**예측 모델 Homework #1**

**학번 : 2019021168 이름 : 조억**

**1. Conditional probability (조건부 확률)**

조건부 확률은 단어 그대로 조건이라는 특정 Condition이 발생했을 때의 확률입니다. 예를 들어 A, B 사건의 확률이 있다고 하고, 전체 N에서 발생할 A사건이 발생할 확률 P(A)는 n(x)를 x가 발생할 경우의 총 개수라고 정의한다면, 가 되며, 마찬가지로 P(B) 또한 이라고 할 수 있습니다. 여기에서 조건부의 확률의 개념을 얘기할 수 있는 단계가 되었는데요, 우선 B의 사건이 발생한 경우의 수중에 A라는 사건이 또한 발생한 경우의 수를 우리는 P(A|B)라고 표현할 수 있습니다. 수식을 표현할려면 B의 개수 중에서 의 경우를 고려하면 됩니다. 그래서 가 조건부 확률의 수식이 됩니다.

전체 확률의 법칙에 의해서

이를 통해서 우리는 베이즈 정리 또한 유도해낼 수 있습니다.

예를 들어 비가 올때의 교통사고의 확률을 같은 것이 조건부의 확률이 되며 비가 온다라는 사건은 여기에서 B 그리고 교통사고의 확률은 A와 매핑되며, 우리는 어떤 가정을 하고 그 가정하에 특정 사건이 일어날 확률을 구하여 현실 세계에 사용할 수 있겠습니다. 그래서 예를 들어보면 우리는 교통사고 중에서 비가 온 날의 교통사고가 가 될 수 있습니다.

**2. Joint probability distribution (결합확률분포), Conditional probability distribution (조건부확률분포), Marginal probability distribution (주변확률분포)**

Joint Probability distribution은 예를 들어, 랜덤 변수가 3개 이상의 경우(multivariate distribution)는 복잡하니 심플하게 bivariate distribution의 경우)를 고려하여 랜덤 변수 X,Y가 있다고 가정을 해보자. 이 2개의 확률 변수가 확률 공간에서 특정 공간(연속적인 경우)나 아니면 이산 집합(discrete 의 값을 가질 경우)에 같이 겹치는 경우를 Joint probability distribution이라고 할 수 있습니다. 이 Joint Probability distribution의 경우에는 cumulative distribution function라고도 불려지며, discrete한 값을 가질 경우에는 Joint probability density function 이며, continuous 한 값을 가질 경우에는 Probability mass function이라고 불려집니다. 이 분포는 또한 다른 2가지의 분포를 찾는데도 사용이 가능합니다. 첫번째는 Marginal probability distribution이며 두개의 변수 중 어느 특정 변수의 범위를 제한하지 않고 다른 변수 하나의 확률 분포의 면적을 구했을 경우에는 해당됩니다. 두번째는 위의 조건부확률과 유사하다고 볼 수 있으며 어떤 변수의 특정 값에 조건부로 하여 다른 변수의 부분 집합의 확률을 구하는 것을 Conditional probability distribution이라고 합니다. Conditional probability distribution의 경우에는 대상이 되는 랜덤 변수의 값에 dependent하고 Marginal Probability distribution의 경우에는 다른 랜덤 변수를 고려하지 않고 대상 랜덤 변수의 확률을 전체 고려하는 것으로 생각 할 수 있겠습니다.

이를 수식으로 적자면 두개의 Random variable X,Y가 있다고 했을 때,

1. Joint cumulative distribution function
2. Conditional probability distribution
3. Marginal probability distribution

**3. Independent random variable (독립확률변수)**

이 변수는 우리가 가정한 어떤 실험에서 다른 어느 확률 변수에 영향을 안 미치는 확률 변수라고 할 수 있다. 즉 이 말은 우리가 X를 이 변수라고 가정한다면 이 변수는 다른 변수 Y를 예측하는데 도움이 전혀 안된다고 볼 수 있겠다. 조건부 확률 또한 관련이 없으니 P(X)자체가 된다고 볼 수 있다.

아래와 같이도 표기가 가능하다.

그리고 동시 확률의 경우에도 두 확률의 곱으로 표현이 가능하다.

Independent random variable X,Y에 대해서 다음 조건이 성립된다.

**4. Correlation (상관계수), Autocorrelation (자기상관계수)**

우리는 어떤 관측치 observation이 두개의 set이 있고 이것에 대한 연관성을 보고 싶다고 했을 경우 두 변인 간의 관계를 측정하고 묘사하기 위해 통계적인 수치인 상관계수를 사용한다 서로 다른 두개의 변인의 관측치를 공분산을 구하게 되면 상관의 정도가 관측이 되나 그 관측치의 범위나 개수에 따라서 공분산에서 각 변인의 표준 편차를 나누어서 일반화를 시킨 값이 상관계수가 되겠다.

그런데 Autocorrelation의 경우는 서로 다른 두개의 변인을 구하는게 아니라 시계열 데이터에 대한 데이터를 분석할 때 이 시계열 데이터가 시계열 차 (lag)에 줌으로써 다른 데이터셋을 shift 시킨 데이터셋을 가지고 서로 비교하는 것을 자신과 비교한다고해서 Autocorrelation이라고 한다. 그래서 시계열 데이터를 확보하여서 이 데이터셋이 시계열에 유의미한지 판단하는 계수로서 사용될 수 있겠다. 그래서 어떤 반복되는 패턴을 찾아내는 것이며 신호처리나 시계열 시그널의 데이터를 처리하는데 사용될 수 있겠다.

