**<데이터 분석을 위한 수리 모델 입문>**

에자키 타카히로

**1장. 데이터 분석과 수리 모델**

**1) 수리 모델의 역할**

**① 데이터를 바라보는 것 이상의 분석이 필요하다면 수리 모델을 사용할 때이다.**

데이터 분석이라는 말을 들으면 가장 먼저 평균값, 표준편차와 값은 값(기술 통계량: descriptive statistics)을 계산하거나 그래프를 그리는 일을 떠올리게 된다. 이런 방식으로 어떤 현상이 변화하는 경향을 읽어내거나 그 배경에 이는 메커니즘을 상상하는 것 즉, 데이터의 특징을 파악하는 일은 매우 중요한 일이다.

그러나 기술 통계량으로만 해결할 수 없는 문제들도 있다. 예를 들어 ‘현상의 메커니즘을 객관적인 방법으로 밝힌다’, ‘데이터로 미래를 예측한다’, ‘컴퓨터에 수준 높은 데이터 처리와 데이터 생성을 맡긴다’ 와 같은 문제를 해결해야 할 때 수리 모델(mathematical model)이 필요하다.

**수리 모델:** 대상이 되는 데이터의 생성 규칙을 수학적인 수단을 사용해서 시뮬레이팅하는 것을 말한다. 분석 대상의 데이터를 자유롭게 생성하거나 내부 조건을 변경해서 처리 할 수 없기 때문에 이렇게 생성 및 변경을 할 수 있게 수식을 가지고 그 복사본을 만드는 것이다. 이렇게 해서 분석하거나 예측하는 데 자유롭게 활용할 수 있다. 또한 수리 모델은 주목할 변수들 사이의 관계성을 수학적으로 표현한 것을 말한다.

**시스템(system):** 분석의 대상. 주목해야 할 한 덩어리. Ex) 고객의 행동, 사회 현상 등

**② 수리 모델은 ‘가정하는 것’이다.**

수리 모델을 만들기 위해서는 시스템을 표현할 수학적인 틀을 정해야 한다. 그리고 ‘이런 틀 안에서 수식을 만들어서 잘 해결된다’라고 가정하고 그것을 기반으로 분석해가는 것이다. 수리 모델을 가정한 후에는 대상이 되는 시스템과 이 수리 모델이 최대한 같은 데이터를 내놓을 수 있게 조정한다.

**2장. 수리 모델을 구성하는 요소와 종류**

**1) 변수 수리의 구조 매개변수**

**① 수리 모델을 구성하는 요소**

변수(variable), 수리의 구조(mathematical structure), 매개변수(parameter)

**② 변수의 종류**

양적변수, 질적변수(카테고리 변수), 관측변수, 잠재변수, 목적변수(종속변수), 설명변수(독립변수)

관측변수(observable variable): 직접 관측해서 데이터를 구할 수 있는 변수

Ex) 어떤 고객이 편의점에서 구입한 상품의 목록이 있다고 했을 때 ‘무엇을 몇 개 구입했는지’는 관측변수이다. 그러나 이것만으로 ‘왜 상품을 구입했는지(디자인이 좋아서? 값이 저렴해서? 등등)는 판단할 수 없다. 이렇게 구입한 이유는 잠재변수(latent variable)이다.

**③ 매개변수는 수리 모델을 ‘움직인다’**

수리 모델에는 그 수리 모델의 구조를 데이터에 잘 맞추기 위한 ‘가동영역’이라는 것이 있다. 수리 모델의 움직임을 제어하는 것이 매개변수이다. 매개변수의 개수가 많으면 여러 방향으로 ‘가동영역’(자유도: degree of freedom)을 마련할 수 있으므로 수리 모델의 표현력이 좋아진다.

**④ 수리구조 = ‘수리 모델의 뼈대’**

주목할 변수들 사이의 관계성을 수학적으로 표현한 것이 수리 모델이다. ‘수학적으로 표현할 때’ 필요한 것(예를 들어 수식 등)을 모두 모아서 수리 구조(mathematical structure) 라고 하고 이 수리 구조가 정해지면 수리 모델의 뼈대가 정해진다. 제대로 분석이 이뤄지려면 적절한 수리 구조를 선택하는 것이 중요하다.

**2) 수리 모델과 자연과학의 기초 이론**

**① 확립된 수리 모델은 기초 이론이 된다**

수리 모델은 미분 방정식(differential equation)에 의해 표현된다. 미분은 ‘어떤 것이 어떤 것에 대해 변화하는 비율을 계산한 것’을 나타내는 중요한 변수이다. 예를 들어 이동하는 물체의 속도는 물체의 위치가 시간에 따라 변화하는 비율을 나타낸 것이며 미분으로 표현된다.

**② 경계조건과 계산의 난이도**

수리 모델에는 문제에 따라 적용할 수 있는 시간적, 공간적인 범위가 존재한다. 이 범위의 경계에서 수리 모델이 만족시켜야 할 조건을 경제조건(boundary condition)이라고 한다. 현재 생각하는 문제의 처음 시점에 만족시켜야 할 조건인 초기조건(initial condition)도 경계조건 중 하나이다.

