**Colon cancer data set of Alon et al. 1999**

**Description**

Response is y=1 or y=-1 according as tissue is normal or tumor There are 2000 gene expressions.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | y=1 | y=–1 | Total |
| train | 14 | 26 | 40 |
| test | 8 | 14 | 22 |

**Usage**

data(Alon)

**Format**

List with 4 named elements, X, y, Xt, yt, which are respectively the training design matrix, training classes, test design matrix and test classes.

**Details**

Colon adenocarcinoma tissues were collected from patients and from some of these patients, paired normal colon tissue also was obtained. Gene expression in 40 tumor and 22 normal colon tissue samples was analyzed with an Affymetrix oligonucleotide array complementary to more than 6500 human genes. The data set contains the expression of the 2000 genes with highest minimal intensity across the 62 tissues. Each gene intensity has been derived from the about 20 feature pairs that correspond to the gene on the chip by using a filtering process. The data is otherwise unprocessed, i.e. no normalization has been performed yet. The training set consists of 40 colon tissues of which 14 are normal and 26 tumor samples. The test set consists of 22 tissues of which 8 are normal and 14 tumor samples. The number of gene expression levels is 2000. The goal here is to classify the tissues as being cancerous or noncancerous.

**Source**

Nathalie Pochet, Frank De Smet, Johan A.K. Suykens and Bart L.R. De Moor (2004). Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. Bioinformatics Advance Access published July 1, 2004. <http://homes.esat.kuleuven.be/~npochet/Bioinformatics/>

**References**

Alon,A., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D., and Levine,A.J. (1999) Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays, Proc. Natl. Acad. Sci. USA, 96,6745-6750.

**Examples**

data(Alon)

[Package *CLEg* version 2.0 [Index]](http://www.stats.uwo.ca/faculty/aim/2015/9850/microarrays/FitMArray/chm/00Index.html)

|  |  |
| --- | --- |
| Singh(CLEg) | R Documentation |

**Breast cancer BRCA1 mutation Hendenfalk et al. 2004**

**Description**

Response is y=1 or y=-1 according as have BRAC1 mutation or not. There are 3226 gene expressions.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | y=1 | y=–1 | Total |
| train | 4 | 10 | 14 |
| test | 3 | 5 | 8 |

Note: the gene expressions in BRAC1, BRAC2 and Sporadic are not identical.

**Usage**

data(BRAC1)

**Format**

List with 4 named elements, X, y, Xt, yt, which are respectively the training design matrix, training classes, test design matrix and test classes.

**Details**

Breast cancer data set (Hedenfalk et al., 2001). RNA from samples of primary breast tumors from 7 carriers of the BRCA1 mutation, 8 carriers of the BRCA2 mutation, and 7 patients with sporadic cases of breast cancer have been hybridized to a cDNA microarray containing 6512 complementary DNA clones of 5361 genes. The goal here is to classify the different mutations, so three combinations are possible in this case. First, tissues with BRCA1 mutations are separated from the tissues with BRCA2 or sporadic mutations. The training set consists of 14 breast cancer tissues of which 4 have a BRCA1 mutation and 10 have not. The test set consists of 8 tissues of which 3 have a BRCA1 mutation and 5 have not. The number of gene expression levels is 3226. Second, tissues with BRCA2 mutations are separated from the tissues with BRCA1 or sporadic mutations. The training set consists of 14 breast cancer tissues of which 5 have a BRCA1 mutation and 9 have not. The test set consists of 8 tissues of which 3 have a BRCA1 mutation and 5 have not. The number of gene expression levels is 3226. Third, tissues with sporadic mutations are separated from the tissues with BRCA1 or BRCA2 mutations. The training set consists of 14 breast cancer tissues of which 4 have a BRCA1 mutation and 10 have not. The test set consists of 8 tissues of which 3 have a BRCA1 mutation and 5 have not. The number of gene expression levels is 3226.

**Source**

Nathalie Pochet, Frank De Smet, Johan A.K. Suykens and Bart L.R. De Moor (2004). Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. Bioinformatics Advance Access published July 1, 2004. <http://homes.esat.kuleuven.be/~npochet/Bioinformatics/>

**References**

Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner, M., Simon,R., Meltzer,P., Gusterson,B., Esteller,M., Raffeld,M., Yakhini,Z., Ben-Dor,A., Dougherty,E., Kononen,J., Bubendorf,L., Fehrle,W., Pittaluga,S., Gruvberger,S., Loman,N., Johannsson,O., Olsson,H., Wilfond,B., Sauter,G., Kallioniemi,O.-P., Borg,A. and Trent,J. (2001) Gene-Expression Profiles in Hereditary Breast Cancer,

**Examples**

data(BRAC1)

[Package *CLEg* version 2.0 [Index]](http://www.stats.uwo.ca/faculty/aim/2015/9850/microarrays/FitMArray/chm/00Index.html)

**Breast cancer BRCA2 mutation Hendenfalk et al. 2004**

**Description**

Response is y=1 or y=-1 according as have BRAC2 mutation or not. There are 3226 gene expressions.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | y=1 | y=–1 | Total |
| train | 5 | 9 | 14 |
| test | 3 | 5 | 8 |

Note: the gene expressions in BRAC1, BRAC2 and Sporadic are not identical.

