심층학습 - 3

생성자	㊊ 재환 김
태그	머신러닝

확률론과 정보 이론

- **확률론과 정보 이론의 개요**
  - 이 장에서는 확률론과 정보 이론을 다루며, 이 두 분야가 과학과 공학의 필수적인 도구임을 강조합니다. 확률론은 불확실한 상황에서 결과를 예측하고 결정을 내리는 데 중요한 역할을 합니다.
  - 확률론(probability theory)은 수학적 기반을 토대로 다양한 문제 해결에 활용됩니다.
- **소프트웨어 공학에서의 확률론**
  - 확률론은 소프트웨어 공학에서도 중요한 역할을 합니다. 소프트웨어 시스템의 안정성, 성능, 보안 등을 평가하고 예측하는 데 확률론적 방법론이 자주 사용됩니다. 예를 들어, 시스템 오류 발생 확률을 분석하여 위험을 줄이기 위한 조치를 설계할 수 있습니다.
- **확률론과 불확실성**
  - 확률론은 불확실한 상황에서 예측력을 제공하는 유용한 도구입니다. 완벽한 예측은 불가능하지만, 확률론을 통해 가능한 결과들의 범위와 가능성을 추정할 수 있습니다. 이로 인해 확률론은 불확실성을 효과적으로 다루는 방법론으로 자리 잡았습니다.
- **정보 이론과의 연관성**
  - 정보 이론(information theory)은 확률론의 개념을 확장하여 정보 전달 관련 문제를 다룹니다. 이는 통신 시스템에서 정보의 효율적 전송과 복구, 데이터 압축, 암호화 등에 적용됩니다.
  - 정보 이론의 주요 과제는 불확실한 상황에서 정보를 효율적으로 전달하는 방법을 고안하는 것입니다. 예를 들어, 잡음이 있는 통신 채널에서 정보를 정확하게 전달하는 방법을 설계할 때 확률론과 정보 이론이 결합하여 사용됩니다.
- **독자를 위한 학습 제안**

    - 이 장은 기술적인 내용을 포함하고 있어 이해하기 어려울 수 있습니다. 따라서 확률론과 정보 이론의 기본 개념을 구체적으로 다루어 독자의 이해를 돕고자 합니다.
    - 책에서 제공하는 추가 내용과 참고 자료는 독자의 연구와 프로젝트에 실질적인 도움이 되도록 구성되어 있습니다. [Daynes, 2003]과 같은 참고자료를 통해 더 깊이 있는 학습을 할 수 있습니다.
- **참고문헌의 중요성**
    - 추가 학습을 위해 관련 자료를 참고하는 것이 중요합니다. [Daynes, 2003]과 같은 자료는 확률론과 정보 이론에 대한 심층적 이해를 돕고, 독자의 연구와 학습 과정에 중요한 역할을 합니다.

1 확률의 필요성

- **확률의 필요성**
    - 이 섹션에서는 컴퓨터 과학 및 다양한 분야에서 확률론이 필수적인 이유를 설명합니다. 주로 확률론은 불확실성을 처리하는 데 유용하며, 완전히 결정론적인 방식으로 다룰 수 없는 문제들을 다룰 때 필요합니다.
    - 예를 들어, 일반적인 컴퓨터 프로그램에서 중앙처리장치(CPU)가 결정론적으로 동작하여 특정 결과를 예측할 수 있지만, 시스템 수준에서 발생하는 불확실성에 대해서는 확률론이 필요합니다. 특히, 컴퓨터 시스템 자체가 하드웨어의 미세한 결함이나 외부 환경 변화로 인해 비결정론적인 성격을 가지기도 하므로, 이러한 불확실성에 대한 처리가 필요합니다.
- **불확실성의 존재**
    - 대부분의 기계 학습 알고리즘, 특히 통계학이나 인공지능에서는 불확실성을 다루는 것이 필수적입니다. 이는 결정론적인 방법론이 모든 상황에서 적합하지 않기 때문입니다. 예를 들어, 하드웨어적인 결함이나 데이터 수집 과정에서의 오차 등이 시스템의 불확실성을 증가시키는 요인이 될 수 있습니다.
    - 기계 학습에서는 확률적(stochastic) 혹은 비결정론적(nondeterministic) 수치들을 다뤄야 할 필요성이 생기며, 이를 통해 보다 현실적인 모델을 구축할 수 있습니다.
- **불확실성의 예시**
    - 불확실성을 다루기 위해서는 수학적 명제를 제외할 때, 그러한 명제를 따르지 않는 경우 반드시 발생하는 것이 불확실성입니다.
    - 불확실성을 처리할 수 있는 주요 개념 중 하나는 모호성(ambiguity)으로, 모형화된 시스템에서 입력 값들이 명확하게 정의되지 않거나 여러 가능성을 가정할 수 있는

경우가 있습니다. 예를 들어, 주사위를 굴릴 때 결과를 완전히 예측할 수 없는 것처럼 불확실성을 다룰 수 있습니다.

- **불완전한 모형과 관측**
  - 불완전한 모형이나 관측 과정에서도 불확실성이 발생합니다. 예를 들어, 주변의 데이터를 측정하거나 예측할 때도, 불확실성이 내재되어 있으며, 이를 확률론적 방식으로 해결해야 하는 경우가 있습니다.
  - 이러한 상황에서는 **이산화(discretization)** 같은 방법을 사용해 문제를 해결할 수 있습니다. 이 과정에서는 연속적인 데이터를 이산적으로 나누어 표현하게 됩니다.
- **키워드와 예측 문제**
  - 예측 과정에서도 불확실성이 존재하며, 이를 처리하기 위해 많은 연구자들이 확률론을 기반으로 한 모델을 사용하고 있습니다. 예를 들어, 항공기 사고 예측에서 비행 경로를 완벽하게 예측하는 것은 불가능하지만, 확률적으로 예측해볼 수 있습니다.
  - 확률론은 이러한 상황에서 "가능성(probability)"이라는 개념을 도입하여 해결할 수 있습니다.
- **몬티 홀 문제와 확률**
  - 확률론의 중요한 사례 중 하나로 **몬티 홀 문제**가 언급됩니다. 이 문제에서는 참가자가 세 개의 문 중 하나를 선택하는 게임을 설명하며, 선택 후 다시 다른 문을 고르는 전략이 확률적으로 이득이 있다는 것을 보여줍니다. 이러한 문제는 단순해 보이지만, 확률론을 통해 해결해야만 그 본질적인 의미를 이해할 수 있습니다.

2 확률변수 (Random Variable)

- 확률변수는 **여러 가능한 값 중 무작위로 특정 값을 가지는 변수**입니다.
  - 확률변수는 불확실성에 따른 사건의 결과를 수치화한 것이라고 볼 수 있습니다.
  - 예를 들어, 주사위를 던지는 경우 확률변수 X는 1에서 6까지의 값을 가질 수 있습니다. 각 값은 주사위가 특정 면에 멈추는 결과를 나타냅니다.
- 확률변수는 이산형(discrete)과 연속형(continuous)으로 나뉩니다.
  - **이산형 확률변수**는 유한한 개수 혹은 셀 수 있는 개수의 값을 가집니다.
    - 예를 들어, 주사위의 눈금이나 동전 던지기의 결과가 이산형 확률변수에 해당됩니다.
  - **연속형 확률변수**는 연속적인 값을 가질 수 있으며, 이 값들은 연속적인 구간에 걸쳐 나타납니다.

        - 예를 들어, 사람의 키나 무게는 특정 범위 안에서 연속적인 값을 가질 수 있습니다.
- 확률변수의 값을 특정 확률에 따라 분포시키는 방법은 확률분포(probability distribution)에서 다루게 됩니다.

3 확률분포 (Probability Distribution)

- **확률분포**는 하나의 확률변수 또는 여러 확률변수의 값들이 특정 상태에서 어떻게 분포되는지를 수치적으로 나타냅니다.
    - 확률변수가 이산형인지 연속형인지에 따라 확률분포를 설정하는 방식이 다르며, 이에 따라 확률질량함수(PMF)와 확률밀도함수(PDF)로 나뉩니다.

3.1 이산 변수와 확률질량함수 (Probability Mass Function, PMF)

- **이산 확률변수**에 대한 확률분포를 나타낼 때 주로 확률질량함수(PMF)를 사용합니다.
    - PMF는 특정 이산 값이 나올 확률을 나타냅니다. 예를 들어, $P(X = x)$는 확률변수X가 특정 값 x를 가질 확률을 의미합니다.
- 주사위를 던졌을 때 나오는 결과를 예로 들면, X가 주사위의 눈금이라면, $P(X = 3)$는 주사위가 3이 나올 확률을 나타냅니다.
    - PMF는 이처럼 각 이산값에 대해 그 값이 발생할 확률을 제공합니다.
- 두 개 이상의 확률변수에 대한 결합 확률 분포는 결합확률분포(joint probability distribution)라고 하며, 이는 두 확률변수 사이의 관계를 나타냅니다.
    - 예를 들어, $P(X = x, Y = y)$는 두 확률변수 X와 Y가 각각 특정 값 x와 y를 가질 확률을 의미합니다.