**3) 이해 지향형 모델링과 응용 지향형 모델링**

**① 목적에 따라 모델링은 크게 달라진다**

수리 모델을 통한 분석은 목적에 따라 크게 두 종류로 나뉜다.

이해 지향형 모델링

-> ‘데이터가 생성되는 구조를 이해하는 것’을 가장 큰 목표로 삼고 실행하는 모델링

응용 지향형 모델링

-> ‘현재 가지고 있는 데이터를 기반으로 미지의 데이터에 관해 예측하고 제어하거나 새로운 데이터를 생성해서 이용하는 것’을 가장 큰 목표로 삼고 실행하는 모델링. 이미지를 판별하거나 생성하는 일, 자동차의 자율 주행이나 기계로 번역하는 일에 응용된다.

**4) 이해 지향형 모델링 (4가지 방법)**

이해 지향형 모델링은 ‘데이터를 잘 설명하는 수리 모델은 데이터가 생성되는 과정을 어느 정도 파악해서 포함하고 있으므로 해당 수리 모델을 조사해 보면 이해할 수 있다’ 라는 기본 개념을 전제로 하며 4가지 방법이 있다.

① 수리 구조를 통해 설명하는 방법

이 방법을 통해 ‘모델이 데이터를 설명할 수 있는 이유는 수리 모델을 만들 때 가정한 수리 구조가 옳았기(가능성이 높기) 때문이다’라는 논리를 세운다. 모델을 만들 때 가정해도 괜찮다고 생각되는 경험 사실과 관측 사실이라는 밑바탕으로부터 위 단계로 진행하면서 논리를 구성한다. 따라서 처음에 설정하는 가정이 좋아야하고 그 대신에 연역적으로 이해되기만 한다면 해당 모델의 정확도는 그다지 중요하지 않다.

② 추정한 매개변수의 값을 통해 설명하는 방법

모델 안에 있는 어떤 매개변수의 의미와 작용이 분명할 경우에는 추정된 값이 데이터의 한 측면을 보여준다. 이렇게 얻어진 값을 사용해 분석 대상을 해석할 수도 있다.

③ 추정된 잠재변수나 내부 표현을 사용해 그 다음 부분을 해석하는 방법

추정된 모델의 잠재변수와 그 잠재변수들의 내부 표현을 사용해 데이터를 변환하고 그것을 사용해 다음 데이터 분석을 실행하는 방법이다. 변수의 개수가 많은 데이터(고차원 데이터)에서는 값을 그냥 관찰하거나 그래프로 그리는 방법으로는 내부에서 무슨 일이 벌어졌는지 알 수 없다. 이 경우에는 일단 잠재변수를 포함한 수리 모델로 표현한 후에 해당 잠재변수를 사용해 저차원 표현으로 변환하는 방법을 이용한다.

④ 수리 모델에 있는 매개변수를 변화시키면서 상황을 시뮬레이팅하는 방법

수리 모델에 있는 매개변수에 일부러 현실의 상황과는 다른 값을 넣은 후에 모델이 동작하는 모습을 시뮬레이팅해서 그 대상을 깊이 이해하는 방법이다.

**5) 응용 지향형 모델링**

응용 지향형 모델링에서는 수리 모델에서 나오는 데이터 출력을 사용한다. 응용방식에는 ‘예측’과 ‘생성’ 두 종류가 있는데 ‘예측’은 수리 모델을 만들 때 사용한 데이터가 있던 상황과는 다른 상황에서 어떤 양을 알아 맞히는지 그 작업을 말한다. 내일의 날씨를 예상하거나 고객의 정보에서 해당 고객이 흥미를 가질 만한 상품을 추정해서 제안하는 일도 예측이다.

‘생성’은 수리 모델을 만들 때 사용한 것과 닮은 데이터를 출력하는 작업을 말한다. 기계번역이나 사진 가공에 유용하게 응용할 수 있다.

**3장. 방정식 모델**

**1) 선형모델**

- 변수 사이의 관계가 덧셈, 뺄셈, 정수배만으로 표현된 모델을 선형모델이라고 하고 변수들 사이에 곱셈, 나눗셈, 삼각함수, 지수함수 등으로 표현된 모델을 비선형 모델이라고 한다.

**2) 실험식과 커브 적합**

어떤 변수의 값이 다름 변수의 값의 몇 제곱에 비례하는 관계를 멱법칙(power law) 또는 눈금 바꿈 법칙(scaling law)이라고 한다. ‘어떤 것이 멱법칙 관계에 있다’는 사실을 밝히려고 할 때 데이터의 개수나 변수의 범위가 충분하지 않으면 틀린 결론에 이르게 될 때가 많으니 주의해야 한다.

**3) 최적화 문제**

변수의 값을 제어해서 다른 어떤 변수의 값을 최대화(또는 최소화)하는 문제를 최적화 문제라고 한다.

**4장. 미분 방정식으로 구성된 모델**

미분 방정식을 사용한 모델은 어떤 대상이 시간에 따라 변화하는 모습을 조사할 때 자주 사용된다.