**Usage**

data(BRAC2)

**Format**

List with 4 named elements, X, y, Xt, yt, which are respectively the training design matrix, training classes, test design matrix and test classes.

**Details**

Breast cancer data set (Hedenfalk et al., 2001). RNA from samples of primary breast tumors from 7 carriers of the BRCA1 mutation, 8 carriers of the BRCA2 mutation, and 7 patients with sporadic cases of breast cancer have been hybridized to a cDNA microarray containing 6512 complementary DNA clones of 5361 genes. The goal here is to classify the different mutations, so three combinations are possible in this case. First, tissues with BRCA1 mutations are separated from the tissues with BRCA2 or sporadic mutations. The training set consists of 14 breast cancer tissues of which 4 have a BRCA1 mutation and 10 have not. The test set consists of 8 tissues of which 3 have a BRCA1 mutation and 5 have not. The number of gene expression levels is 3226. Second, tissues with BRCA2 mutations are separated from the tissues with BRCA1 or sporadic mutations. The training set consists of 14 breast cancer tissues of which 5 have a BRCA1 mutation and 9 have not. The test set consists of 8 tissues of which 3 have a BRCA1 mutation and 5 have not. The number of gene expression levels is 3226. Third, tissues with sporadic mutations are separated from the tissues with BRCA1 or BRCA2 mutations. The training set consists of 14 breast cancer tissues of which 4 have a BRCA1 mutation and 10 have not. The test set consists of 8 tissues of which 3 have a BRCA1 mutation and 5 have not. The number of gene expression levels is 3226.

**Source**

Nathalie Pochet, Frank De Smet, Johan A.K. Suykens and Bart L.R. De Moor (2004). Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. Bioinformatics Advance Access published July 1, 2004. <http://homes.esat.kuleuven.be/~npochet/Bioinformatics/>

**References**

Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner, M., Simon,R., Meltzer,P., Gusterson,B., Esteller,M., Raffeld,M., Yakhini,Z., Ben-Dor,A., Dougherty,E., Kononen,J., Bubendorf,L., Fehrle,W., Pittaluga,S., Gruvberger,S., Loman,N., Johannsson,O., Olsson,H., Wilfond,B., Sauter,G., Kallioniemi,O.-P., Borg,A. and Trent,J. (2001) Gene-Expression Profiles in Hereditary Breast Cancer,

**Examples**

data(BRAC2)

[Package *CLEg* version 2.0 [Index]](http://www.stats.uwo.ca/faculty/aim/2015/9850/microarrays/FitMArray/chm/00Index.html)

**Gaussian data with two classes**

**Description**

Two predictor variables generated from a bivariate normal distribution. The covariance matrix is different for each class. In the training sample, there is a separating hyperplane.

**Usage**

data(Gaussian)

**Format**

The format is: List of 4 $ X : num [1:200, 1:2] 1.142 0.589 0.946 1.327 1.052 ... $ y : num [1:200] 1 1 1 1 1 1 1 1 1 1 ... $ Xt: num [1:200, 1:2] 1.20 -0.63 1.28 1.02 1.17 ... $ yt: num [1:200] 1 1 1 1 1 1 1 1 1 1 ...

**Details**

**Examples**

data(Gaussian)

[Package *CLEg* version 1.0 [Index]](http://www.stats.uwo.ca/faculty/aim/2015/9850/microarrays/FitMArray/chm/00Index.html)

**Iizuka carcinoma data set**

**Description**

Response is y=1 or y=-1 according as early intrahepatic recurrence or not There are 7129 gene expressions.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | y=1 | y=–1 | Total |
| train | 12 | 21 | 33 |
| test | 8 | 19 | 27 |

**Usage**

data(Iizuka)

**Format**

List with 4 named elements, X, y, Xt, yt, which are respectively the training design matrix, training classes, test design matrix and test classes.