3.2 연속 변수와 확률밀도함수 (Probability Density Function, PDF)

- 연속 변수(continuous variable)는 이산 변수와 달리 연속적인 값들을 가집니다. 이때, 확률분포를 설명하기 위해 확률밀도함수(PDF)를 사용합니다.
    - 연속 확률변수는 특정한 하나의 값에 대한 확률이 0이므로, 특정 구간에서의 확률을 계산하는 것이 주된 관심사입니다.
    - 예를 들어, $P(a \leq X \leq b)$는 연속 변수 X가 구간 $[a, b]$에서 값을 가질 확률을 나타냅니다.

확률밀도함수의 성질

- 확률밀도함수 $p(x)$는 연속 변수의 분포를 수학적으로 나타내는 함수입니다. 특정 구간에서 확률을 계산하기 위해서 이 함수를 적분합니다.
  - 예를 들어, 구간 $[a, b]$에서 확률변수 X가 나타날 확률은 $\int_a^b p(x)dx$로 표현됩니다.
- 확률밀도함수는 다음의 조건을 만족해야 합니다:
  1. **정의역**: $p(x)$는 x의 가능한 모든 상태에 대해 정의되어야 합니다.
  2. **비음수 조건**: $p(x) \geq 0$이어야 합니다. 이는 확률이 음수가 될 수 없다는 것을 의미합니다.
  3. **정규화 조건**: 전체 구간에서 확률이 1이 되어야 합니다.

     즉, $\int_{-\infty}^{\infty} p(x)dx = 1$이어야 합니다.

     이 조건은 모든 가능성을 합하면 반드시 1이어야 한다는 것을 나타냅니다.

예시: 균등분포 (Uniform Distribution)

- 균등분포는 특정 구간 내에서 모든 값이 동일한 확률로 나타나는 분포를 의미합니다.
  - 예를 들어, 구간 $[a, b]$에서 균등분포를 따른다면, 확률밀도함수 $p(x)$는 $\frac{1}{b-a}$로 주어집니다.
  - 이 경우, $p(x)$는 구간 $[a, b]$ 외부에서는 0이 되고, 내부에서는 일정한 값을 가집니다. 이를 수학적으로 표현하면 $p(x) = \frac{1}{b-a}$로 나타낼 수 있습니다.

3.4 주변확률 (Marginal Probability)

- **주변확률**은 여러 변수가 있을 때, 그 중 일부 변수에 대한 확률분포를 구하는 방법입니다. 이는 전체 확률분포에서 일부 변수를 통합하여, 관심 있는 특정 변수에 대한 확률을 얻는 과정입니다.

수학적 정의

- 예를 들어, 두 변수 X와 Y에 대한 결합확률분포 $P(X, Y)$가 주어져 있을 때, 변수 X의 주변확률$P(X)$는 다음과 같이 계산됩니다:

$$P(X = x) = \sum_y P(X = x, Y = y)$$

- 이때, Y에 대한 모든 값을 합산하여 X의 확률만 남깁니다. 이를 주변화(marginalization)라고 부릅니다.

주변확률의 예시

- 주사위 두 개를 굴려서 각각의 눈금을 X와 Y라고 할 때, 이 두 변수의 결합확률은 $P(X = x, Y = y)$로 나타낼 수 있습니다. 이제 X의 주변확률을 구하려면, Y의 모든 가능한 값에 대해 합산하여 $P(X = x)$를 얻습니다.

활용 예시

- 주변확률은 특히 다변수 확률분포에서 중요한 역할을 합니다. 예를 들어, 다차원 데이터의 분석에서 관심 있는 변수에 대한 확률을 계산할 때, 나머지 변수들을 제거하고 특정 변수에 대한 확률분포를 얻는 데 사용됩니다.

3.5 조건부 확률 (Conditional Probability)

- **조건부 확률**은 어떤 사건이 발생했을 때, 다른 사건이 발생할 확률을 구하는 방법입니다. 이는 사건 A가 이미 일어났다는 정보를 바탕으로 사건 B가 일어날 확률을 의미합니다.
  - 예를 들어, $P(A \mid B)$는 사건 B가 발생한 조건에서 사건 A가 발생할 확률을 나타냅니다.
- 조건부 확률은 다음과 같은 수식으로 계산됩니다:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

  - 여기서 $P(A \cap B)$는 A와 B가 동시에 발생할 확률이고, P(B)는 사건 B가 발생할 확률입니다.
  - 조건부 확률은 $P(B) > 0$일 때만 정의되며, B가 일어나지 않았다면 A가 발생할 조건부 확률은 계산할 수 없습니다.
- 조건부 확률은 특히 인과관계를 분석하는 데 사용되며, 상황에 따라 사건 간의 종속성 또는 독립성을 평가하는 중요한 도구입니다.

3.6 조건부 확률의 연쇄법칙 (Chain Rule of Probability)

- **연쇄법칙**은 여러 사건에 대한 조건부 확률을 순차적으로 계산하는 방법을 설명합니다. 이는 곱의 법칙(product rule)으로도 불립니다.
  - 예를 들어, 세 사건 A, B, C에 대한 조건부 확률을 연쇄적으로 계산할 수 있습니다:

$$P(A, B, C) = P(A|B, C) \cdot P(B|C) \cdot P(C)$$

  - 각 사건이 다음 사건에 조건부로 의존할 때, 이를 곱의 형태로 나타내는 것입니다.
- 연쇄법칙은 복잡한 확률 계산에서 여러 사건이 차례로 조건부 독립일 때 매우 유용합니다.

3.7 독립과 조건부 독립 (Independence and Conditional Independence)

- 독립(Independence)은 두 사건이 서로 영향을 미치지 않는 경우를 말합니다.
  - 두 사건 X와 Y가 독립이라면 $P(X, Y) = P(X) \cdot P(Y)$로 표현됩니다.
  - 즉, 하나의 사건이 발생했을 때, 다른 사건의 발생 확률에 아무런 영향을 미치지 않는 경우를 의미합니다.
- 조건부 독립(Conditional Independence)은 세 사건이 있을 때, 한 사건의 발생이 다른 두 사건 간의 상관관계에 영향을 미치지 않는 경우를 말합니다.
  - 예를 들어, X와 Y가 Z에 대해 조건부 독립이라면, $P(X, Y \mid Z) = P(X \mid Z) \cdot P(Y \mid Z)$가 성립합니다.
  - 이는Z가 주어진 상황에서 X와 Y가 독립적으로 발생함을 의미합니다.

3.8 기대값, 분산, 공분산 (Expectation, Variance, Covariance)

- 기대값(Expectation)은 확률변수의 평균적인 값을 나타내며, 확률분포의 중심을 의미합니다. 이산 변수의 경우, 기대값은 다음과 같이 계산됩니다:

$$E[X] = \sum_i x_i P(x_i)$$

  - 이는 각 확률변수 값에 그 값이 발생할 확률을 곱한 후, 그들의 합을 구하는 방식입니다.
- 분산(Variance)은 확률변수 값들이 기대값을 중심으로 얼마나 퍼져 있는지를 나타냅니다. 이는 확률변수의 변동성을 나타내는 중요한 지표입니다.

$$\mathrm{Var}(X) = E[(X - E[X])^2]$$

  - 분산이 클수록 확률변수 값들이 넓게 퍼져 있음을 의미합니다.
- 공분산(Covariance)은 두 확률변수 간의 상관관계를 나타내는 척도입니다. 공분산이 양수면 두 변수가 같은 방향으로 변동하며, 음수면 반대 방향으로 변동합니다.

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

3.8 기대값 (Expectation)

- **기대값**은 확률변수의 평균적인 값을 나타내며, 확률분포에서 중요한 개념입니다.
  - **이산 변수**의 경우, 기대값은 다음과 같이 계산됩니다:

$$E[X] = \sum x_i P(x_i)$$

        - 각 값 x_i에 그 값이 발생할 확률 $P(x_i)$를 곱한 후, 이들의 합을 구하는 방식입니다.
    - **연속 변수**의 경우, 기대값은 적분을 이용하여 계산됩니다:

$$E[X] = \int_{-\infty}^{\infty} xp(x)dx$$

        - $p(x)$는 확률밀도함수로, 확률변수가 구간 $[a, b]$에서 어느 값에 집중되는지를 나타냅니다.
- 기대값의 성질로는 **선형성**이 있습니다. 즉, 두 함수의 선형 결합에 대한 기대값은 다음과 같이 계산됩니다:

$$E[aX + bY] = aE[X] + bE[Y]$$

    - 이는 확률변수들의 결합을 고려할 때 중요한 성질입니다.

3.8 분산 (Variance)

- **분산**은 확률변수 값들이 기대값을 중심으로 얼마나 퍼져 있는지를 나타냅니다. 이는 변동성을 측정하는 지표로 사용됩니다.
    - 분산은 다음과 같이 정의됩니다:

$$\text{Var}(X) = E[(X - E[X])^2]$$

    - 이는 확률변수 X의 값들이 평균에서 얼마나 떨어져 있는지를 제곱하여 측정한 후, 그 평균을 구하는 방식입니다.
- 분산이 작을수록 확률변수의 값들이 평균에 가까이 모여 있음을 의미하고, 분산이 클수록 값들이 평균에서 멀리 퍼져 있음을 나타냅니다.