**1) 미분 방정식을 통해 시간 변화를 나타낸다.**

대상이 시간에 따라 변화하는 모습을 바닥부터 차근차근 모델화하려고 할 때 먼저 그 후보로 떠오르는 것이 미분 방정식을 사용한 모델링이다. 미분은 ‘대상이 되는 변수가 어떤 것에 대해 변화하는 비율’을 나타낸다. 변수가 시간에 따라 변화하는 모습을 직접 구현한 모델을 역학계(dynamical system)이라고 한다.

**2) 비선형 미분 방정식 모델**

미분 방정식은 일반적으로 어떤 변수나 변수의 미분을 곱한 항이 식에 포함돼 있으면 해석적으로 풀 수 없다. 그러나 실제 모델링에서는 이런 요소를 포함한 복잡한 상황을 표현해야 할 때가 자주 있다.

① 로트카-볼테라 방정식

순환 현상을 나타내는 미분 방정식 중 하나로, 생태학에서 사용되고 특히 자연계에서 인구의 변화를 모델링하는데 사용된다. 개체 수가 늘어나는 속도와 줄어드는 속도가 서로의 상대편 생물의 개체 수에 의존한다는 형태의 식이다. 예를 들어 사자와 얼룩말처럼 먹고 먹히는 관계에서 포식자(사자)는 피식자(얼룩말)를 잡아먹으므로 피식자의 수가 줄어들면 그에 따라 포식자의 수도 줄어든다.

② 룽게-쿠타 방법

룽게-쿠타 방법은 미분 방정식의 수치 해법 중 하나로, 주어진 초기값에서 시작하여 단계별로 근사적인 해를 계산하는 방법이며 미분 방정식의 해가 해석적으로 구하기 어려울 때 유용하다.

**3) 풀 수 있는 모델과 풀 수 없는 모델**

선형인 미분 방정식은 기본적으로 풀린다. 하지만 풀 수 있는 미분 방정식은 많지 않다. 변수가 하나인 비선형 미분 방정식은 시간에 관한 함수로 표현되는 변수가 하나일 때는 방정식이 다소 비선형이라고 풀리는 경우가 있다. 하지만 미분 방정식을 사용해서 수리 모델을 만들 때 변수가 하나만 필요한 상황은 많지 않다. 또한 편미분 방정식도 그 형태가 선형이면 풀 수 있을 때도 있지만 비선형이면 해석적으로는 풀 수 없는 경우가 대부분이다.

① 분석 소프트웨어를 이용한다

미분 방정식은 풀기 어렵기에 소프트웨어를 이용한다. 미분 방정식을 해석적으로 풀려고 할 때는 상용 소프트웨어인 Mathematica나 Maple, 그리고 무료 소프트웨어인 Maxima, 웹서비스로 실행되는 Wolfram Alpha가 있다. 수치 시뮬레이션이 필요하다면 위에서 열거한 프로그램에 추가해서 사용 소프트웨어인 MATLAB, IMSL, NAG, 무료 소프트웨어인 Scilab, GNU, Octave, 파이썬의 SciPy를 이용할 수 있다.

**4) 제어이론(control theory)**

제어이론(control theory): 어떤 변수의 값을 자유롭게 변화시킬 수 있을 때 그 변수가 다른 변수에 주는 영향을 평하거나 제어할 수 있는 개념이다. 예를 들어 가정이나 사무실에서 온도가 일정하게 유지되기 위해서는 히터나 에어컨과 같은 시스템이 사용되는데 이때, 온도 조절을 위한 제어 시스템이 제어이론을 활용한다.

① 미분 방정식을 풀기 위해 사용되는 도구

라플라스 변환(Laplace transform): 어떤 변수(이 책에서는 시간에 관한 함수로 표현된다)에 적분 계산을 적용해서 다른 함수로 변환하는 절차를 말한다.

**5장 확률모델**

**1) 확률과정**

**① 확률적인 상황을 생각해보자.**

현실 세계에서 수집된 데이터를 다룰 때 변수가 반드시 매번 동일한 움직임을 보이지는 않으며 어느 정도 분산된다. 분산이 비교적 무시할 수 있을 정도로 작을 때는 미분 방정식을 사용하지만 분산이 무시할 수 없을 정도로 클 때는 확률적인 요소를 포함한 모델이 해당 대상의 움직임을 잘 파악하기도 한다.

**② 확률분포는 확률의 정보를 모아 놓은 것이다**

**확률분포(probability distribution):** 어떤 사건이 발생할 확률을 전부 모아서 어떤 확률변수에 부여한 것을 확률분포라고 한다.

예) 주사위를 던졌을 때 그 결과로 몇 개의 눈이 나왔는지를 주사위를 던지기 전에는 어떤 숫자가 나왔는지 알 수 없다. 주사위의 숫자 개수를 X라고 한다면 X를 확률변수(random variable)이라고 한다. 그리고 주사위를 던졌을 때 발생할 확률을 전부 모아서 어떤 확률변수에 부여한 것을 확률분포라고 한다.