**Details**

Hepatocellular carcinoma data set (Iizuka et al., 2003). mRNA expression profiles in tissue specimens from a training set comprising 33 patients with hepatocellular carcinoma have been hybridized with high-density oligonucleotide microarrays representing about 6000 genes. The same has been done for a blinded set of samples from 27 newly enrolled patients. Since hepatocellular carcinoma has a poor prognosis because of the high intrahepatic recurrence rate, the goal here is to predict early intrahepatic recurrence or non-recurrence. The training set consists of 33 hepatocellular carcinoma tissues of which 12 suffer from early intrahepatic recurrence and 21 not. The test set consists of 27 hepatocellular carcinoma tissues of which 8 suffer from early intrahepatic recurrence and 19 not. The number of gene expression levels is 7129.

**Source**

Nathalie Pochet, Frank De Smet, Johan A.K. Suykens and Bart L.R. De Moor (2004). Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. Bioinformatics Advance Access published July 1, 2004. <http://homes.esat.kuleuven.be/~npochet/Bioinformatics/>

**References**

Iizuka,N., Oka,M., Yamada-Okabe,H., Nishida,M., Maeda,Y., Mori,N., Takao,T., Tamesa,T., Tangoku,A., Tabuchi,H., Hamada,K., Nakayama,H., Ishitsuka,H., Miyamoto,T., Hirabayashi,A., Uchimura,S. and Hamamoto,Y. (2003) Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection, The Lancet, 361,923-929.

**Examples**

data(Iizuka)

[Package *CLEg* version 2.0 [Index]](http://www.stats.uwo.ca/faculty/aim/2015/9850/microarrays/FitMArray/chm/00Index.html)

**HTF Mixture Classifcation Example**

**Description**

HTF example. Data generated for each class by a mixture of 10 different normal distribution. The training sample is as given by HTF but the test sample was generated independently.

The Bayes error rate for this model is about 21

**Usage**

data(Mixture)

**Format**

The format is: List of 4 $ X :'data.frame': 200 obs. of 2 variables: ..$ V1: num [1:200] 2.5261 0.3670 0.7682 0.6934 -0.0198 ... ..$ V2: num [1:200] 0.3211 0.0315 0.7175 0.7772 0.8673 ... $ y : num [1:200] 1 1 1 1 1 1 1 1 1 1 ... $ Xt:'data.frame': 1000 obs. of 2 variables: ..$ V1: num [1:1000] -0.584 1.604 0.475 1.264 2.362 ... ..$ V2: num [1:1000] -0.446 0.617 1.376 0.159 0.892 ... $ yt: num [1:1000] 1 1 1 1 1 1 1 1 1 1 ...

**Details**

**Source**

HTF

**References**

**Examples**

data(Mixture)

[Package *CLEg* version 1.0 [Index]](http://www.stats.uwo.ca/faculty/aim/2015/9850/microarrays/FitMArray/chm/00Index.html)

**High-grade glioma dataset Nutt et al. 2003**

**Description**

Response is y=1 or y=-1 according as glioblastomas or anaplastic oligodendrogliomas. There are 12625 gene expressions.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | y=1 | y=–1 | Total |
| train | 14 | 7 | 21 |
| test | 14 | 15 | 29 |

Note: the gene expressions in BRAC1, BRAC2 and Sporadic are not identical.

**Usage**

data(Nutt)

**Format**

List with 4 named elements, X, y, Xt, yt, which are respectively the training design matrix, training classes, test design matrix and test classes.

**Details**

50 high-grade glioma samples were carefully selected, 28 glioblastomas and 22 anaplastic oligodendrogliomas, all were primary tumors sampled before therapy. The classic subset of tumors were cases diagnosed similarly by all examining pathologists, and each case resembled typical depictions in standard textbooks. A total of 21 classic tumors was selected, and the remaining 29 samples were considered nonclassic tumors, lesions for which diagnosis might be controversial. Affymetrix arrays are used to determine the expression of about 12000 genes. The goal here is to separate the glioblastomas from the anaplastic oligodendrogliomas, which allows appropriate therapeutic decisions and prognostic estimation. The training set consists of 21 gliomas with classic histology of which 14 are glioblastomas and 7 anaplastic oligodendrogliomas. The test set consists of 29 gliomas with non-classic histology of which 14 are glioblastomas and 15 are anaplastic oligodendrogliomas. The number of gene expression levels is 12625.

**Source**

Nathalie Pochet, Frank De Smet, Johan A.K. Suykens and Bart L.R. De Moor (2004). Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. Bioinformatics Advance Access published July 1, 2004. <http://homes.esat.kuleuven.be/~npochet/Bioinformatics/>

**References**

Nutt,C.L., Mani,D.R., Betensky,R.A., Tamayo,P., Cairncross, J.G., Ladd,C., Pohl,U., Hartmann,C., McLaughlin, M.E., Batchelor,T.T., Black,P.M., von Deimling,A., Pomeroy,S.L., Golub,T.R. and Louis,D.N. (2003) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification, Cancer Research, 63(7),1602-1607.

