

```
In [7]: import numpy as np
        from sklearn import datasets, naive_bayes
        from sklearn.model_selection import cross_val_score
```

```
In [33]: breast_cancer = datasets.load_breast_cancer()
print breast_cancer.DESCR
```

```
Breast Cancer Wisconsin (Diagnostic) Database
=====
```

```
Notes
```

```
-----
```

```
Data Set Characteristics:
```

```
  :Number of Instances: 569
```

```
  :Number of Attributes: 30 numeric, predictive attributes and the class
```

```
  :Attribute Information:
```

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter² / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

```
  - class:
```

- WDBC-Malignant
- WDBC-Benign

```
  :Summary Statistics:
```

```
=====
```

	Min	Max
radius (mean):	6.981	28.11
texture (mean):	9.71	39.28

```
=====
```

perimeter (mean):	43.79	188.5
area (mean):	143.5	2501.0
smoothness (mean):	0.053	0.163
compactness (mean):	0.019	0.345
concavity (mean):	0.0	0.427
concave points (mean):	0.0	0.201
symmetry (mean):	0.106	0.304
fractal dimension (mean):	0.05	0.097
radius (standard error):	0.112	2.873
texture (standard error):	0.36	4.885
perimeter (standard error):	0.757	21.98
area (standard error):	6.802	542.2
smoothness (standard error):	0.002	0.031
compactness (standard error):	0.002	0.135
concavity (standard error):	0.0	0.396
concave points (standard error):	0.0	0.053
symmetry (standard error):	0.008	0.079
fractal dimension (standard error):	0.001	0.03
radius (worst):	7.93	36.04
texture (worst):	12.02	49.54
perimeter (worst):	50.41	251.2
area (worst):	185.2	4254.0
smoothness (worst):	0.071	0.223
compactness (worst):	0.027	1.058
concavity (worst):	0.0	1.252
concave points (worst):	0.0	0.291
symmetry (worst):	0.156	0.664
fractal dimension (worst):	0.055	0.208
=====	=====	=====

:Missing Attribute Values: None

:Class Distribution: 212 - Malignant, 357 - Benign

:Creator: Dr. William H. Wolberg, W. Nick Street, Olvi L. Mangasarian

:Donor: Nick Street

:Date: November, 1995

This is a copy of UCI ML Breast Cancer Wisconsin (Diagnostic) datasets.
<https://goo.gl/U2Uwz2> (<https://goo.gl/U2Uwz2>)

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.

The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in:
[K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

This database is also available through the UW CS ftp server:

```
ftp ftp.cs.wisc.edu
cd math-prog/cpo-dataset/machine-learn/WDBC/
```

References

- W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.
- O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pages 570-577, July-August 1995.
- W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters 77 (1994) 163-171.

```
In [40]: print 'BernoulliNB classifier:', \
          cross_val_score(naive_bayes.BernoulliNB(), breast_cancer.data,
                          breast_cancer.target).mean()
print 'MultinomialNB classifier:', \
      cross_val_score(naive_bayes.MultinomialNB(), breast_cancer.data,
                      breast_cancer.target).mean()
print 'GaussianNB classifier:', \
      cross_val_score(naive_bayes.GaussianNB(), breast_cancer.data, \
                      breast_cancer.target).mean()
```

```
BernoulliNB classifier: 0.627420402859
MultinomialNB classifier: 0.894579040193
GaussianNB classifier: 0.936749280609
```

```
In [34]: digits = datasets.load_digits()  
print digits.DESCR
```

```
Optical Recognition of Handwritten Digits Data Set  
=====
```

Notes

Data Set Characteristics:

```
:Number of Instances: 5620  
:Number of Attributes: 64  
:Attribute Information: 8x8 image of integer pixels in the range 0..16.  
:Missing Attribute Values: None  
:Creator: E. Alpaydin (alpaydin '@' boun.edu.tr)  
:Date: July; 1998
```

This is a copy of the test set of the UCI ML hand-written digits datasets
<http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits> (<http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits>)

The data set contains images of hand-written digits: 10 classes where each class refers to a digit.

Preprocessing programs made available by NIST were used to extract normalized bitmaps of handwritten digits from a preprinted form. From a total of 43 people, 30 contributed to the training set and different 13 to the test set. 32x32 bitmaps are divided into nonoverlapping blocks of 4x4 and the number of on pixels are counted in each block. This generates an input matrix of 8x8 where each element is an integer in the range 0..16. This reduces dimensionality and gives invariance to small distortions.

For info on NIST preprocessing routines, see M. D. Garris, J. L. Blue, G. T. Candela, D. L. Dimmick, J. Geist, P. J. Grother, S. A. Janet, and C. L. Wilson, NIST Form-Based Handprint Recognition System, NISTIR 5469, 1994.

References

- C. Kaynak (1995) Methods of Combining Multiple Classifiers and Their Applications to Handwritten Digit Recognition, MSc Thesis, Institute of

Graduate Studies in Science and Engineering, Bogazici University.

- E. Alpaydin, C. Kaynak (1998) Cascading Classifiers, Kybernetika.
- Ken Tang and Ponnuthurai N. Suganthan and Xi Yao and A. Kai Qin. Linear dimensionality reduction using relevance weighted LDA. School of Electrical and Electronic Engineering Nanyang Technological University. 2005.
- Claudio Gentile. A New Approximate Maximal Margin Classification Algorithm. NIPS. 2000.

```
In [39]: print 'BernoulliNB classifier:', \
          cross_val_score(naive_bayes.BernoulliNB(), digits.data,
                          digits.target).mean()
print 'MultinomialNB classifier:', \
      cross_val_score(naive_bayes.MultinomialNB(), digits.data,
                      digits.target).mean()
print 'GaussianNB classifier:', \
      cross_val_score(naive_bayes.GaussianNB(), digits.data, \
                      digits.target).mean()
```

```
BernoulliNB classifier: 0.825823650778
MultinomialNB classifier: 0.870877148974
GaussianNB classifier: 0.818600380355
```

Ответы на вопросы

1. Максимальное качество на датасете breast_cancer получилось примерно 0.937 у классификатора GaussianNB
2. Максимальное качество на датасете digits получилось примерно 0.871 у классификатора MultinomialNB
3. c, d