**③ 연속변수의 확률 분포: 확률밀도함수**

연속변수의 값이 발생하는 확률을 나타내기 위해 확률밀도함수(probability density function)을 사용한다. ‘확률변수 X가 a에서 b까지의 구간에 들어갈 확률’을 P(a b)라고 표기한다. 확률밀도함수를 히스토그램(도수분포표)와 같은 것 이라고 생각하면 된다.

**확률과정(stochastic process):** 시간의 변화에 대한 확률변수 (예: 주사위를 여러 번 던져서 매번 나오는 눈의 개수)

**2) 마르코프 과정(Markov process)**

**① 마르코프 과정은 과거의 상태를 무시한다**

마르코프 과정은 시스템의 확률변수(상태라고도 한다)를 변화시킬 때 현재 상태의 정보만을 사용해 다음 상태를 결정하는 확률과정이다. 마르코프 과정을 수리적으로 해석하기 쉽고 널리 응용할 수 있다는 장점이 있다.

예시1) 점심 메뉴 정하기

E씨는 매일 회사 근처에서 점심식사를 한다. 회사 근처에는 라면 가게, 스테이크 가게, 국수 가게 이렇게 세 가게 뿐이며 매일 이 매일 이 중에 하나를 골라서 식사를 하는 것이 번거롭게 느껴졌다. 그래서 E씨는 점심 메뉴를 확률적으로 결정하기로 했다. 그러나 완전히 무작위로 결정하면 운이 나쁠 때는 며칠 동안 연속으로 같은 가게에서 식사하게 될 수 도 있다. 그렇다고 순서를 정해놓고 그대로 따르는 것도 재미없을 거 같아서 다음과 같은 규칙으로 가게를 결정하기로 했다.

그 규칙은 바로 오늘 간 가게에 따라 내일 갈 가게가 결정된다. E씨는 라면을 좋아하기 때문에 주사위를 던졌을 때 1, 2, 3 중 하나가 나오면 내일도 라면 가게에 가고 4, 5 중 하나가 나오면 스테이크 가게에 가고 6이 나오면 국수 가게에 간다.

이처럼 상태가 변화해가는 것을 *상태천이(state translation)*라고 하고 각각의 상태천이가 발생할 확률을 *천이확률(transition probability)*라고 한다.

**3) 대기행렬이론**

확률 과정을 사용해서 현상을 모델링하는 방법인 대기행렬이론은 일정한 시간 혹은 공간에서 발생하는 서비스 요청이나 이벤트를 모델링하고 분석하는 수학적인 도구이며 특히, 서비스 시스템이나 대기열이 발생하는 상황에서 시스템의 성능을 평가하고 최적화하는 데 사용된다.

**① 포아송 과정(Poisson Process)**

포아송 과정은 일정한 시간 또는 공간에서 발생하는 사건들을 모델링하는 확률과정 중 하나로 독립적인 사건들이 일정한 평균 발생률로 발생하는 상황을 나타낸다.

예를 들어 특정 도로에서 일정 시간 동안 교통사고가 발생하는 경우 만약 교통사고가 평균적으로 시간당 2번 발생한다고 가정하면, 포아송 과정을 사용하여 이를 모델링할 수 있다.

**② 출생사멸과정(birth-death process)**

출생사멸과정은 인구 통계학에서 사용되는 모델 중 하나로, 출생과 사망 사이의 인구 변화를 설명하는 확률 과정이며 주어진 기간 동안 얼마나 많은 사람이 태어나고 죽을지를 모델링한다.

예를 들어 어떤 지역에서 특정 기간 동안 출생율이 평균적으로 연간 10명/천명이라고 가정하고 사망율이 연간 5명/천명이라고 가정하자. 이때, 출생사멸과정을 사용하여 해당 지역의 인구 변화를 추정할 수 있다.

**6장 통계모델**

**1) 정규분포**

**① 균등분포(uniform distribution)**

발생할 수 있는 모든 사건에 관해 확률 분포가 같은 값. 예를 들어 주사위를 한 번 던지면 말이 몇 칸 째에 있을까를 생각하면 ‘각각 동일한 1/6의 확률로 첫 번째 칸에서 여섯 번째 칸 중의 한 곳에 있게 된다’라고 말할 수 있다.

주사위를 10번 던지고, 100번 던지고 이런 작업을 계속하면 확률분포는 *정규분포(normal distribution)*라는 분포에 가까워진다. 이런 현상을 *중심극한정리(central limit theorem)*라고 말한다. 분포가 어떤 것이든지 동일한 분포에서 독립적으로 생성된 확률변수를 계속 더해가면 그 합을 나타내는 분포가 정규분포에 가까워진다고 말할 수 있다. 이런 사실이 있기 때문에 현실 세계의 다양한 현상이 정규분포에 가까운 분포를 따른다고 알려져 있고 그런 의미에서 정규분포는 확률분포의 왕과 같은 존재라고 할 수 있다.