3.8 공분산 (Covariance)

- **공분산**은 두 확률변수 간의 관계를 나타냅니다. 두 변수가 같은 방향으로 변동하는지, 반대 방향으로 변동하는지를 측정합니다.
    - 공분산은 다음과 같이 정의됩니다:

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

    - 공분산이 양수이면 두 변수가 같은 방향으로 움직이며, 음수이면 반대 방향으로 움직입니다. 공분산이 0이면 두 변수 간에는 상관관계가 없음을 의미합니다.
- **상관계수**는 공분산을 표준화한 값으로, 두 변수 간의 선형 관계를 더욱 명확하게 보여줍니다.

공분산 행렬 (Covariance Matrix)

- 다변수 확률변수의 경우, 각 변수들 간의 공분산을 **공분산 행렬**로 나타냅니다. 공분산 행렬은 대칭 행렬로, 다음과 같은 성질을 가집니다:

  $$\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$$

  - 대각선 성분은 각 변수의 분산을 나타내고, 나머지 성분은 각 변수 간의 공분산을 나타냅니다.

3.9 흔히 쓰이는 확률분포들

이 섹션에서는 확률론에서 자주 사용되는 몇 가지 분포 중 **베르누이 분포**를 설명합니다.

3.9.1 베르누이 분포 (Bernoulli Distribution)

- **베르누이 분포**는 이진 사건(성공 또는 실패)에서 사용되는 가장 기본적인 확률분포입니다. 확률변수 X가 성공 확률 ϕ로 성공(1)을, 실패 확률$1 - \phi$로 실패(0)를 가질 때 이를 베르누이 분포라고 합니다.
  - 베르누이 분포는 다음과 같이 정의됩니다:

    $$P(X = 1) = \phi, \quad P(X = 0) = 1 - \phi$$

  - 예를 들어, 동전을 던져 앞면이 나올 확률이 $\phi = 0.5$라면, X는 1 또는 0의 값을 가질 수 있으며, 각 사건의 확률은 각각 0.5입니다.
- 베르누이 분포는 단일 시도에 대한 확률을 설명할 때 사용되며, 이를 확장하면 이항 분포(binomial distribution)로 이어집니다.

3.9.2 멀티누이 분포 (Multinomial Distribution)

- **멀티누이 분포**는 여러 범주 중 하나를 선택할 때 나타나는 분포로, 범주형 분포(categorical distribution)의 확장판이라고 볼 수 있습니다. 즉, k개의 범주 중 하나의 범주에 해당할 확률을 나타냅니다.
  - 멀티누이 분포는 각 범주가 선택될 확률이 p로 주어질 때, $p = [p_1, p_2, ..., p_k]$로 표현됩니다. 이때 각 확률의 합은 1이 되어야 하며, $\sum_{i=1}^{k} p_i = 1$을 만족해야 합니다.
  - 예를 들어, 주사위를 10번 던졌을 때 각각의 면이 나오는 횟수에 대한 분포를 나타낼 때 멀티누이 분포를 사용할 수 있습니다. 각 면이 나올 확률은 $\frac{1}{6}$이므로, 6개의 범주가 있고, 그 각각의 확률이 동일한 멀티누이 분포를 따르게 됩니다.

- **멀티누이 분포의 사용 사례**는 주로 분류 문제에서 많이 사용됩니다. 특히, 카테고리 데이터를 다룰 때 주로 사용됩니다. 예를 들어, 자연어 처리에서 단어의 출현 빈도와 같은 데이터를 다룰 때 멀티누이 분포가 적용될 수 있습니다.
- 멀티누이 분포의 특징은 다항분포(multinomial distribution)와 밀접하게 관련이 있습니다. 다항분포는 여러 범주에 속하는 데이터를 다룰 때 사용되며, n번의 시도에서 각각의 범주에 해당하는 횟수를 나타냅니다.

3.9.3 가우스 분포 (Gaussian Distribution)

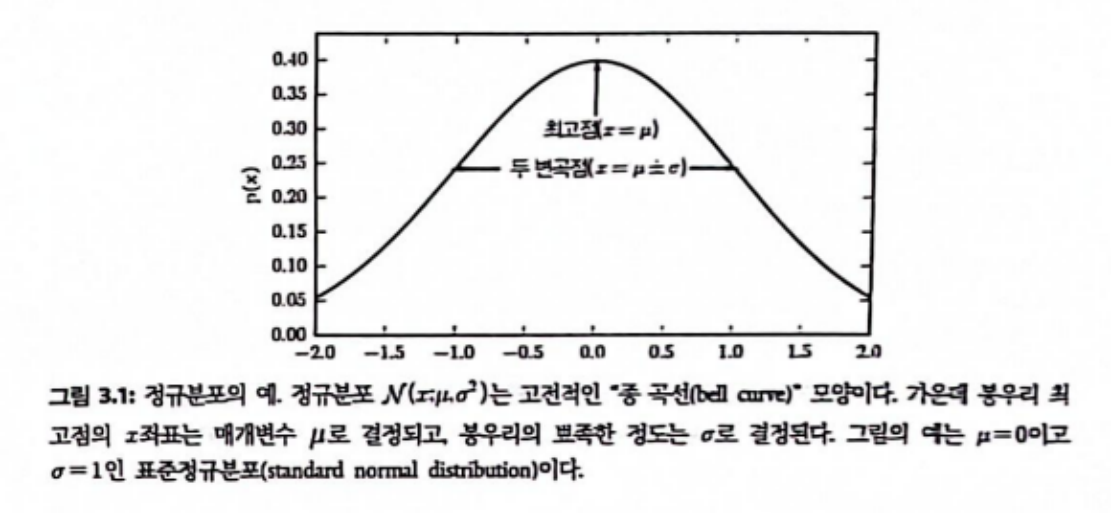

그림 3.1: 정규분포의 예. 정규분포 $N(x;\mu,\sigma^2)$는 고전적인 "종 곡선(bell curve)" 모양이다. 가운데 봉우리 최고점의 x좌표는 매개변수 μ로 결정되고, 봉우리의 뾰족한 정도는 σ로 결정된다. 그림의 예는 $\mu=0$이고 $\sigma=1$인 표준정규분포(standard normal distribution)이다.

- **가우스 분포**, 또는 정규분포(normal distribution)는 확률론과 통계학에서 가장 많이 사용되는 분포 중 하나입니다. 실수 값들을 다룰 때 그 값들이 평균 주변에 어떻게 분포되는지를 나타냅니다.
  - 정규분포는 다음과 같은 확률밀도함수로 정의됩니다:

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

    - 여기서 μ는 평균(mean), $\sigma 2$는 분산(variance)을 나타냅니다. 평균은 데이터가 집중되는 중심 위치를 나타내며, 분산은 데이터가 얼마나 퍼져 있는지를 나타냅니다.
- **정규분포의 특징**은 종형 곡선을 그리며, 대칭적인 형태를 가지고 있습니다. 이때, 평균 주변에 데이터가 많이 몰려 있으며, 멀어질수록 그 확률이 점점 낮아집니다.
  - **표준정규분포**는 평균이 0이고 분산이 1인 정규분포를 의미하며, 이를 기준으로 다른 정규분포들도 변환할 수 있습니다.
- **가우스 분포의 사용 사례**는 매우 다양합니다.

    - 실세계에서 발생하는 많은 현상들이 정규분포를 따르기 때문에, 이는 모델링과 예측에 자주 사용됩니다. 예를 들어, 사람의 키, 시험 점수, 제품의 수명 등은 정규분포를 따르는 경향이 있습니다.
    - 또한, 중심극한정리(Central Limit Theorem)에 따라, 여러 개의 독립적인 변수들의 합은 정규분포에 가까워지므로, 실제로 많은 데이터가 정규분포를 따른다고 가정할 수 있습니다.
- **정규분포의 활용**은 기계 학습에서도 매우 중요합니다. 특히, 최소 제곱법(least squares)을 사용할 때나 회귀 분석에서 오류가 정규분포를 따른다는 가정 하에 진행됩니다. 이는 기계 학습의 성능을 평가하고 모델을 설명하는 데 유용한 분포입니다.

3.9.4 지수분포와 라플라스 분포

지수분포 (Exponential Distribution)

- **지수분포**는 주로 사건이 발생하는 **시간 간격**을 모델링할 때 사용되는 분포입니다. 예를 들어, 전화가 걸려오는 시간 간격, 버스가 도착하는 시간 간격 등의 데이터를 설명할 때 자주 사용됩니다.
    - 지수분포의 확률밀도함수(PDF)는 다음과 같이 정의됩니다:

      $$p(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

        - 여기서 λ는 분포의 속도 매개변수(rate parameter)로, 사건이 발생할 빈도를 조절하는 역할을 합니다. x는 사건이 발생할 때까지의 시간을 나타냅니다.
- **지수분포의 특징**:
    - 사건 간의 간격이 기억이 없다는 기억 없음의 성질(memoryless property)을 가집니다. 이는 사건이 발생하는 시간이 얼마나 지났든 앞으로 발생할 사건의 시간은 동일한 분포를 따른다는 의미입니다.
    - 예를 들어, 버스가 오기까지 기다린 시간이 10분이 지났더라도, 앞으로 5분 더 기다려야 할 확률은 처음부터 기다리기 시작했을 때와 동일합니다.

