相似性度量

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

1．基于范数的度量

1.1 L_1范数——Manhattan Distance(曼哈顿距离)

1.2 L_2范数——Euclidean Distance(欧氏距离)

1.3 L_∞范数——Chebyshev Distance(切比雪夫距离)

1.4 L_p范数——Minkowski Distance(闵可夫斯基距离)

1.5 $L_{2,1}$范数

1.1 L_1 范数——Manhattan Distance(曼哈顿距离)

$$d(x,y)=\sum_{i=1}^{N}|x_i-y_i|$$

1.2 L_2 范数——Euclidean Distance(欧氏距离)

$$d(x,y)=\left(\sum_{i=1}^{N}(x_i-y_i)^2\right)^{\frac{1}{2}}$$

1.3 L_∞范数——Chebyshev Distance(切比雪夫距离)

$$d(x,y)=\max_i|x_i-y_i|$$

1.4 L_p 范数——Minkowski Distance(闵可夫斯基距离)

$$d(x,y)=\left(\sum_{i=1}^{N}(x_i-y_i)^p\right)^{\frac{1}{p}}$$

1.5 $L_{2,1}$ 范数

$$d(x,y)=\sum_{i=1}^{N}\left(\sum_{j=1}^{M}(x_i-y_j)^2\right)^{\frac{1}{2}}$$

2．基于协方差的度量

2.1 Mahalanobis Distance(马氏距离)

2.2 Correlation Distance(相关距离)

2.1 Mahalanobis Distance(马氏距离)

$$d(x,y)=\left((x-y)^T S^{-1}(x-y)\right)^{\frac{1}{2}}$$

2.2 Correlation Distance(相关距离)

$$d(x,y)=1-\rho_{xy}=1-\frac{\mathrm{cov}(x,y)}{\sigma_x\sigma_y}=1-\frac{E((x-E(x))(y-E(y)))}{\sqrt{\mathrm{var}(x)}\sqrt{\mathrm{var}(y)}}$$

3．基于幅度的度量

3.1 Cosine Similarity(余弦距离)

3.2 Tonimoto系数

3.1 Cosine Similarity(余弦距离)

$$\cos\theta = \frac{x^T y}{\|x\| \cdot \|y\|} = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}}$$

3.2 Tonimoto 系数

$$T(x,y) = \frac{x^T y}{\|x\|^2 + \|y\|^2 - x^T y}$$

4. Jaccard Distance

$$J_\delta(A,B) = 1 - J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

5. 基于概率分布的度量

5.1 互信息

5.2 Kullback-Leibler Divergence (KL散度)

5.3 Jensen-Shannon divergence(JS散度)

5.4 Wasserstein distance(推土机距离)

5.1 互信息

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

5.2 Kullback–Leibler Divergence (KL 散度)

$$D_{KL}(p\|q)=\sum_{i=1}^{N}p(x_i)\log\frac{p(x_i)}{q(x_i)}$$

5.3 Jensen–Shannon divergence(JS 散度)

$$JS(p\|q)=\frac{1}{2}D_{KL}\left(p\|\frac{p+q}{2}\right)+\frac{1}{2}D_{KL}\left(q\|\frac{p+q}{2}\right)$$

5.4 Wasserstein distance(推土机距离)

$$W[p,q]=\inf_{\gamma\in\prod[p,q]}\iint\gamma(x,y)d(x,y)\mathrm{d}x\mathrm{d}y$$

6．基于核函数的度量

6.1 高斯核

6.2 q次多项式核

6.3 Maximum mean discrepancy(最大均值差异)

6.1 高斯核

$$d(x,y)=1-K(x,y)=1-\mathrm{e}^{-\frac{\|x-y\|^2}{2\sigma^2}}$$

6.2 q 次多项式核

$$d(x,y)=1-K_q(x,y)=1-(c+x^Ty)^q$$

6.3 Maximum mean discrepancy(最大均值差异)

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

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

7. Hamming Distance(汉明距离)

$$d(x,y)=\sum_{i=0}^{n}x[i]\oplus y[i]$$

8. 参考

[1] 范数：向量范数与矩阵范数

[2] 最大均值差异：MATLAB最大均值差异(Maximum Mean Discrepancy)

[3] 马氏距离：MATLAB求马氏距离(Mahalanobis distance)

[4] 相关系数：MATLAB实例：求相关系数、绘制热图并找到强相关对

[5] 互信息：MATLAB聚类有效性评价指标（外部）

[6] Jaccard Distance：MATLAB聚类有效性评价指标（外部 成对度量）

[7] KL散度与JS散度：MATLAB小函数：计算KL散度与JS散度

[8] 余弦相似度：Python小练习：向量之间的距离度量