**② 정규분포**

정규분포에는 평균(mean)과 표준편차(standard deviation)라는 두 개의 매개변수가 있다. 평균은 해당 분포를 따르는 확률변수의 평균적인 크기를 나타내는 양이고 표준편차는 확률 변수의 값이 얼마나 분산돼 있는가를 나타내는 양이다. 확률분포 그래프의 모양을 가지고 이야기하자면 평균은 산의 꼭대기가 어디에 있는가에 해당되고 표준편차는 산이 완만한 정도에 해당된다.

평균을 라고 표기하고 표준편차를 라고 표기한다. 그리고 평균이 이고 분산이 인 정규분포를 *N(, )*라고 표기한다. 특히 평균이 0이고 표준편차가 1인 정규분포(=N(0, 1))는 기본적 정규분포로 다양한 곳에 사용된다. 이것을 *표준정규분포(standard normal distribution)*라고 한다.

**③ 정규분포에는 재생 성질이 있다 (다른 정규분포끼리 덧셈 가능)**

재생 성질이란 무엇인가? 예를 들어 A씨와 B씨가 컨베이어 벨트에서 함께 작업하고 있는 상황을 생각해보자. A씨가 작업하는데 걸리는 시간을 측정했더니 평균이 100분이고 표준편차가 15분인 정규분포를 따르고 B씨가 작업하는 데 걸리는 시간은 평균이 200분이고 표준편차가 20분인 정규분포를 따른다고 가정하다. 여기서 A씨가 자신의 작업을 마치고 B씨에게 넘겨주고 그 시점부터 B씨가 작업을 시작해서 얼마 후에 마쳤는데, 이 과정에 소요된 총 시간을 측정했더니 평균이 3시간이고 표준편차가 25분인 정규분포를 따랐다. 다시 말하면 A씨의 정규분포 *N(, )* 와 B씨의 정규분포 *N(, )* 을 합하면 이라는 식으로 단순히 평균과 분산을 각각 덧셈한 정규분포 *N(, )*를 따른다.

**④ 정규분포를 제곱해서 덧셈하면 어떻게 될까?**

위에서 보듯이 정규분포를 따르는 확률변수끼리의 덧셈은 또 다른 정규분포를 따른다는 사실을 알게됐다. 그럼 이 정규분포들을 제곱해서 덧셈하면 어떻게 될까?

이것은 정규분포가 되지 않는다. 이렇게 계산한 분포는 *(카이제곱분포)*가 된다. 그리고 카이제곱분포를 따르는 확률변수의 비율은 *F-분포*라는 확률분포를 따른다.

**2) 통계적 검정**

현실 세계에서 얻는 데이터에는 항상 분산과 오차가 존재한다. 통계 과학 분야에서는 이 분산이 발생할 확률에 초점을 맞추는데 데이터에서 보이는 특징이 우연히 발생한 것인지 또는 우연으로는 설명할 수 없는(의미가 있는) 것인지에 대해 평가한다.

우리의 데이터는 한정되어 있고 현재 가지고 있는 즉, 한정된 데이터에서 도출된 결과 우연이 아닐 가능성이 높음을 이야기해야한다.

예를 들어 어떤 동전을 20번 던졌더니 앞면이 16번 나왔다. 일반적으로 동전을 던지면 1/2의 확률로 앞면이 나오고 1/2의 확률로 뒷면이 나와야 정상인데, 이렇게 16번이나 앞면이 나왔다는 결과를 ‘우연’이라고 말할 수 있을까? 그래서 우리는 우연히 그렇게 된 것이 아님을 밝혀야하고 이때 가설을 사용한다. (귀무가설, 대립가설, p-value, 유의수준)

**① 귀무가설과 대립가설**

동전 예제에서 귀무가설은 ‘우연히 그렇게 됐다’ 이고 대립가설은 ‘우연히 그렇게 된 게 아니다’ 이다. 대립가설은 일반적으로 내세우고 싶은 주장에 들어간다. 이 문제에서 귀무가설을 기각하기 위해서는 해당 사건이 우연히 발생할 확률을 계산한다. 동전을 20번 던져서 16번 앞면이 나올 확률을 인데, 이 확률이 충분히 큰지 작은지를 평가할 기준이 없다. 그 대신 ‘16번 미만의 값이 나올 확률은 몇 가지인가’를 계산하면 p=0.9941이 나온다. 다시 말하면 무게중심에 이상이 없는 동전을 20번 던지면 약 99.4%의 확률로 1에서 15 사이의 횟수만큼 앞면이 나와야 하는데, 이번 실험에서는 그 이외의 사건이 발생했다고 생각할 수 있다. 이런 사건이 발생할 확률은 약 0.6%이고 이처럼 귀무가설로 설명했을 때 데이터의 사상이 발생할 확률을 p값이라고 한다. 귀무가설을 채택하면 우연으로는 설명할 수 없을 정도로 나은 확률의 사상이 발생했다고밖에는 말할 수 없기 때문에 대립가설이 맞다고 생각할 수 있고 이때 동전의 앞면이 나올 확률은 1/2과는 유의미하게 다르다는 결론을 얻게된다.