**Examples**

data(Nutt)

[Package *CLEg* version 2.0 [Index]](http://www.stats.uwo.ca/faculty/aim/2015/9850/microarrays/FitMArray/chm/00Index.html)

**Smoker dataset Spira et al. 2004**

**Description**

Number of genes: 9968. Other covariates including age and gender are available in the original data. The training set consists of 38 subjects and the test data 37.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | y=1 | y=–1 | Total |
| train | 26 | 12 | 38 |
| test | 26 | 11 | 37 |

The response y=1 corresonds to a smoker or previous smoker while y=-1 corresponds to an individual who never smoked.

The original data was not divided into training and test. There are 27 males in the training and test datasets. There are 10 females in the training dataset and 9 in the test.

**Usage**

data(Spira)

**Format**

List with 4 named elements, X, y, Xt, yt, which are respectively the training design matrix, training classes, test design matrix and test classes.

**Details**

**Source**

[www.pnas.org/cgi/doi/10.1073/pnas.0401422101](http://www.stats.uwo.ca/faculty/aim/2015/9850/microarrays/FitMArray/chm/www.pnas.org/cgi/doi/10.1073/pnas.0401422101)

**References**

Avrum Spira, Jennifer Beanea, Vishal Shah, Gang Liu, Frank Schembri, Xuemei Yang, John Palma, and Jerome S. Brody (2004). Effects of cigarette smoke on the human airway epithelial cell transcriptome. PNAS July 6, 2004 vol. 101 no. 27

**Examples**

data(Spira)

[Package *CLEg* version 2.1 [Index]](http://www.stats.uwo.ca/faculty/aim/2015/9850/microarrays/FitMArray/chm/00Index.html)

**Breast cancer Sporadic mutation Hendenfalk et al. 2004**

**Description**

Response is y=1 or y=-1 according as have BRAC2 mutation or not. There are 3226 gene expressions.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | y=1 | y=–1 | Total |
| train | 4 | 10 | 14 |
| test | 3 | 5 | 8 |

Note: the gene expressions in BRAC1, BRAC2 and Sporadic are not identical.

**Usage**

data(BRAC2)

**Format**

List with 4 named elements, X, y, Xt, yt, which are respectively the training design matrix, training classes, test design matrix and test classes.

**Details**

Breast cancer data set (Hedenfalk et al., 2001). RNA from samples of primary breast tumors from 7 carriers of the BRCA1 mutation, 8 carriers of the BRCA2 mutation, and 7 patients with sporadic cases of breast cancer have been hybridized to a cDNA microarray containing 6512 complementary DNA clones of 5361 genes. The goal here is to classify the different mutations, so three combinations are possible in this case. First, tissues with BRCA1 mutations are separated from the tissues with BRCA2 or sporadic mutations. The training set consists of 14 breast cancer tissues of which 4 have a BRCA1 mutation and 10 have not. The test set consists of 8 tissues of which 3 have a BRCA1 mutation and 5 have not. The number of gene expression levels is 3226. Second, tissues with BRCA2 mutations are separated from the tissues with BRCA1 or sporadic mutations. The training set consists of 14 breast cancer tissues of which 5 have a BRCA1 mutation and 9 have not. The test set consists of 8 tissues of which 3 have a BRCA1 mutation and 5 have not. The number of gene expression levels is 3226. Third, tissues with sporadic mutations are separated from the tissues with BRCA1 or BRCA2 mutations. The training set consists of 14 breast cancer tissues of which 4 have a BRCA1 mutation and 10 have not. The test set consists of 8 tissues of which 3 have a BRCA1 mutation and 5 have not. The number of gene expression levels is 3226.