**5. Eigenvalue, Eigenvector**

이 정리를 통해서 고유값과 고유벡터가 중요한 기초를 담당하고 있다 생각이 들었습니다. 그래서 많은 이론에서 이 개념이 등장하고 사람들이 중요하다고 하는지를 깨닫는 계기가 되었습니다. SVD(특이값분해), PCA(주성분분석)에서 이 개념을 이용애서 값을 분해하고 축소를 해서 저차원으로 피쳐를 줄이는 바탕이 된다는 이해와 직관의 시간을 가졌습니다. 정의를 말씀드리면, 우선 행렬 A가 있고 이에 대한 벡터를 이용한 선형 변환을 했더니 그 A의 어떤 상수를 곱한 값과 같을 때 우리는 행렬 A는 그것의 Eigenvalue와 Eigenvector를 말할 수 있습니다. 여기에서 Eigenvalue는 scaler역할을 해서 선형변환시에 해당 값을 스케일링 처리하여 값을 늘이거나 줄이는 역할을 하는 것이고 Eigenvector는 기존의 선형 변환을 또 다른 선형 변환으로 처리하는 열벡터이며 이 열벡터로 인한 선형변환의 결과는 기존 선형변환의 상수배로 표현이 될 수 있는 고유한 벡터라고 할 수 있겠습니다. 이를 기하학적인 관점으로 보았을 때 A의 선형변환에 변형을 가했을 때 방향은 보존되고 크기만 변경이 되는 방향 벡터를 나타내는게 eigenvecor이고 이 선형변환이 이루어져도 실제로 변환 전과 후의 axis가 변하지 않는 base라고 볼 수 있겠다. 그리고 그 크기의 정도를 라고 그것을 eigenvalue라고 할 수 있겠다. 이 2가지의 값을 통해서 우리는 고유값분해를 이용한 대각화(eigendecomposition)을 할 수 있다. 먼저 대각행렬과 행렬곱을 살펴보자. 대각행렬 뒤에 곱하면 행렬의 열벡터들이 대각원소의 크기만큼 상수배가 된다.

행렬 A의 고유값 고유벡터들을 , 그리고 i는 1,2,…,n이라고 하면,

이식을 정리하면,

행렬 A의 고유벡터를 열벡터로 하는 행렬을 P, 고유값들을 대각원소로 하는 대각행렬을 라 하면 다음 식이 성립이 된다.

이렇게 A라는 행렬을 3가지 고유벡터들을 열벡터로 하는 행렬과 고유값을 대각원소로 하는 행렬의 곱으로 대각화 분해가 가능한데 이를 eigencomposition이라고 하며, 이것을 하게 되면 det(A), A의 거듭제곱 , 역행렬, 대각합, 행렬의 다항식등을 매우 손쉽게 계산할 수 있다.

**6. Positive definite matrix, Positive semidefinite matrix**

대칭 행렬(대각을 기준으로 서로 대칭의 값을 가지는 행렬이라 Transpose해도 값이 같다)중에 속하면서 특수한 케이스인 Positive Definite Matrix 는 아주 유용하게 쓰이는 행렬이라고 한다. 이 행렬이 필요한 이유는 예를 들어 변수가 두개인 이차 함수를 기준으로 이를 행렬로 나타낼 수 있고 이 함수가 극소값을 가지는지 극대값을 구하는지 판단하기 위해서 이 행렬을 사용할수 있기 때문이다.

위의 식을 통해

양의 정부호일려면, 제곱항의 앞에 붙은 값들이 0보다 커야한다. 그래서 조건은 i) a > 0 ii) c > 이라는 조건이면 성립한다. 즉 이 된다. 음의 정부호일 경우에는 반대로 되는 경우이며, 이면, 은 안장이다.

이제 쉽게 양/음의 정부호 그리고 안장을 결정하기 위해 행렬의 식으로 표현을 하면 라고 나타낼 수 있고 A는 2x2대칭행렬이다.

위와 같이 행렬을 구할 수가 있고 여기에서 위의 A라는 매트릭스를 가지고 이 행렬이 양의 정부호인지 음의 정부호인지 아니면 안장인지를 결정한다. 여기에 a,b,c,d의 값을 가지고 위의 식을 대입해보면 구할 수 있다는 얘기이며 여기에서 사용한 대칭 행렬이 이면 양의 정부호 행렬이며 극소점을 가지는 케이스라고 생각할수 있다.

즉 정리하면 양의 정부호 행렬의 조건은

1. 영이 아닌 모든 실수 벡터 x에 대해 을 만족한다.
2. 행렬A의 모든 고유값들이 0 보다 크다.
3. 행렬 A의 모든 서브행렬등(상위 왼쪽 서브행렬들)의 행렬식들이 0보다 크다.
4. 행렬 A의 모든 피봇들이 0보다 크다
5. A = 인 독립 열들을 지닌 행렬 R이 존재한다.

그리고 여기에서 Positive semidefinite matrix은 Positive definite matrix와 동일하나 실수 벡터들이 0보다 크다는 제약보다는 좀 덜한 제약을 가진 matrix이며 음의 실수도 허용한다.