**② 1종 오류와 2종 오류**

사실은 귀무가설이 맞았는데 잘못 판단하여 귀무가설을 파기해버리는 일을 1종 오류라고 하고 대립가설이 사실은 맞았는데, 귀무가설을 파기하지 못한 오류를 2종 오류라고 한다.

**3) 회귀분석**

데이터가 선형관계를 얼마나 잘 설명하는지를 나타내는 지수로 *피어슨 상관계수(Pearson’s correlation coefficient)*가 자주 사용되는데 이것은 두 변수가 얼마나 같은 패턴으로 변동하는지를 정량화한 지표이다. R은 -1에서 1 사이의 값을 취하는데 절댓값이 클수록 선형모델이 잘 맞음을 나타낸다. 이 r값의 크기는 선형회귀로 구해진 직선의 기울기와는 직접적인 관계가 없고 단지 ‘선형이라는 관계를 가지고 데이터를 얼마나 잘 설명할 수 있는가’를 나타낸다는 점에 주의해야한다

**7장 시계열 모델**

**트렌드:** 시계열에서 비교적 긴 시간을 잡고 봤을 때 완만히 증가하거나 감소하는 경향.

문제에 따라서 시계열 데이터를 파동의 총합으로 보고 정보를 얻어내는 경우가 있다.

‘파동’은 수학적으로 삼각함수 사인(코사인) 또는 사인함수와 등가적으로 표현할 수 있는 복소수의 지수함수로 표현된다.

주파수(frequency)는 소리로 말하자면 음정의 높이에 해당되는데 어떤 특정한 원인에 의해 발생되는 진동이나 노이즈는 특유의 주파수를 가지고 있기도 한다. 그렇기에 데이터에 어떤 주파수의 파동이 얼마나 포함돼 있는지를 파악하는 일은 매우 유용하다.

**1) 예측에 사용할 수 있는 모델들**

시계열 데이터를 분석할 때 필요한 문제 설정 중 하나는 과거의 데이터를 기반으로 해서 미래를 예측하는 일이다. 이렇게 하기 위해서는 시계열을 모델로 직접 표현하고 그 모델을 기반으로 해서 예측값을 구해야 한다.

**① AR 모델 (자기회귀의 관계성을 구현)**

시계열 데이터에는 데이터점들 사이에 어떤 시간적인 관계가 있으므로 이 관계를 모델화하자는 이야기는 자연스러운 발상이다. 어떤 시점에서 변숫값이 이고 이 값이 그 바로 전 시점에서의 변숫값 과 관계가 있음을 나타내는 관계식은 다음과 같이 쓸수있다.

이때 는 상수인 매개변수이고 는 노이즈 항이다. 노이즈는 평균이 0이고 분산이 인 확률분포에서 시점마다 독립적으로 결정된다. 이것을 화이트 노이즈(white noise)라고 한다. 이 모델은 확률모델이라는 시점에서 보면 평균이 이고 분산이 ()인 노이즈 분포에서 가 생성된다고 해석할 수도 있다. 이런 관계성을 가정한 모델을 자기회귀모델(autoregressive model)이라고 하거나 앞 글자를 따서 AR 모델이라고 말한다. 이 모델은 변수를 다변수 벡터로 확장할 수 있는데, 그것을 *벡터자기회귀(VAR, vector autoregressive) 모델*이라고 부른다.

**② ARMA 모델 (이동평균 효과를 추가)**

AR 모델에서는 노이즈가 매 시각 독립적으로 변수에 가산된다는 전제를 뒀는데 ARMA 모델은 어떤 시점에 있는 노이즈가 q 시점 후까지 영향을 준다고 가정한다.

과거의 노이즈를 나타내는 항이 새로 추가됐는데, 이 항을 *이동평균(moving average)*라고 한다. 일반적인 선형회귀로 표현할 수 없는 노이즈 사이의 관계성을 이렇게 표현할 수 있다. 기호를 사용해 ARMA(p, q)라는 방식으로 표현할 수 있다.

**③ ARIMA 모델 (트렌드 효과를 추가)**

본래의 시계열에서 정상성이 만족되지 않을 경우에는 AR 모델이나 ARMA 모델을 적용할 수 없다. 그러나 각각의 시각에서 앞뒤 값의 차이를 구해서(차분) 각각의 시각에 나타난 변동만큼을 시계열로 만들면 근사적으로 정상이라고 간주할 수 있을 때가 자주 있다 (주로 경제 데이터). 이것은 트렌드를 제거하는 작업에 해당된다. 이처럼 차분을 구해서 ARMA 모델에 적용하는 방법을 ARIMA(autoregressive integrative moving average) 모델이라고 한다.

**④ SARIMA 모델 (주기(계절) 변동 효과를 추가)**

트렌드와 함께 주기의 변동도 있을 경우 주기 변동을 제거한 후에 분석하는 방법을 생각할 수 있다. 예를 들어 1년 중 발생하는 계절 변동을 제거하고 싶다면 전년 같은 시기의 값과 차분을 구해서 시계열을 만든다. 이렇게 만든 시계열에 ARIMA 모델을 적용하는 방법을 SARIMA(seasonal autoregressive integrative moving average) 모델이라고 한다.