라플라스 분포 (Laplace Distribution)

- **라플라스 분포**는 **두 모드**를 가지는 분포로, **대칭적인 모양**을 띄는 확률분포입니다. 이는 주로 데이터의 중심점에서 급격한 변화가 있는 경우에 사용됩니다.
    - 라플라스 분포는 다음과 같은 확률밀도함수로 정의됩니다:

      $$p(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

        - 여기서 μ는 **중심**을, b는 척도(scale)를 나타냅니다.
- **라플라스 분포의 특징**:
    - 데이터가 중심 근처에 몰려 있고, 중심에서 멀어질수록 확률이 급격히 감소하는 분포입니다.
    - 라플라스 분포는 **평활한 변화**를 모델링할 때 주로 사용됩니다. 예를 들어, 신호 처리에서 급격한 변화를 포착하는 데 유용합니다.

3.9.5 디락 분포와 경험분포

디락 델타 분포 (Dirac Delta Function)

- **디락 델타 함수**는 특정한 지점에서만 값을 가지는 특이한 함수입니다. 물리학 및 공학에서 **단위 질량이 특정한 점에 집중**되어 있는 현상을 설명할 때 사용됩니다.
    - 디락 델타 함수는 $x = a$에서 값이 무한대가 되며, 그 외의 지점에서는 0이 되는 함수로 정의됩니다. 이를 수식으로 표현하면 다음과 같습니다:

      $$p(x) = \delta(x - a)$$

        - 이는 $x = a$일 때 함수값이 무한대로 발산하는 대신, 그 값이 무한히 작은 영역에서 1을 가지게 하여 전체 면적이 1이 되도록 조정됩니다.
- 디락 델타 함수는 확률밀도함수(PDF)로 직접 사용되지는 않지만, 특정 값에서 집중되는 확률을 설명할 때 매우 유용합니다. 예를 들어, 아주 짧은 시간 동안 발생하는 충격을 모델링할 때 사용됩니다.

경험분포 (Empirical Distribution)

- **경험분포**는 데이터의 관측값에 따라 확률을 할당하는 분포입니다. 즉, **실제로 관측된 데이터**를 기반으로 만들어진 분포로, 데이터를 직접 반영하는 분포입니다.
    - 경험분포는 다음과 같이 정의됩니다:

      $$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$$

        - 여기서 x_i는 데이터의 관측값이고, n은 데이터의 총 개수를 나타냅니다.
- 경험분포는 **이론적인 분포가 없는 경우** 또는 **데이터 자체를 기반으로 통계를 계산**하고자 할 때 사용됩니다. 데이터의 개별 관측값에 모두 집중된 확률을 할당하기 때문에, 데이터가 추가될수록 점진적으로 개선될 수 있습니다.

3.9.6 분포의 혼합 (Mixture of Distributions)

- **혼합 분포**는 여러 개의 단순한 분포들을 결합하여 더 복잡한 분포를 만드는 방법입니다. 이는 각 단순 분포를 성분(component)이라고 하며, 각 성분이 전체 분포에 일정한 비율로 기여하는 방식입니다.
- **혼합 분포의 정의**는 다음과 같은 수식으로 표현됩니다:

$$P(x) = \sum_i P(c = i)P(x|c = i)$$

  - 여기서 $P(c = i)$는 성분 i에 대한 가중치를 나타내며, 각 성분 분포의 비율입니다. 각 성분 분포 $P(x \mid c = i)$는 성분 i가 선택되었을 때의 확률분포입니다.
  - 이를 통해 여러 성분 분포가 서로 다른 가중치를 가지며 결합되어, 하나의 복잡한 분포를 형성하게 됩니다.

혼합 분포의 예

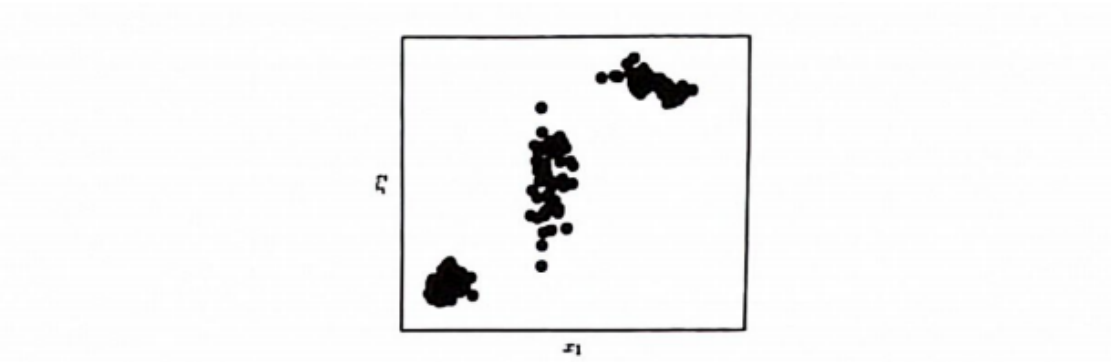

그림 3.2: 가우스 혼합모형의 표본들. 이 예에서 혼합분포의 성분은 세 개이다. 그림의 왼쪽 성분은 공분산행렬이 등방성이다. 이는 각 방향의 분산이 같음을 뜻한다. 가운데 성분의 공분산행렬은 대각행렬이다. 이는 각 축 방향의 분산을 따로 제어할 수 있음을 뜻한다. 그림의 예에서는 x_2축 방향의 분산이 x_1축 방향의 분산보다 크다. 오른쪽 성분의 공분산행렬은 최대 계수(full rank; 또는 완전 위수) 행렬이다. 따라서 임의의 방향의 기저에 따라 분산을 따로 제어할 수 있다.

- 가우스 혼합 모델(Gaussian Mixture Model, GMM)은 가장 많이 사용되는 혼합 분포 중 하나입니다. 이는 여러 개의 가우스 분포를 결합하여 더 복잡한 분포를 만드는 방법입니다.
  - 예를 들어, 여러 개의 군집(cluster)을 가지고 있는 데이터를 분석할 때, 각각의 군집을 가우스 분포로 모델링할 수 있습니다. 그리고 이 가우스 분포들을 혼합하여 전체 데이터를 설명하는 복합적인 분포를 만들 수 있습니다.
- **혼합 분포의 활용**:
  - 혼합 분포는 주로 **군집화(clustering)** 문제에서 사용됩니다. 데이터가 여러 개의 하위 집단(군집)으로 나누어질 때, 각 군집을 개별 분포로 모델링하고, 그 분포들을 혼합하여 전체 데이터를 설명할 수 있습니다.

    - 또한, 숨겨진 변수(latent variable)가 존재하는 경우, 이 변수는 직접적으로 관찰되지 않지만 데이터의 생성 과정에 영향을 미칠 수 있습니다. 이러한 숨겨진 변수를 설명하기 위해 혼합 분포가 사용되며, 이는 잠재 변수 모델(latent variable model)이라고도 불립니다.

- **잠재 변수와 EM 알고리즘**:
    - 혼합 분포에서 잠재 변수는 보이지 않는 상태에서 데이터를 생성하는 과정에서 영향을 미칩니다. 이러한 잠재 변수를 추정하기 위해 EM 알고리즘(Expectation-Maximization Algorithm)이 자주 사용됩니다.
    - EM 알고리즘은 먼저 데이터를 기반으로 잠재 변수를 추정한 후, 그 추정 값을 기반으로 분포의 매개변수를 업데이트하는 방식으로 작동합니다. 이를 반복적으로 수행하여 최종적으로 혼합 분포의 매개변수를 최적화합니다.

혼합 분포의 장점과 단점

- **장점**:
    - 여러 개의 단순한 분포를 결합하여 복잡한 데이터 구조를 모델링할 수 있습니다.
    - 다양한 패턴이 있는 데이터를 효율적으로 설명할 수 있으며, 군집화, 이상치 탐지 등 여러 분야에서 활용될 수 있습니다.
- **단점**:
    - 혼합 분포를 정확하게 학습하려면 많은 데이터가 필요하며, 계산 비용이 많이 들 수 있습니다.
    - 성분 분포의 개수와 초기 값 설정에 민감할 수 있습니다. 즉, 성분의 수를 적절하게 설정하지 않으면 학습 성능이 저하될 수 있습니다.

3.10 흔히 쓰이는 함수들의 유용한 성질들

3.10.1. 로지스틱 함수 (Logistic Sigmoid Function)

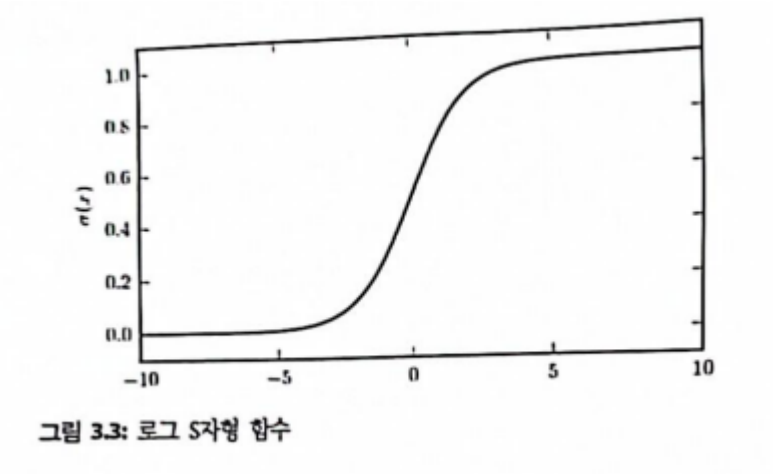

그림 3.3: 로그 S자형 함수

- **로지스틱 함수**는 주로 분류 문제에서 이진 분류에 사용되는 함수로, 확률적인 결과를 출력할 때 많이 사용됩니다. 이 함수는 입력 값을 0과 1 사이의 값으로 변환합니다.
  - 로지스틱 함수의 수식은 다음과 같습니다:

    $$\sigma(x) = \frac{1}{1+\exp(-x)}$$
  - 이 함수는 x가 매우 큰 값이면 출력이 1에 가까워지고, x가 매우 작은 값이면 출력이 0에 가까워집니다. 중간 값인 $x = 0$일 때는 0.5를 출력합니다.
- **로지스틱 함수의 성질**:
  - **매우 부드러운 형태**를 가지며, 급격한 변화 없이 완만하게 0에서 1로 변환합니다.
  - **미분 가능성**: 로지스틱 함수는 미분 가능하며, 그 미분 값은 다음과 같습니다:

    $$\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x))$$
  - 이 성질은 기계 학습에서 역전파(backpropagation)를 사용하여 파라미터를 업데이트할 때 유용하게 사용됩니다.

