Distance

Euclidean distance

유클리드 거리(Euclidean distance)는 두 점간의 거리를 계산할 때 주로 사용된다. 일반적으로 많이 사용되는 공식이다.

$$d(\mathbf{a},\mathbf{b}) = \|\mathbf{a}-\mathbf{b}\| = \sqrt{(\mathbf{a}-\mathbf{b})^T(\mathbf{a}-\mathbf{b})} = \sqrt{\sum_j (a_j - b_j)^2}$$

$\|\mathbf{a}\|$는 $\mathbf{a}$의 2-norm 이고, a_j와 b_j는 벡터 $\mathbf{a}$와 $\mathbf{b}$의 j번째 원소이다.

Generalized Euclidean

양정치행렬(positive-definite matrix) $\mathbf{W}$를 고려한 generalized Euclidean distance는 다음과 같다.

$$d_{\mathbf{W}}^2(\mathbf{a},\mathbf{b}) = (\mathbf{a}-\mathbf{b})^T\mathbf{W}(\mathbf{a}-\mathbf{b})$$

χ^2 distance

$$\chi^2 = (\mathbf{a}-\mathbf{b})^T\mathbf{W}(\mathbf{a}-\mathbf{b})$$

$$\mathbf{W} = (\mathrm{diag}\{\mathrm{c}\})^{-1}$$

$\mathbf{W}$가 대각행렬로서 c가 각 열의 가중치를 나타낸다.

Mahalanobis distance

다변수 가우시안 분포 $P \sim N(\mathbf{a},\Sigma)$와 점 $\mathbf{b}$에 대한 마할라노비스 거리는 다음과 같다.

$$d_{\Sigma}^2(\mathbf{a},\mathbf{b}) = (\mathbf{a}-\mathbf{b})^T\Sigma^{-1}(\mathbf{a}-\mathbf{b})$$

즉, 가중치 행렬 $\mathbf{W}$가 공분산(covariance) 행렬 Σ의 역이 될 때($\mathbf{W}=\Sigma^{-1}$)이다.

Minskowki's distance

$$d_p(\mathbf{a},\mathbf{b}) = \|\mathbf{a}-\mathbf{b}\|_p = \left(\sum_j |a_j - b_j|^p\right)^{\frac{1}{p}}$$

가장 일반적으로 사용되는 Minkowski 거리의 차수는 1, 2, ∞ 이다.

Hellinger distance

$$d(\mathbf{a},\mathbf{b}) = \left(\sum_j \left(\sqrt{a_j} - \sqrt{b_j}\right)^2\right)^{\frac{1}{2}}$$

Squared Hellinger distance

두 정규분포를 가지는 점 $P \sim N(\mu_1, \sigma_1^2)$ 와 $Q \sim N(\mu_2, \sigma_2^2)$ 에 대한 squared Hellinger 거리는 다음과 같다.

$$H^2(P,Q) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}} \exp\left(-\frac{1}{4}\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}\right)$$

두 다변수 가우시안 분포 $P \sim N(\mu_1, \Sigma_1)$ 와 $Q \sim N(\mu_2, \Sigma_2)$ 에 대한 squared Hellinger 거리는 다음과 같다.

$$H^2(P,Q) = 1 - \sqrt{\frac{|\Sigma_1|^{1/2}|\Sigma_2|^{1/2}}{|\overline{\Sigma}|}} \exp\left(-\frac{1}{8}(\mu_1-\mu_2)^T \overline{\Sigma}^{-1}(\mu_1-\mu_2)\right) \quad \text{where } \overline{\Sigma} = \frac{\Sigma_1+\Sigma_2}{2}$$