**2) 상태공간 모델**

상태공간 모델은 시계열을 분석하는 데 사용되는 강력한 기법 중 하나이다. AR모델, ARMA모델, ARIMA모델, SARIMA모델은 실제로 데이터를 관측할 수 있는 변수인 관측 변수들 사이의 관계를 구현한 모델인데, 이 모델에 상태변수(state variable)라는 잠재 변수를 추가하여 모델링하는 것을 상태공간 모델이라고 한다. 자기회귀 모델에 필요한 정상성을 만족하지 않는 시계열에도 모델을 적용할 수 있다.

상태공간 모델은 범용성이 매우 뛰어난 기법이라고 할 수 있다.

**3) 그 밖의 다양한 시계열 분석법**

**① 푸리에 변환(Fourier transform):** 이 기법을 통해 데이터에 어떤 파동이 얼마나 포함돼 있는지를 분석할 수 있다. 기본 개념은 주어진 데이터를 삼각함수의 합으로 나타내고 그 내부에서 각각의 주파수가 얼마나 사용됐는지를 보면 데이터가 가진 특징을 파악할 수 있다.

**② 라푸노프 지수(Lyapunov exponent):** 시계열이 카오스적인지를 평가하는 지표

시계열 데이터를 분석하는 기법에는 현재 주목하는 성질에 따라 주파수 해석, 비선형 해석, 인과성 해석과 같은 다양한 기법이 있다.

**8장 머신러닝 모델**

**1) 분류와 회귀 문제**

**① 결정 트리(decision tree)**

**② 랜덤 포레스트(random forest)**

결정 트리를 많이 생성해서 이 결정 트리의 다수결로 분류 결과를 예측하는 기법이다. 여러 결정 트리를 생성하려면 본래의 데이터셋에서 무작위로 샘플을 추출해서 그 샘플을 가지고 결정 트리를 만든다.

**③ 서포트 벡터 머신(support vector machine)**

모여 있는 데이터점 그룹 사이에 선을 긋고, 그 선 위에 있으면 클래스1으로 분류하고 선 아래에 있으면 클래스2로 분류하는 것이 기본 개념이다. 이 기법에서는 선을 그을 때 각 클래스에 있는 데이터까지의 거리(마진이라고 한다)가 최대가 되게 경계선을 결정한다.

**④ 신경망(neural network)**

신경망은 뇌구조를 흉내내서 개발한 모델이며 단순한 계산을 실행하는 요소(노드)를 네트워크상에 조합해서 예측값을 출력한다. 각각의 노드(퍼셉트론)에서는 화살표로 연결된 노드에서 값을 받아오고 그 값에 적절한 계수를 곱한 후에 총합을 구한다. 그리고 이 값을 활성화 함수(activation function)라는 함수에 대입한다. 이 함수는 받아온 모든 입력을 그 다음 노드에 얼마나 전달할지를 표현하는(비선형) 함수이다. 목적에 따라 시그모이드 함수 또는 다양한 다른 함수가 이용된다. 그렇게 계산한 값을 화살표로 연결한 다음 노드에 넘겨준다. 이처럼 화살표를 따라 순서대로 계산해서 마지막에 나온 값을 예측값으로 정하고 학습할때는 이 예측값과 실제값의 오차가 작아지게 계수 값을 조정한다.

**2) 클러스터링**

데이터점의 값이 분산된 정도를 보고 가까운 데이터들을 같은 카테고리에 넣는 작업을 클러스터링이라고 한다. 어느 데이터가 어느 카테고리에 속하는지 모르는 상태에서 실행하는 비지도학습이다.

**① k-평균법(k-means)**

데이터점에서 각 클러스터의 중심까지의 거리를 비교해서 가장 가까운 클러스터에 소속시키는 알고리즘이다.

**② 혼합분포 모델**

각 클러스터의 데이터를 생성하는 확률분포를 가정해서 그 확률분포 중 하나를 통해 데이터가 생성된다라고 생각하는 방법을 혼합분포 모델이라고 하고 특히 가우스 분포를 조합해서 데이터의 확률분포를 표현한 모델을 혼합 가우스 모델이라고 한다.

**③ 계층적 클러스터링 기법**

고차원 데이터를 다룰 때는 클러스터 사이의 관계성에도 관심을 가져야 한다. 예를 들어 어떤 클러스터와 다른 클러스터가 얼마나 닮았는지 또는 얼마나 닮지 않았는지가 데이터 전체의 특징을 파악할 때 중요한 정보가 된다. 이런 클러스터 사이의 유사도를 기반으로 해서 클러스터를 나누는 방법을 말한다.

**3) 차원축소**

고차원의 데이터를 적은 수의 변수로 구성된 데이터로 치환하는 작업을 차원축소(dimensionality reduction)라고 한다.