3.10.2. 소프트플러스 함수 (Softplus Function)

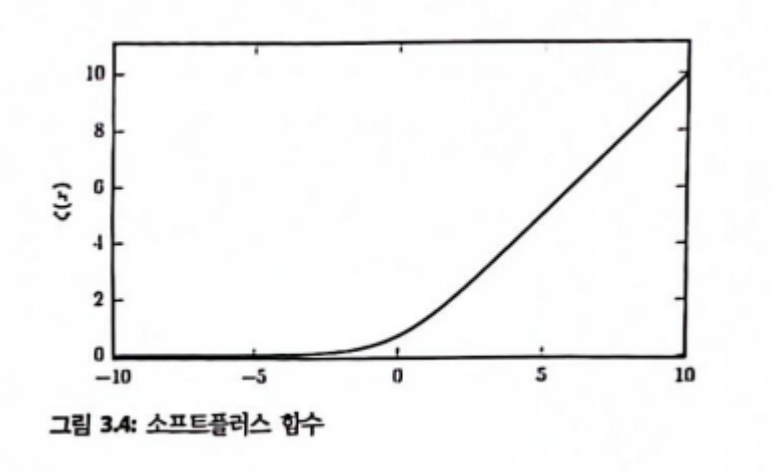

그림 3.4: 소프트플러스 함수

- **소프트플러스 함수**는 로지스틱 함수와 유사한 성질을 가지고 있지만, 이 함수는 더 부드럽게 증가하는 형태를 가집니다. 이는 주로 **ReLU 함수**의 매끄러운 버전으로 사용됩니다.
  - 소프트플러스 함수는 다음과 같은 수식으로 정의됩니다:

    $$\zeta(x) = \log(1 + \exp(x))$$

  - 이 함수는 입력 값이 매우 클 때는 선형적으로 증가하며, 작을 때는 0에 가까워집니다.
- **소프트플러스 함수의 성질**:
  - **미분 가능성**: 소프트플러스 함수는 모든 구간에서 미분 가능하며, 그 미분 값은 다음과 같습니다:

    $$\frac{d}{dx}\zeta(x) = \sigma(x)$$

    - 여기서 $\sigma(x)$는 로지스틱 함수입니다. 즉, 소프트플러스 함수의 미분은 로지스틱 함수와 동일합니다.
  - **ReLU와의 관계**: 소프트플러스 함수는 ReLU 함수의 매끄러운 대체 함수로 사용되며, 신경망에서 비선형성을 도입하는 데 사용됩니다. ReLU는 $x > 0$일 때 x를 출력하고, $x \leq 0$일 때는 0을 출력하는 반면, 소프트플러스는 더 부드럽게 동작하여 과도한 비선형성을 방지합니다.

3.10.3. 로지 함수 (Log Function)

- **로지 함수**는 로지스틱 함수의 역함수로, **로지스틱 변환**을 설명하는 데 사용됩니다. 이 함수는 확률값을 입력으로 받아 로그 오즈(log odds)를 반환합니다.

    - 로지 함수의 정의는 다음과 같습니다:

      $$\log\left(\frac{x}{1-x}\right)$$

    - 여기서 x는 (0,1) 사이의 확률값입니다.

- **로지 함수의 성질**:
    - 확률값을 오즈 비율로 변환하여, 선형 회귀와 같은 선형 모델에서 사용될 수 있습니다. 이는 로지스틱 회귀에서 매우 중요한 함수입니다.

3.10.4. 소프트맥스 함수 (Softmax Function)

- **소프트맥스 함수**는 다중 클래스 분류에서 사용되며, 여러 클래스 중 각 클래스에 속할 확률을 계산하는 데 사용됩니다.
    - 소프트맥스 함수의 수식은 다음과 같습니다:

      $$\sigma(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

    - 이는 입력된 각 값 z_i를 0과 1 사이의 확률로 변환하고, 모든 확률의 합이 1이 되도록 정규화하는 역할을 합니다.

- **소프트맥스 함수의 성질**:
    - **확률 분포**: 소프트맥스 함수는 입력된 값들을 기반으로 확률 분포를 생성합니다. 즉, 각각의 클래스가 선택될 확률을 계산할 수 있습니다.
    - **미분 가능성**: 소프트맥스 함수 역시 미분 가능하며, 역전파 과정에서 사용됩니다.

3.11 베이즈 법칙 (Bayes' Theorem)

- **베이즈 법칙**은 조건부 확률을 구하는 데 매우 중요한 역할을 합니다. 특정 사건이 발생했을 때, 그 사건에 대한 추가적인 정보를 바탕으로 확률을 갱신하는 방법입니다.
    - 베이즈 법칙은 다음과 같이 정의됩니다:

      $$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

        - 여기서$P(y \mid x)$는 사건 x가 발생했을 때, 사건 y가 일어날 확률을 나타냅니다.
        - $P(x \mid y)$는 사건 y가 일어난 조건에서 x가 발생할 확률을 의미합니다.
        - $P(y)$는 사건 y가 발생할 사전 확률(prior probability)이고, $P(x)$는 주변 확률(marginal probability)로 사건 x가 발생할 전체 확률을 나타냅니다.

- 베이즈 법칙은 사전 지식을 바탕으로 사건에 대한 정보를 갱신하는 데 사용됩니다. 이는 통계적 추론에서 매우 중요한 역할을 하며, 특히 기계 학습과 데이터 분석에서 사용되는 **베이즈 추론**의 핵심이 됩니다.
- **베이즈 법칙의 응용**:
  - 베이즈 법칙은 특히 **분류 문제**에서 자주 사용됩니다. 예를 들어, 스팸 이메일 분류기에서 특정 단어가 포함된 이메일이 스팸일 확률을 계산하는 데 사용될 수 있습니다. 이때 사전 확률과 단어 등장 확률을 이용해 이메일이 스팸인지 여부를 추론할 수 있습니다.
  - **베이즈 네트워크**와 같은 확률적 그래픽 모델에서 각 노드 간의 확률 관계를 모델링하는 데도 베이즈 법칙이 활용됩니다.

3.12 연속 변수의 특별한 세부 사항

- **연속 확률변수**는 이산 변수와 달리 무한히 많은 값 중 하나를 가질 수 있습니다. 이러한 연속 변수를 다루기 위해서는 측도론(measure theory)의 개념이 필요합니다.
  - 측도론은 수학에서 **길이, 면적, 부피** 등의 개념을 일반화하여 특정 구간에서 확률을 계산하는 방법을 제공합니다.
- **측도론과 확률론**:
  - 확률론에서 연속 변수의 확률을 계산하려면 특정 구간에서 확률밀도함수(PDF)를 적분하여 해당 구간의 확률을 구해야 합니다. 측도론은 각 구간에서 확률을 정의하는 기초적인 개념을 제공합니다.
  - 예를 들어, **르베그 측도(Lebesgue measure)**를 이용하여 사건 발생 확률을 정의할 수 있습니다. 이는 특정 구간에서 사건이 일어날 확률을 더 일반적으로 다룰 수 있는 방법입니다.
- **속도와 연속성의 문제**:
  - 연속 변수에서 변화율(rate of change)은 매우 중요한 개념입니다. 특히, 특정 구간에서 변수의 값이 어떻게 변화하는지 측정하는 과정에서 이 변화율에 대한 정보는 중요한 통계적 결과를 도출하는 데 기여합니다.
  - 이는 물리학에서 속도를 정의하는 개념과 유사하며, 연속적인 데이터 분석에도 비슷하게 적용됩니다.
- **특정 사건들의 확률 계산**:
  - 연속 확률변수에서 특정 사건이 발생할 확률은 정확히 0이 되는 경우가 많습니다. 예를 들어, **연속 확률변수**에서 특정한 하나의 값을 가질 확률은 0입니다. 대신, 특정 구간에서의 확률을 계산하는 것이 의미 있습니다.

- 이러한 이유로, 특정 점에서의 사건 발생 확률이 아닌 구간에서의 발생 확률을 다루기 위해 **확률밀도함수**와 **적분**이 중요한 역할을 합니다.