**Source**

Nathalie Pochet, Frank De Smet, Johan A.K. Suykens and Bart L.R. De Moor (2004). Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. Bioinformatics Advance Access published July 1, 2004. <http://homes.esat.kuleuven.be/~npochet/Bioinformatics/>

**References**

Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner, M., Simon,R., Meltzer,P., Gusterson,B., Esteller,M., Raffeld,M., Yakhini,Z., Ben-Dor,A., Dougherty,E., Kononen,J., Bubendorf,L., Fehrle,W., Pittaluga,S., Gruvberger,S., Loman,N., Johannsson,O., Olsson,H., Wilfond,B., Sauter,G., Kallioniemi,O.-P., Borg,A. and Trent,J. (2001) Gene-Expression Profiles in Hereditary Breast Cancer,

**Examples**

data(Sporadic)

[Package *CLEg* version 2.0 [Index]](http://www.stats.uwo.ca/faculty/aim/2015/9850/microarrays/FitMArray/chm/00Index.html)

**Prostate cancer data set Singh et al. 2002**

**Description**

Response is y=1 or y=-1 according as tissue is normal/tumor. There are 12600 gene expressions.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | y=1 | y=–1 | Total |
| train | 50 | 52 | 102 |
| test | 9 | 25 | 34 |

**Usage**

data(Singh)

**Format**

List with 4 named elements, X, y, Xt, yt, which are respectively the training design matrix, training classes, test design matrix and test classes.

**Details**

High-quality expression profiles were successfully derived from 52 prostate tumors and 50 nontumor prostate samples from patients undergoing surgery. Oligonucleotide microarrays containing probes for approximately 12600 genes and ESTs. Since prostate tumors are among the most heterogeneous of cancers, both histologically and clinically, the goal here is to classify tumor and nontumour samples. The training set consists of 102 prostate tissues of which 50 are normal and 52 tumor samples. The test set consists of 34 tissues of which 9 are normal and 25 tumor samples. The number of gene expression levels is 12600.

**Source**

Nathalie Pochet, Frank De Smet, Johan A.K. Suykens and Bart L.R. De Moor (2004). Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. Bioinformatics Advance Access published July 1, 2004. <http://homes.esat.kuleuven.be/~npochet/Bioinformatics/>

**References**

Singh,D., Febbo,P.G., Ross,K., Jackson,D.G., Manola,J., Ladd,C., Tamayo,P., Renshaw,A.A., D'Amico,A.V., Richie,J.P., Lander,E.S., Loda,M., Kantoff,P.W., Golub,T.R. and Sellers,W.R. (2002) Gene expression correlates of clinical prostate cancer behavior, Cancer Cell, 1(2),203-209.

**Examples**

data(Singh)

[Package *CLEg* version 2.0 [Index]](http://www.stats.uwo.ca/faculty/aim/2015/9850/microarrays/FitMArray/chm/00Index.html)

|  |  |
| --- | --- |
| Veer(CLEg) | R Documentation |

**Breast cancer data set Van't Veer et al. 2002**

**Description**

Response is y=1 or y=-1 according as develop metastases or disease free. There are 24188 gene expressions.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | y=1 | y=–1 | Total |
| train | 34 | 44 | 78 |
| test | 3 | 5 | 8 |

Note: the gene expressions in BRAC1, BRAC2 and Sporadic are not identical.

**Usage**

data(Veer)

**Format**

List with 4 named elements, X, y, Xt, yt, which are respectively the training design matrix, training classes, test design matrix and test classes.

**Details**

78 primary breast cancers (34 from patients who developed distant metastases within 5 years and 44 from patients who continue to be disease-free after a period of at least 5 years) have been selected from patients who were lymph node negative and under 55 years of age at diagnosis. Two hybridizations were carried out for each tumour using a fluorescent dye reversal technique on microarrays containing approximately 25000 human genes synthesized by an inkjet oligonucleotide technology. The goal here is to predict the presence of subclinical metastases in order to provide a strategy to select patients who would benefit from adjuvant therapy. The training set consists of 78 breast cancer patients of which 34 develop metastases within 5 years and 44 remain disease-free within 5 years. The test set consists of 19 patients of which 12 develop metastases within 5 years and 7 remain disease-free within 5 years. The number of gene expression levels is 24188. This data set contained some missing values. Gene expression levels lacking for all patients are left out. The rest of the missing values is estimated based on the correlations between the gene expressions.

**Source**

Nathalie Pochet, Frank De Smet, Johan A.K. Suykens and Bart L.R. De Moor (2004). Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. Bioinformatics Advance Access published July 1, 2004. <http://homes.esat.kuleuven.be/~npochet/Bioinformatics/>

**References**

Van't Veer,L.J., Dai,H., Van De Vijver,M.J., He,Y.D., Hart,A.A.M., Mao,M., Peterse,H.L., Van Der Kooy,K., Marton, M.J., Witteveen,A.T., Schreiber,G.J., Kerkhoven,R.M., Roberts,C., Linsley,P.S., Bernards,R. and Friend,S.H. (2002) Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer, Nature, 415,530-536.

**Examples**

data(Veer)

[Package *CLEg* version 2.0 [Index]](http://www.stats.uwo.ca/faculty/aim/2015/9850/microarrays/FitMArray/chm/00Index.html)