Chapter 3
Distance Metrics

1

목차

1

Motivation

- 2장에서는 Correlation matrix에 대해서 확인
  - 하지만 Correlation은 codependence를 측정하는데 있어 중요한 단점을 지니고 있음
- 정보이론을 통해 이를 해결
  - 정보이론의 핵심은 확률변수와 연계된 불확실성을 계량화하는데 있음
- 머신러닝은 불확실성과 연계되어 있음
  - 머신러닝의 목적은 문제 상황의 해결책에 내재한 불확실성을 줄여가는데 있기 때문이다
- 본 장의 목적
  - 1. 의사결정나무의 목적함수 설정
  - 2. 분류 문제의 loss function 설정
  - 3. 두 확률변수의 거리 계산
  - 4. 클러스터 비교
  - 5. 특징 추출

2

A Correlation-Based Metric

- **상관관계는 선형 관계를 표상하는데 적절한 기준임**
- **상관관계 행렬이** denoiesd**되고** detoned**된다면, 시스템 내의 정보 구조를 보여주는데 적합함**
  - **서로 높은 상관관계가 있는 클러스터를 발견하는데 상관관계를 이용할 수 있음**
- **하지만** correlation**은 척도가 될 수 없다**
  - Nonnegativity**와 삼각부등식 조건을 만족하지 않기 때문**
    - $|x+y| \leq |x| + |y| \quad \forall x, y \in R$

2

A Correlation-Based Metric

- Metric
  - $d : X \times X \ \rightarrow [0, \infty]$
  - **특성**
    - $d(x, y) = 0 \leftrightarrow x = y$
    - $d(x, y) = d(y, x)$
    - $d(x, y) + d(y, z) \geq d(x, z)$

2

A Correlation-Based Metric

- **유클리드 거리 계산**

  - $d_\rho[X,Y] = \sqrt{\frac{1}{2}(1-\rho[X,Y])}$
  - $d[X,Y] = \sqrt{\sum(X_t - Y_t)^2}$
  - $x = \frac{X - \bar{X}}{\sigma[X]}, y = \frac{Y - \bar{Y}}{\sigma[Y]}$
  - $d[x,y] = \sqrt{\sum(x_t - y_t)^2} = \sqrt{\sum x_t^2 + \sum y_t^2 - 2\sum x_t y_t} = \sqrt{T + T - 2T\sigma[x,y]} = \sqrt{2T(1-\rho[x,y])} = \sqrt{4T} d_\rho[X,Y]$
  - $\rho[X,Y] \in [-1,1] \rightarrow d_\rho[X,Y] \in [0,1]$

3

Marginal and Joint Entropy

- 상관관계
  - 두 확률변수의 선형 관계 포착
    - 비선형 관계는 포착하지 않음
  - Outlier에 민감함
  - 다변량 정규분포가 아닌 경우에 효용성을 입증하기 힘듦
    - 두 변수간의 상관관계는 이변량 정규분포를 가정함

3

Marginal and Joint Entropy

- **엔트로피**
  - **수식**
    - $H[x] = -\sum_{x \in S_x} p[x]log\left[p[x]\right]$
    - $0log[0] = 0$
    - $lim_{p \to 0^+} p\ log[p] = 0$
  - Entropy는 놀라운 값의 기댓값
  - $\frac{1}{p[x]}$: 관측값이 얼마나 놀라운 결과인지 측정
    - 확률이 낮을수록 놀라운 결과이기 때문
    - log함수는 $p[x]$가 $\frac{1}{p[x]}$상쇄하는 것 막음
  - 모든 결과값이 하나의 값으로 통일될 때 가장 낮으며, X가 균등분포일 때 가장 높다
    - $p[x] = \frac{1}{||S_X||} \rightarrow \log\left[||S_X||\right]$

4 Conditional Entropy

- **엔트로피**
  - **공동 엔트로피**
    - $H[X,Y] = -\sum p[x,y] log\, [p[x,y]]$
    - $H[X,Y] = H[Y,X],\ \ H[X,X] = H[X], H[X,Y] \geq max\{H[X], H[Y]\}, H[X,Y] \leq H[X] + H[Y]$
  - **조건부 엔트로피**
    - $H[X|Y] = H[X,Y] - H[Y] = -\sum p[y] \sum p[x|Y - y] log\, [p[x|Y = y]]$
      - Y 정보를 알게 되었을 때의 X 예상에 대한 불확실성
      - $H[X|X] = 0, H[X] \geq H[X|Y]$