3.13 정보 이론 (Information Theory)

- **정보 이론**은 주로 통신 분야에서 발생하는 정보의 양을 수학적으로 다루는 이론입니다. 신호가 전달될 때 그 신호에 포함된 정보를 어떻게 효율적으로 표현할 수 있을지, 또는 불확실성 속에서 정보를 어떻게 처리할 수 있을지를 연구합니다.
  - 예를 들어, 컴퓨터 과학, 물리학, 통신 공학 등 다양한 분야에서 정보 이론이 활용되며, 데이터 압축, 암호화, 통신 시스템 최적화 등에 응용됩니다.
- 정보 이론의 핵심 개념 중 하나는 자기 정보(self-information)입니다. 특정 사건 x가 발생했을 때 그 사건에 포함된 정보의 양을 측정하는 방법으로, 정보의 양은 사건의 발생 가능성에 반비례합니다. 즉, 발생하기 어려운 사건일수록 더 많은 정보를 포함하고 있습니다.
  - **자기 정보의 정의**는 다음과 같습니다:

    $$I(x) = -\log P(x)$$

    - 여기서 $P(x)$는 사건 x가 발생할 확률이며, 이 확률이 작을수록 정보의 양은 커집니다. 이는 사건이 얼마나 예측 불가능한지에 따라 그 사건이 전달하는 정보의 양이 결정된다는 것을 의미합니다.
- **정보량과 확률의 관계**:
  - 사건이 발생할 가능성이 낮을수록 그 사건에 대한 정보량은 커지며, 자주 발생하는 사건은 정보량이 적습니다. 예를 들어, 일상적인 대화에서 자주 쓰이는 단어는 큰 정보량을 전달하지 않지만, 드물게 쓰이는 단어는 상대적으로 더 많은 정보를 전달합니다.
- **정보 이론의 응용**:
  - 정보 이론은 **데이터 압축** 및 **통신 시스템**에서 매우 중요한 역할을 합니다. 데이터 압축에서는 정보를 잃지 않으면서 데이터를 얼마나 효율적으로 표현할 수 있을지를 고민합니다. 통신 시스템에서는 정보를 전달할 때 발생하는 잡음(noise)을 최소화하고, 정확하게 데이터를 전송하는 방법을 연구합니다.
  - 또한, 기계 학습에서도 정보 이론이 중요한 역할을 합니다. 예를 들어, 교차 엔트로피(cross-entropy)는 분류 문제에서 모델이 예측한 확률 분포와 실제 정답 간의 차이를 측정하는 지표로, 정보 이론의 개념에서 유래한 것입니다.

확률 분포와의 수학적 관계

이미지의 왼쪽 부분에는 특정 함수에 대한 확률 분포와 관련된 수학적 변환 및 계산 방법이 나와 있습니다. 이 수학적 관계들은 정보 이론과 관련된 다양한 계산을 위한 기초가 됩니다.

확률밀도 함수 변환

- 확률밀도 함수 $p_y(y)$와$p_x(x)$의 관계를 나타내는 수식은 함수 g의 변환을 고려하는 경우에 사용됩니다.

$$p_y(y) = p_x(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$$

  - 이 식은 x에서 y로의 변환이 이루어질 때 확률밀도 함수가 어떻게 변화하는지를 설명합니다. 이를 통해, 변수 변환 후에도 확률 분포의 형태를 유지하거나 추정할 수 있습니다.

- **야코비안 행렬(Jacobian matrix)**:
  - 다변수 함수의 변환에서 사용되는 행렬로, 각 변수의 변화에 따른 함수의 변화율을 나타냅니다. 이는 고차원 데이터 변환에서 핵심적인 역할을 합니다.
  - 특히 확률밀도 함수 변환 시, 야코비안 행렬의 행렬식(determinant)을 활용하여 변수 변환 후의 새로운 확률밀도 함수를 도출할 수 있습니다.

샤논 엔트로피 (Shannon Entropy)

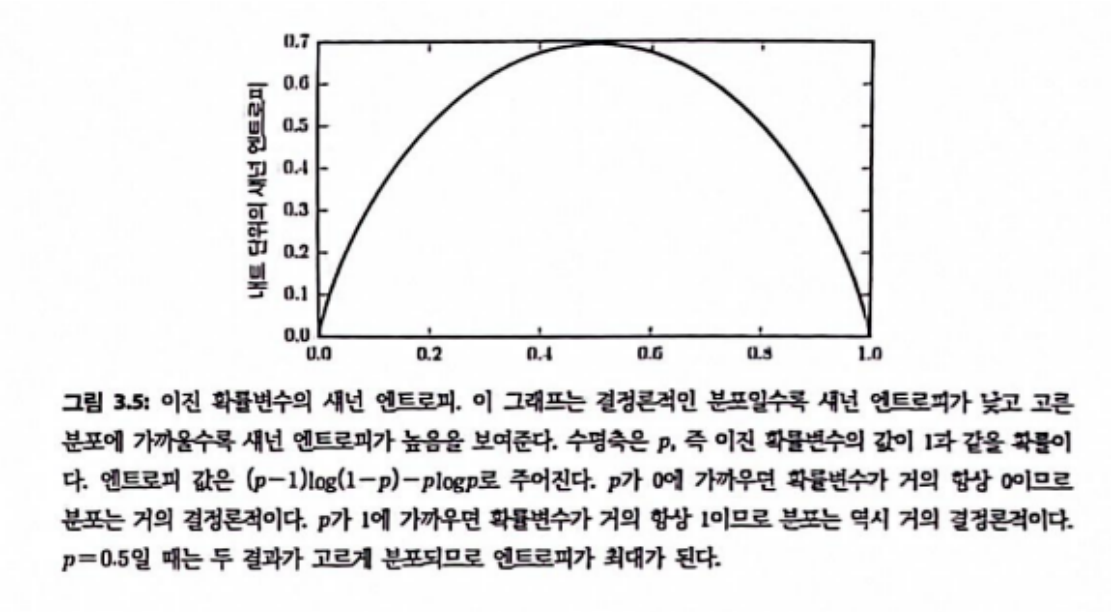

그림 3.5: 이진 확률변수의 섀넌 엔트로피. 이 그래프는 결정론적인 분포일수록 섀넌 엔트로피가 낮고 고른 분포에 가까울수록 섀넌 엔트로피가 높음을 보여준다. 수평축은 p, 즉 이진 확률변수의 값이 1과 같을 확률이다. 엔트로피 값은 $(p-1)\log(1-p)-p\log p$로 주어진다. p가 0에 가까우면 확률변수가 거의 항상 0이므로 분포는 거의 결정론적이다. p가 1에 가까우면 확률변수가 거의 항상 1이므로 분포는 역시 거의 결정론적이다. $p=0.5$일 때는 두 결과가 고르게 분포되므로 엔트로피가 최대가 된다.

- **엔트로피**는 확률분포의 불확실성을 측정하는 개념입니다. 이는 정보 이론의 핵심 개념으로, 사건이 발생할 확률에 따라 그 사건이 전달하는 정보량을 수치화합니다. 불확실성이 클수록, 즉 사건이 발생할 가능성이 낮을수록 엔트로피는 커집니다.
- **샤논 엔트로피**는 다음과 같이 정의됩니다:

$$H(x) = -\sum_x P(x) \log P(x)$$

  - 여기서 $P(x)$는 사건 x가 발생할 확률입니다. 로그 함수는 사건의 발생 확률이 작을수록 더 큰 정보를 전달한다는 것을 수학적으로 반영합니다.
  - 엔트로피는 확률분포 전체에 걸쳐 사건의 불확실성을 측정하며, 발생할 가능성이 낮은 사건일수록 정보량은 증가합니다.

- **엔트로피의 해석**:
  - 엔트로피가 클수록 시스템의 불확실성이 커짐을 의미하며, 이는 더 많은 정보를 필요로 한다는 뜻입니다.
  - 예를 들어, 균등한 확률분포는 엔트로피가 높습니다. 반면, 하나의 사건만 발생할 확률이 1이라면, 그 확률분포의 엔트로피는 0이 됩니다. 이는 불확실성이 전혀 없음을 의미합니다.

- **엔트로피의 응용**:
  - **데이터 압축**: 엔트로피는 데이터 압축에서 매우 중요한 역할을 합니다. 데이터의 불확실성을 줄이고 효율적으로 표현하는 방법을 결정할 때, 엔트로피를 기준으로 압축률을 평가할 수 있습니다.
  - **통신**: 정보 전송에서 엔트로피는 통신 시스템의 효율성을 측정하는 데 사용됩니다. 잡음이 많은 시스템에서 정보 손실을 최소화하고 데이터를 정확히 전송하기 위해 엔트로피를 고려한 시스템 설계가 필요합니다.

KL 발산 (Kullback-Leibler Divergence)

- **KL 발산**은 두 확률분포 간의 차이를 측정하는 방법입니다. 이는 **P**라는 실제 분포와 **Q**라는 기준 분포가 있을 때, 두 분포가 얼마나 다른지를 수치화하는 데 사용됩니다.
- **KL 발산의 정의**는 다음과 같습니다:

$$D_{KL}(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

  - 여기서 $P(x)$는 실제 분포, $Q(x)$는 기준 분포입니다. $\log \frac{P(x)}{Q(x)}$는 두 분포 간의 상대적인 차이를 나타내며, 이를 확률 분포 전체에 대해 계산한 값이 KL 발산입니다.

