Scikit-learn ゼミ

1.11 Feature selection
  1.11.1 Removing features with low variance
  1.11.2 Univariate feature selection

新納 浩幸

Feature selection

素性選択、識別に有効な素性を選出

最も簡易な手法
　　分散が小さい素性を削除する

すべてのデータのある次元（素性）の値が
全部０だとすると分散は０、、、
この次元（素性）はなくてもOK

01分布の分散

$$P(x)=\begin{cases} p & (x=1) \\ 1-p & (x=0) \end{cases}$$

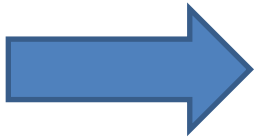
$$V(X)=p(1-p)$$

素性の値が 0 か 1 のとき、1 になる割合 p’ を
調べて、p’ > p か p’ < 1 – p のときに、
その素性を削除する

実行例

```
>>> from sklearn.feature_selection import VarianceThreshold
>>> X = [[0,0,1],[0,1,0], [1,0,0],[0,1,1],[0,1,0],[0,1,1]]
>>> sel = VarianceThreshold(threshold=( 0.8 * ( 1 - 0.8 )))
>>> sel.fit_transform(X)
array([[0, 1],
       [1, 0],
       [0, 0],
       [1, 1],
       [1, 0],
       [1, 1]])
>>>
```

第１次元の１の割合は 1/6 < 0.2



第１次元を削除

Univariate feature selection

単変量の検定を用いて素性を選択

例えば、χ2 検定を使えば、独立性の検定ができる。つまり、素性とラベルの独立の度合いの強い素性を削除する。

実行例

```
>>> from sklearn.datasets import load_iris
>>> from sklearn.feature_selection import SelectKBest
>>> from sklearn.feature_selection import chi2
>>> iris = load_iris()
>>> X, y = iris.data, iris.target
>>> X.shape
(150L, 4L)
>>> X_new = SelectKBest(chi2, k=2).fit_transform(X,y)
>>> X_new.shape
(150L, 2L)
>>>
```

選択する素性の数
削除する数ではない