5 KullBack-Leibler Divergence

- **특징**
  - p와 q가 같은 확률 공간 상의 서로 다른 확률 분포
  - $D_{KL}[p||q] = -p[x]\log\sum\left[\frac{q[x]}{p[x]}\right] = p[x]\log\sum[\left[\frac{p[x]}{q[x]}\right]]$
    - p가 참고 분포 q로부터 얼마나 떨어져 있는지에 대한 측정
      - 측도가 아님
        - Nonnegative 만족($D_{KL}[p||q \geq 0)$
        - Symmetry 만족 못함 ($D_{KL}[p||q] \neq D_{KL}[q||p]$)
        - Triangle Inequality 만족 못함
- **비교**
  - Joint entropy의 경우 같은 확률 공간 하에 있지 않아도 되지만, KLD는 같은 공간 하에 존재해야 한다.

6 Cross-Entropy

- Cross Entropy
  - 수식
    - $\boldsymbol{H_C[p||q] = -\sum p[x] log\left[q[x]\right] = H[x] + D_{KL}[p||q]}$
  - 해석
    - $\boldsymbol{X}$와 연계된 불확실성
    - 참 분포 $\boldsymbol{p}$ 대신 잘못된 분포 $\boldsymbol{q}$를 활용하여 측정
  - 특징
    - 분류 문제에서 자주 사용하는 척도
    - 금융 영역에서 의미가 깊음

7 Mutual Information

- **수식**
  - $I[X,Y] = H[X] - H[X|Y] = H[X] + H[Y] - H[X,Y] = \sum\sum p[x,y] log\left[\frac{p[x,y]}{p[x]p[y]}\right]$
  - $D_{KL}[p[x,y]||p[x]p[y]] = \sum p[y] \sum p[x|y] log\left[\frac{p[x|y]}{p[x]}\right]$
  - $E_Y[D_{KL}[p[x|y]||p[x]]] = \sum p[x] \sum p[y|x] log\left[\frac{p[y|x]}{p[y]}\right]$
  - $E_X[D_{KL}[p[y|x]||p[y]]]$
- **해석**
  - Y를 알게 되면서 줄어드는 X에 대한 불확실성

7

Mutual Information

- 수식
  - $\boldsymbol{I[X,Y] \geq 0, I[X,Y] = I[Y,X], I[X,X] = H[X]}$
- 특징
  - 독립일 경우
    - $\boldsymbol{p[x,y] = p[x]p[y]}$
    - $\boldsymbol{I[X,Y] = 0}$
  - Metric 아님
    - $\boldsymbol{I[X,Z] \nleq I[X,Y] + I[Y,Z]}$

8

Variation of information

- 수식
  - $VI[X,Y] = H[X|Y] + H[Y|X] = H[X] + H[Y] - 2I[X,Y]$
  - $= 2H[X,Y] - H[X] - H[Y] = H[X,Y] - I[X,Y]$
- 범위
  - Lower bound
    - $VI[X,Y] = 0 \leftrightarrow X = Y$
  - Upper bound
    - $VI[X,Y] \leq H[X,Y]$
- 특징
  - Metric
- 대안
  - Bound Metric
    - $\widetilde{VI}[X,Y] = \frac{max\{H[X|Y],H[Y|X]\}}{max\{H[X],H[Y]\}} = 1 - \frac{I[X,Y]}{max\{H[X],H[Y]\}}$

9

Discretization

- **수식**
  - $\boldsymbol{H}[\boldsymbol{X}] = -\int$

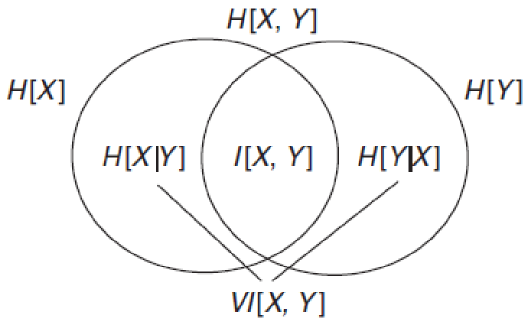

Figure 3.1 Correspondence between joint entropy, marginal entropies, conditional entropies, mutual information, and variation of information.

Thank you