MATLAB最大均值差异(Maximum Mean Discrepancy)

作者：凯鲁嘎吉 - 博客园 http://www.cnblogs.com/kailugaji/

更多内容，请看标签：MATLAB、聚类

注：X与Y数据维度必须一致！

1．MMD介绍

➢ *Maximum Mean Discrepancy* ——最大均值差异

- MMD：用来度量源域与目标域数据之间的距离
- MMD值越小，源域与目标域越相似

$$MMD^2(X,Y)=\left\|\frac{1}{m}\sum_{i=1}^{m}\phi(x_i)-\frac{1}{n}\sum_{j=1}^{n}\phi(y_j)\right\|_H^2$$

$$=\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}K(x_i,x_j)-\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}K(x_i,y_j)+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(y_i,y_j)$$

$$K(x,y)=\mathrm{e}^{-\frac{\|x-y\|^2}{2\sigma^2}}$$

- Gretton A , Borgwardt K M , Rasch M , et al. A Kernel Two-Sample Test[J]. Journal of Machine Learning Research, 2012, 13:723-773.

2．MATLAB程序

数据

注：数据集仅供参考，并不能真正用于研究中。

源域：
2.1789 1.7811 5.079   4.9312
0.8621 2.1287 4.9825 2.3388
2.6347 1.9563 4.5392 4.8442
2.7179 2.9001 4.9027 4.8582
2.6686 1.6799 4.3792 4.6411
1.6736 2.3081 4.8384 3.2979
1.5666 2.6467 5.0504 4.459
-0.5611 2.2365 4.3925 5.1316
5.6693 1.7355 4.5335 4.6407
3.2032 2.103   4.1948 5.2605
3.3525 2.8301 4.6383 5.6972
-1.0407 3.5198 4.7106 4.9243
3.9229 2.1161 4.5666 1.772
2.5607 3.802   4.2681 4.6322
3.3072 2.5083 4.6095 2.2236
2.7121 2.4338 4.136   2.2348
5.3547 2.1088 4.402   4.9884
1.8302 1.4921 4.6216 3.5862
2.8891 2.1286 4.6419 3.8606
-0.0896 2.6894 3.6843 6.6392
3.1404 1.9461 4.2604 5.9859
2.3406 3.1988 5.0872 4.7518
2.5067 2.9704 4.2749 4.3441
8.2153 1.7592 5.2409 3.8201
0.3027 2.7589 3.9826 4.8484
4.0223 1.7566 4.6219 4.92
6.1367 2.1098 4.7832 5.4567
4.9795 2.418   4.7726 3.1959
-1.0746 2.4311 4.7683 4.5599
5.4939 2.6046 4.4663 5.1159
4.5709 1.9838 4.9596 4.9317
1.3746 2.6845 5.1921 3.2068
1.7178 0.7976 4.6948 3.7012
目标域：
1.9584 2.0242 4.7594 2.587
-2.8342 3.4594 4.4371 5.2375
1.6251 2.7737 5.0145 6.3262

```
0.7016  2.5265  4.8881  3.2105
3.5579  2.5773  4.856   4.283
4.3282  2.7581  4.7095  6.715
3.1619  2.5427  4.1323  5.5883
4.9933  2.2985  3.8455  3.8381
3.2214  2.6478  4.3276  2.5246
-0.2848 2.5853  4.6481  3.4857
2.876   1.5096  3.9921  2.4505
0.8559  2.5633  5.483   3.0589
4.2149  2.6618  4.2017  3.3713
```

MMD

```
function mmd_XY=my_mmd(X, Y, sigma)
%Author：kailugaji
%Maximum Mean Discrepancy 最大均值差异 越小说明X与Y越相似
%X与Y数据维度必须一致，X, Y为无标签数据，源域数据，目标域数据
%mmd_XY=my_mmd(X, Y, 4)
%sigma is kernel size, 高斯核的sigma
[N_X, ~]=size(X);
[N_Y, ~]=size(Y);
K = rbf_dot(X,X,sigma); %N_X*N_X
L = rbf_dot(Y,Y,sigma);  %N_Y*N_Y
KL = rbf_dot(X,Y,sigma);  %N_X*N_Y
c_K=1/(N_X^2);
c_L=1/(N_Y^2);
c_KL=2/(N_X*N_Y);
mmd_XY=sum(sum(c_K.*K))+sum(sum(c_L.*L))-sum(sum(c_KL.*KL));
mmd_XY=sqrt(mmd_XY);
```

Guassian Kernel

```
function H=rbf_dot(X,Y,deg)
%Author：kailugaji
%高斯核函数/径向基函数 K(x, y)=exp(-d^2/sigma), d=(x-y)^2, 假设X与Y维度一样
%Deg is kernel size,高斯核的sigma
[N_X,~]=size(X);
[N_Y,~]=size(Y);
G = sum((X.*X),2);
H = sum((Y.*Y),2);
Q = repmat(G,1,N_Y(1));
R = repmat(H',N_X(1),1);
H = Q + R - 2*X*Y';
H=exp(-H/2/deg^2);  %N_X*N_Y
```

结果

```
>> mmd_XY=my_mmd(x, y, 4)

mmd_XY =

    0.1230
```

3. 参考文献

Gretton, A., K. Borgwardt, M. Rasch, B. Schoelkopf and A. Smola: A Kernel Two-Sample Test. JMLR 2012.

Gretton, A., B. Sriperumbudur, D. Sejdinovic, H, Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu: Optimal kernel choice for large-scale two-sample tests. NIPS 2012.