# Bayes分类器

### 一、原理

Naive Bayes（朴素贝叶斯）方法是一类监督学习算法。“朴素”的含义是假设特征之间两两独立。

由条件概率的定义：P(A|B) = P(AB)/P(B), P(B|A) = P(AB)/P(A)，于是有:

P(A|B) = P(B|A)\*P(A)/P(B)

将样本(x, y)的类别作为一个事件，x的每个特征作为一个事件。则有：

P(y \mid x_1, \dots, x_n) = \frac{P(y) P(x_1, \dots x_n \mid y)}
                                 {P(x_1, \dots, x_n)} （1）

由独立性假设，公式（1）可转化为：

P(y \mid x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}
                                 {P(x_1, \dots, x_n)}

由于在给定训练集中，P(x1, ..., xn)是一个常量，因此有如下分类规则：

P(y \mid x_1, \dots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)

\Downarrow

\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),

不同的bayes分类器使用不同的P(x_i \mid y)分布假设。

假设模型符合高斯分布：p(y|X,w,\alpha) = \mathcal{N}(y|X w,\alpha)，则α可以作为要估计的参数。

### 二、scikit-learn：Bayesian Ridge Regression

[BayesianRidge](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.BayesianRidge.html#sklearn.linear_model.BayesianRidge) estimates a probabilistic model of the regression problem as described above. The prior for the parameter wis given by a spherical Gaussian:

p(w|\lambda) =
\mathcal{N}(w|0,\lambda^{-1}\bold{I_{p}})

The priors over \alpha and \lambda are chosen to be [gamma distributions](http://en.wikipedia.org/wiki/Gamma_distribution), the conjugate prior for the precision of the Gaussian.

The resulting model is called Bayesian Ridge Regression, and is similar to the classical [Ridge](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html#sklearn.linear_model.Ridge). The parameters w, \alpha and \lambda are estimated jointly during the fit of the model.

## 三、实现

（参考《使用决策树算法在印第安人糖尿病数据实现分类器》。）实验数据依旧采用印第安人糖尿病数据。训练模型的代码：

from sklearn.linear\_model import BayesianRidge

def **trainModel**(x, y):

# Fit the Bayesian Ridge Regression

clf = BayesianRidge(compute\_score=True)

clf = clf.fit(x, y)

return clf

测试：

# DT result

print (*'DT result(Max-depth=3):'*)

reg1 = DT.trainRegressionModel(x, y, 3)

MLUtil.testBinaryModelFloat(xc, yc, reg1)

println ()

# Bayesian result

print (*'Bayesian result:'*)

reg2 = bayesian.trainModel(x, y)

MLUtil.testBinaryModelFloat(xc, yc, reg2)

## 四、结果

DT result(Max-depth=3):

predict precision:

0.762711864407

Bayesian result:

predict precision:

0.78813559322

实验结果说明该组数据上贝叶斯回归分类效果略好于决策树。