- **KL 발산의 성질**:
  - **비대칭성**: KL 발산은 대칭적이지 않으므로 $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$입니다. 이는 두 분포의 차이를 대칭적으로 계산하지 않음을 의미합니다.
  - **0보다 크거나 같다**: KL 발산의 값은 항상 0 이상이며, P와 Q가 동일할 때 KL 발산은 0이 됩니다. 즉, 두 분포가 완전히 일치할 때만 KL 발산이 0이 됩니다.

- **KL 발산의 응용**:

- **정보 손실**: KL 발산은 정보 손실을 측정할 때 사용됩니다. 예를 들어, 기계 학습 모델이 실제 분포를 얼마나 잘 추정하는지를 평가할 때 사용될 수 있습니다. 분포가 다를수록 KL 발산 값이 커지며, 이는 모델이 정보를 얼마나 잘 보존하고 있는지를 나타냅니다.
- **최적화 문제**: KL 발산은 최대 우도 추정(Maximum Likelihood Estimation, MLE)이나 변분 추정(Variational Inference)에서 사용되며, 두 분포 간의 차이를 최소화하는 방향으로 모델을 최적화하는 데 사용됩니다.

KL 발산(Kullback-Leibler Divergence)과 교차 엔트로피(Cross-Entropy)

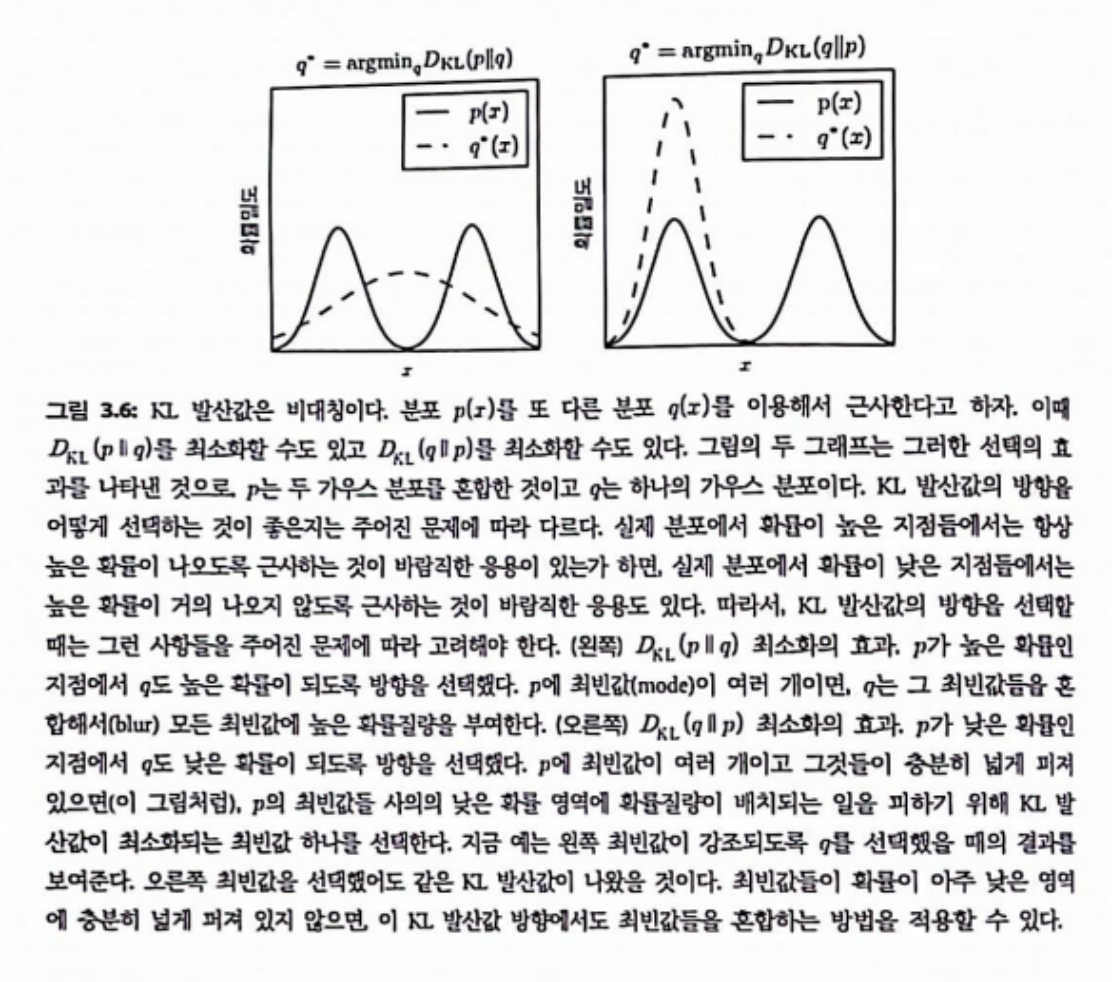

그림 3.6: KL 발산값은 비대칭이다. 분포 $p(x)$를 또 다른 분포 $q(x)$를 이용해서 근사한다고 하자. 이때 $D_{KL}(p\|q)$를 최소화할 수도 있고 $D_{KL}(q\|p)$를 최소화할 수도 있다. 그림의 두 그래프는 그러한 선택의 효과를 나타낸 것으로, p는 두 가우스 분포를 혼합한 것이고 q는 하나의 가우스 분포이다. KL 발산값의 방향을 어떻게 선택하는 것이 좋은지는 주어진 문제에 따라 다르다. 실제 분포에서 확률이 높은 지점들에서는 항상 높은 확률이 나오도록 근사하는 것이 바람직한 응용이 있는가 하면, 실제 분포에서 확률이 낮은 지점들에서는 높은 확률이 거의 나오지 않도록 근사하는 것이 바람직한 응용도 있다. 따라서, KL 발산값의 방향을 선택할 때는 그런 사항들을 주어진 문제에 따라 고려해야 한다. (왼쪽) $D_{KL}(p\|q)$ 최소화의 효과. p가 높은 확률인 지점에서 q도 높은 확률이 되도록 방향을 선택했다. p에 최빈값(mode)이 여러 개이면, q는 그 최빈값들을 혼합해서(blur) 모든 최빈값에 높은 확률질량을 부여한다. (오른쪽) $D_{KL}(q\|p)$ 최소화의 효과. p가 낮은 확률인 지점에서 q도 낮은 확률이 되도록 방향을 선택했다. p에 최빈값이 여러 개이고 그것들이 충분히 넓게 퍼져 있으면(이 그림처럼), p의 최빈값들 사의의 낮은 확률 영역에 확률질량이 배치되는 일을 피하기 위해 KL 발산값이 최소화되는 최빈값 하나를 선택한다. 지금 예는 왼쪽 최빈값이 강조되도록 q를 선택했을 때의 결과를 보여준다. 오른쪽 최빈값을 선택했어도 같은 KL 발산값이 나왔을 것이다. 최빈값들이 확률이 아주 낮은 영역에 충분히 넓게 퍼져 있지 않으면, 이 KL 발산값 방향에서도 최빈값들을 혼합하는 방법을 적용할 수 있다.

- **KL 발산**은 두 확률 분포 P와 Q 사이의 차이를 측정하는 방법으로, 특히 Q가 실제 분포 P에 비해 얼마나 다른지, 얼마나 많은 정보가 손실되었는지를 수치화하는 데 사용됩니다.
- 교차 엔트로피(Cross-Entropy)는 **KL 발산**과 밀접한 관련이 있으며, 두 확률 분포 P와 Q 사이의 거리를 측정하는 수치입니다. 교차 엔트로피는 다음과 같이 정의됩니다:

$$H(P, Q) = -\sum_x P(x) \log Q(x)$$

  - 이 값은 두 분포가 얼마나 다른지를 나타내며, 두 분포가 완전히 동일하다면 교차 엔트로피는 최소가 됩니다.

- **KL 발산과 교차 엔트로피의 관계**:

$$H(P, Q) = H(P) + D_{KL}(P \parallel Q)$$

  - 여기서 $H(P)$는 실제 분포 P의 엔트로피이고, $D_{KL}(P \parallel Q)$는 P와 Q 사이의 KL 발산입니다.
  - 즉, 교차 엔트로피는 P의 엔트로피에 두 분포 간의 KL 발산을 더한 값입니다. 이는 실제 분포와 모델 분포 간의 차이가 클수록 교차 엔트로피가 커진다는 것을 의미합니다.

- **최소화 문제에서의 사용**:
  - KL 발산을 최소화하는 것은 실제 분포 P와 모델 분포 Q 간의 차이를 줄이는 방법입니다. 기계 학습에서는 이 값을 최소화하는 방향으로 학습이 이루어지며, 분류 문제에서도 모델의 성능을 평가할 때 자주 사용됩니다.

3.14 구조적 확률 모형 (Structured Probabilistic Models)

- **구조적 확률 모형**은 확률 변수를 여러 개의 부분으로 나누어 각각의 관계를 모델링하는 방법입니다. 하나의 복잡한 확률 분포를 단일 함수로 모델링하는 대신, 여러 변수를 구조적으로 표현하여 더 간단한 형태로 분해하는 방식입니다.

