[4,]	0.2791	0.2800	0.3063	1.0000

Median: 5 3.4 1.5 0.2

Mean: 5.006 3.418 1.464 0.244

MCD-estimated:

MDC-0.975-Mean: 4.975 3.3893 1.4429 0.2

MDC-0.750-Mean: 4.975 3.3893 1.4429 0.2

MDC-0.500-Mean: 4.975 3.3893 1.4429 0.2

Class2: n= 50

Covariance matrix:

	[,1]	[,2]	[,3]	[,4]
[1,]	0.2664	0.0852	0.1829	0.0558
[2,]	0.0852	0.0985	0.0827	0.0412
[3,]	0.1829	0.0827	0.2208	0.0731
[4,]	0.0558	0.0412	0.0731	0.0391

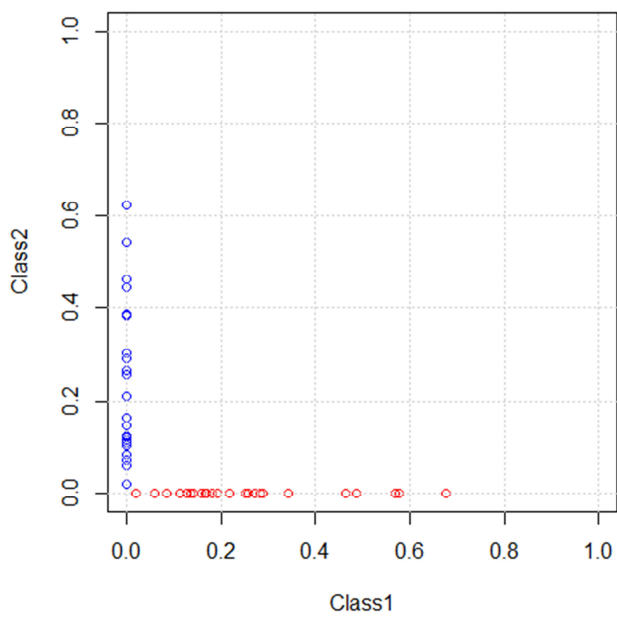
Correlation matrix:

	[,1]	[,2]	[,3]	[,4]
[1,]	1.0000	0.5259	0.7540	0.5465
[2,]	0.5259	1.0000	0.5605	0.6640
[3,]	0.7540	0.5605	1.0000	0.7867
[4,]	0.5465	0.6640	0.7867	1.0000

Median: 5.9113 2.7996 4.2731 1.3255
 Mean: 5.936 2.77 4.26 1.326
 MCD-estimated:
 MDC-0.975-Mean: 5.9205 2.8154 4.2026 1.3051
 MDC-0.750-Mean: 5.9146 2.8098 4.2268 1.3073
 MDC-0.500-Mean: 5.9146 2.8098 4.2268 1.3073

Measures:
 Mah.Dist: 10.1937
 Mah.Dist-MCD-0.975: 9.6235
 Mah.Dist-MCD-0.750: 9.9827
 Mah.Dist-MCD-0.500: 9.6777

DD-Plot (zonoid): iris_setosavsversicolor



DD-Plot (random Tukey): iris_setosavsversicolor

