基于libSvm

- 总样本数**28056**， 其中正样本**2796**, 负样本**25260**。
- 随机取**5000**个样本训练，其余测试。
- 样本归一化，在训练样本上，求出每个维度的均值和方差，在训练和测试样本上同时归一化。

$$newX = \frac{X - mean(X)}{std(X)}$$

高斯核: $e^{\frac{-\|x-y\|^2}{\delta}}$

- 高斯核
- **5-fold cross validation**，在

CScale = [2^(-5), 2^15]; gamma = [2^(-15),2^3];

上遍历求识别率的最大值。

上述**C**和**gamma**的区间设置参见**LIBSVM**自带的介绍：

a practical guide to support vector classification

- 训练参数设置 **svmtrain(yTraining, xTraining, cmd)**

cmd参数如下：

svm-type

(1) –s 0 "-s svm_type : set type of SVM (default 0)\n"
" 0 -- C-SVC (multi-class classification)\n"
" 1 -- nu-SVC (multi-class classification)\n"
" 2 -- one-class SVM\n"
" 3 -- epsilon-SVR (regression)\n"
" 4 -- nu-SVR (regression)\n"

kernel-type

(2) –t 2
"-t kernel_type : set type of kernel function (default 2)\n"
" 0 -- linear: u'*v\n"
" 1 -- polynomial: (gamma*u'*v + coef0)^degree\n"
" 2 -- radial basis function: exp(-gamma*|u-v|^2)\n"
" 3 -- sigmoid: tanh(gamma*u'*v + coef0)\n"
" 4 -- precomputed kernel (kernel values in training_instance_matrix)\n"

→ $K(x_1, x_2)$

输入矩阵（5000个样本）

$$\begin{bmatrix} K(x_1,x_1) & \cdots & K(x_1,x_{5000}) \\ \vdots & & \\ K(x_{5000},x_1) & \cdots & K(x_{5000},x_{5000}) \end{bmatrix}$$

$R^{5000\times 5000}$

(3) –c CVALUE
"-c cost : set the parameter C of C-SVC, epsilon-SVR, and nu-SVR (default 1)\n"

(4) –g gammaValue
"-g gamma : set gamma in kernel function (default 1/num_features)\n"

(5) –v 5
"-v n : n-fold cross validation mode\n"

- 训练后获得的参数

（1）**C = 16， gamma = 0.0825**

（2）支持向量（即**alpha**不为**0**的向量）：**358**个 **(162**个正样本，**196**个负样本**)**

（3）**b = 6.2863**

对于一个测试样本X：

若 $\sum_{i=1}^{n} a_i y_i K(x_i, x) + b \geq 0$ $y=1$

经验：支持向量在总样本 20%~30%

否则 $y=-1$

交叉验证 (cross validation)

5000个样本分5组 每组1000个

(5-fold)

① (a b c d) 训练 e 测试

⋮

C_5^4 训练 $5-C_5^4$ 测试

总样本28056 正样本2796 负样本25260

随机取5000个样本训练 其余测试

样本归一化 在训练样本上 求出每个维度的均值和方差
在训练和测试样本上同时归一化

$$newX = \frac{X - mean(X)}{std(X)}$$

高斯核

5-fold cross validation 在
$C\ Scale = [2^{-5}, 2^{15}]$
$gamma = [2^{-15}, 2^{3}]$

混淆矩阵

		预测	
		正样本	负样本
实际	正样本	TP	FN
	负样本	FP	TN

取好不要发生！ FP=0 TP越好

TP

FP

ROC曲线 (Receiver Operating Characteri)

TP + FN = 1 ①

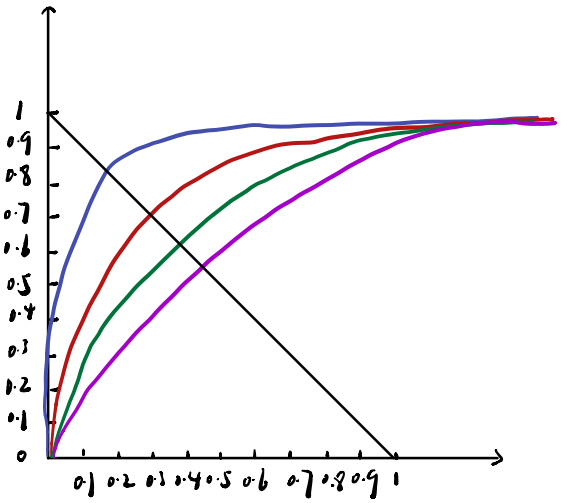
FP + TN = 1 ②

对同一个系统来说 若TP增加 FP也增加 ③

TP增加 ③ FP增加 ② TN减少

① ↕

FN减少



等错误率(Equal Error Rate, EER)

是两类错误 FP和 FN 相等时的错误率

可以直观的表示系统性能