- **조건부 독립성**:
  - 구조적 확률 모형은 **조건부 독립성**을 활용하여 변수들 간의 관계를 단순화합니다. 조건부 독립성이란, 한 변수가 주어졌을 때 다른 두 변수가 서로 독립임을 의미합니다. 예를 들어, 확률 분포 $P(a, b, c)$가 있을 때, b와 c가 a에 대해 조건부 독립이라면 다음과 같은 관계식을 사용하여 분해할 수 있습니다:

$$P(a, b, c) = P(a)P(b|a)P(c|b)$$

  - 이는 복잡한 확률 분포를 더 단순한 형태로 쪼개어, 계산과 분석을 용이하게 합니다.

- **그래픽 모형(Probabilistic Graphical Models)**:
  - 구조적 확률 모형의 한 예로 **확률적 그래픽 모형**이 있습니다. 이 모형은 변수들 간의 관계를 **그래프**로 표현하여, 각 변수 간의 조건부 독립성을 시각적으로 나타냅니다. 각 변수는 노드로 표현되고, 변수 간의 관계는 엣지로 나타냅니다.
  - 베이즈 네트워크(Bayesian Network)와 마르코프 랜덤 필드(Markov Random Field)가 대표적인 그래픽 모형의 예입니다. 이러한 모형들은 복잡한 확률 분포를

효율적으로 모델링하는 데 유용합니다.

구조적 확률 모형의 장점

- **계산 효율성**: 복잡한 확률 분포를 여러 개의 간단한 분포로 분해할 수 있기 때문에 계산량이 크게 줄어듭니다.
- **해석 가능성**: 변수들 간의 관계를 시각적으로 표현할 수 있어, 데이터 분석이나 해석이 용이해집니다.
- **모델 확장성**: 새로운 변수가 추가되더라도 구조적인 형태를 유지하면서 모델을 확장할 수 있어, 다양한 문제에 유연하게 대처할 수 있습니다.

유향 그래프 모형 (Directed Graphical Models)

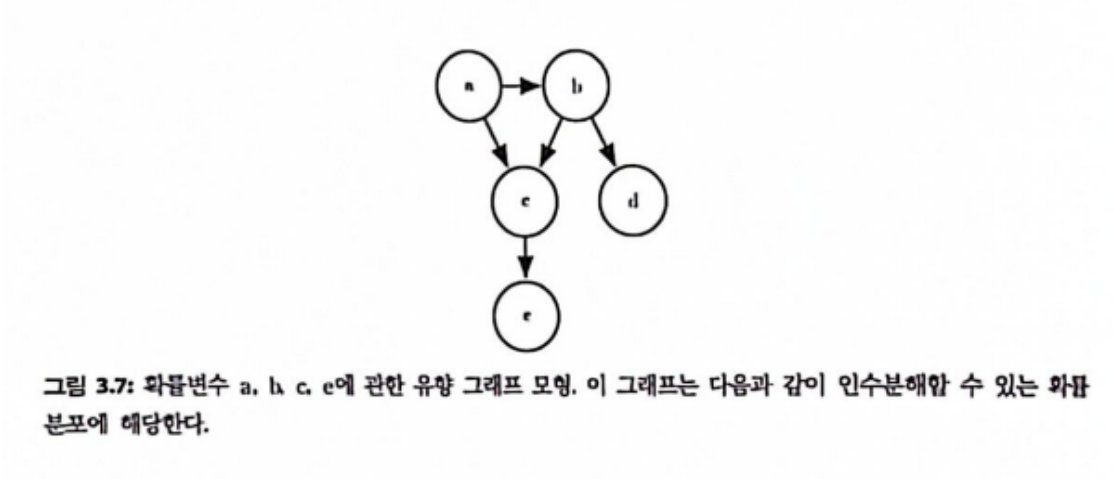

그림 3.7: 확률변수 a, b, c, e에 관한 유향 그래프 모형. 이 그래프는 다음과 같이 인수분해할 수 있는 확률 분포에 해당한다.

- 유향 그래프 모형에서는 확률 변수가 방향을 가지는 화살표로 연결됩니다. 이는 한 변수가 다른 변수에 영향을 미친다는 의미로, **조건부 확률**을 기반으로 각 변수를 표현합니다.
- **유향 그래프 모형에서 확률 분포를 표현하는 방법**:
  - 유향 그래프에서 각 변수는 그 부모 변수를 조건으로 한 **조건부 확률**로 표현됩니다. 이는 다음과 같이 수식으로 나타낼 수 있습니다:

    $$P(x) = \prod_i P(x_i|Pa(x_i))$$

    - 여기서 $Pa(x_i)$는 변수 x_i의 부모 변수(해당 변수에 영향을 미치는 변수)를 나타냅니다.
    - 예를 들어, 그림 3.7의 유향 그래프에서는 각 변수가 부모 변수를 조건으로 확률을 표현하며, 이러한 관계는 변수 간의 인과적 관계(causal relationship)를 설명하는 데 자주 사용됩니다.

- **예시**:
  - 그림 3.7에 나오는 유향 그래프에서 a, b, c, d, e 변수가 서로 연결된 방식은 각 변수 간의 **조건부 독립성**을 나타냅니다. 예를 들어, e는 c와 d에 의해 영향을 받으며, 이는 확률적으로 $P(e \mid c, d)$로 표현됩니다.
- 유향 그래프 모형은 베이즈 네트워크(Bayesian Networks)와 같은 모델에서 자주 사용되며, 변수 간의 인과관계를 모델링할 때 매우 유용합니다.

무향 그래프 모형 (Undirected Graphical Models)

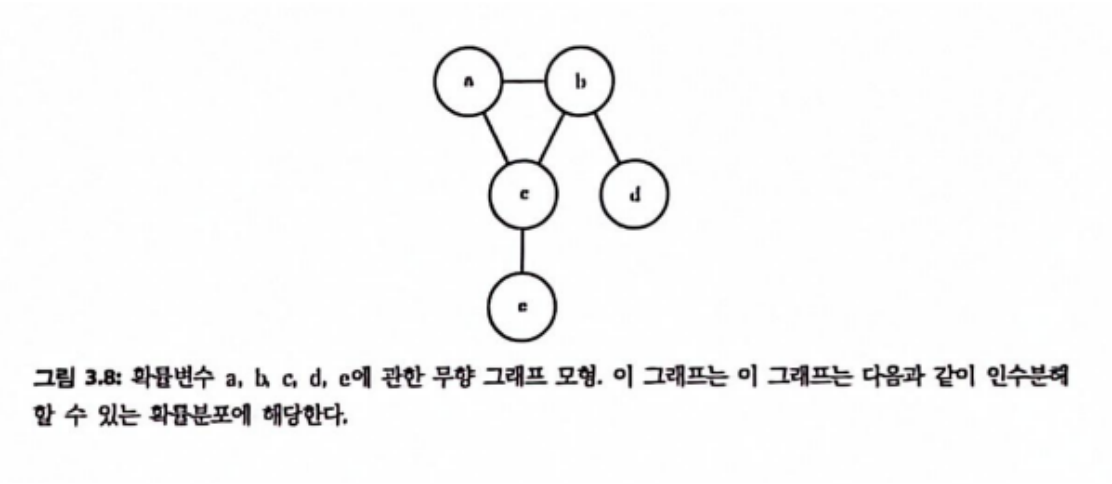

그림 3.8: 확률변수 a, b, c, d, e에 관한 무향 그래프 모형. 이 그래프는 이 그래프는 다음과 같이 인수분해 할 수 있는 확률분포에 해당한다.

- **무향 그래프 모형**에서는 변수 간의 관계를 나타내는 엣지가 방향성을 갖지 않습니다. 즉, 변수들 간의 **대칭적 관계**를 나타내며, 이는 **조건부 독립성**보다는 변수 간의 결합 확률을 표현하는 데 중점을 둡니다.
- **무향 그래프 모형에서 확률 분포를 표현하는 방법**:
  - 무향 그래프에서는 각 변수를 조건부 확률이 아닌 클릭(potential functions)을 사용하여 표현합니다. 클릭은 서로 연결된 변수들의 집합을 나타내며, 이는 결합 확률로 표현됩니다. 수식은 다음과 같이 표현됩니다:

    $$P(x) = \frac{1}{Z} \prod_C \phi_C(x_C)$$

    - 여기서 $\phi_C(x_C)$는 클릭에서 정의된 잠재 함수이고, Z는 정규화 상수(normalizing constant)로 전체 확률의 합이 1이 되도록 조정합니다.
- **예시**:
  - 그림 3.8의 무향 그래프에서는 a, b, c, d, e변수가 상호작용하며, 이러한 변수들의 관계는 상호 대칭적인 관계로 표현됩니다. 이때 각 변수는 인접한 변수들과의 결합 확률로 나타냅니다.

- 무향 그래프 모형은 마르코프 랜덤 필드(Markov Random Fields)와 같이 변수 간의 대칭적인 상호작용을 모델링하는 데 자주 사용됩니다. 이 모형은 특히 이미지 처리와 신경망 분야에서 중요한 역할을 합니다.

유향 그래프와 무향 그래프의 차이

- **유향 그래프**는 변수 간 방향성 있는 관계를 나타내며, 인과관계 모델링에 적합합니다. 각 변수를 **조건부 확률**로 표현하며, 베이즈 네트워크 등의 모델에 활용됩니다.
- **무향 그래프**는 변수 간 상호 대칭적 관계를 나타내며, 각 변수를 **클릭 함수**로 표현합니다. 마르코프 랜덤 필드와 같은 모델에서 사용되며, 주로 변수 간 상호작용을 다룹니다.