**① 주성분 분석(PCA, principal component analysis)**

데이터가 분산된 방향으로 그릴 수 있는 직선을 찾는 기법이다. 주성분 분석에서는 직선적인 데이터밖에 파악할 수 없으므로 비선형적인 특징을 차원 축소할 수는 없다.

**② 독립성분 분석(ICA, independent component analysis)**

성분의 개수를 미리 지정해서 데이터를 그 개수만큼의 독립된 성분으로 표현한다. 주성분 분석과는 달리 이 성분들 사이에 서열 관계가 없다.

**4) 딥러닝**

앞서 소개한 신경망에서 중간층의 개수를 늘린 것을 심층 신경망이라고 하고 이 심층 신경망을 사용한 머신러닝을 딥러닝이라고 한다. 딥러닝은 복잡한 모델로 대응하는 기법이라고 할 수 있다.

**① 컨볼루션 신경망(CNN, convolutional neural network)**

이미지 식별 등에 사용되며 고성능을 자랑하는 신경망이다. CNN은 뇌의 시각영역이 작용하는 모습에 힌트를 얻어 고안된 컨볼루션층과 풀링층이라는 중간층을 포함하는데 어떤 대상의 패턴이 있는 위치가 달라져도 동일한 방식으로 처리할 수 있다.

**② 순환 신경망(RNN, recurrent neural network)**

시계열 모델링에 자주 사용되는 신경망으로 일반적인 순전파(신경망의 입력에 각각의 시각에 대응되는 데이터를 입력으로 받으면 매 시각 출력을 얻는데, 서로 다른 시각에 대응되는 데이터 사이의 관계성이 모델에 반영되지 않음)의 문제를 해결하기 위해 순방향뿐만 아니라 이전으로 되돌아가는 경로를 네트워크에 추가한다. 과거에 있는 중간층의 값을 유지하거나 과거에 나온 출력을 중간층에 되돌리는 등 다양한 방법으로 실행된다.

**③ 오토인코더(autoencoder)**

입력과 출력이 같아지게 학습시키는 신경망이다.

**④ 적대적 생성 네트워크(GAN, generative adversarial network)**

이 모델은 이미지 등을 생성할 때 사용된다. 생성기(generator)와 식별기(discriminator)라는 모델을 동시에 학습시킨다. 생성기는 데이터를 생성하는 신경망이고 식별기는 생성기로 생성한 데이터와 실제로 학습한 데이터의 차이를 알아내는 신경망이다. 생성기는 자신이 생성한 데이터가 식별기에 의해 실제 데이터(실제로 학습한 데이터)라고 오판되는 것을 목표로 해서 학습한다. 그리고 식별기는 그 거짓을 알아채는 것을 목표로 해서 학습한다. 이처럼 라이벌 관계에 있는 두 모델을 학습시키면 최종적으로는 진품과 구분이 되지 않는 데이터를 생성할 수 있다.

**9장 강화학습 모델**

강화학습이란 환경에서 오는 피드백에서 가장 적합한 반응을 찾는 프레임워크이다. 강화학습은 알 수 없는 환경의 상황에 따라 의사결정하는 주체(에이전트)가 적절한 행동을 학습하는 모습을 모델화한다. 머신러닝 분야에서의 최적의 행동을 기계에 학습시키는 방법으로도 이용된다.

**1) 강화(reinforcement)**

인간이나 동물이 어떤 행동을 일으키고 그 결과로 얻어진 보수에 따라 그 행동을 실행하는 횟수를 늘리는 특성을 가리킨다.

**2) 강화학습(reinforcement learning)**

강화를 통해 시행착오를 끊임없이 반복하면서 적절한 행동을 학습해나가는 모습(시간 변화)을 수리 모델로 표현한 것을 말한다.

**3) 에이전트(agent)**

강화학습 모델이 모델화되는 데 필요한 의사결정을 하는 주체를 에이전트(agent)라고 한다. 에이전트는 내부에 수리 모델을 포함하고 그 수리 모델에 따라 결정한다. 행동이 결정된 후에는 그 행동이 잘 됐을 때는 보수를 받지만 반대로 잘 되지 않았을때는 보수를 받지 못하거나 마이너스의 보수를 받는다. 이떄 ‘행동이 잘 되는지 안되는지’를 결정하는 곳을 환경(environment)라고 한다. 에이전트는 그 다음 의사결정을 위해 그 결과를 받아 직접 의사결정 모델을 수정하고 거기서 행동을 생성해서 결과를 보는 작업을 반복한다.

**10장 다체계 모델(행위자 기반 모델)**

인간 사회나 생물 개체, 뇌와 같은 복잡한 구조를 갖춘 시스템에서는 ‘그 시스템을 구성하는 각각의 요소의 집합’이라는 논리가 본질적인 역할을 수행한다. 이처럼 각각의 요소가 큰 집합을 이룰 때 전체적으로 보이는 구성을 분석하기 위해 다체계 모델을 사용한다.

**1) 다체계 모델이란**

각각의 물체(에이전트)가 행동하는 모습을 구현하는 모델이 있다고 했을 때 이 물체를 많이 만들어서 물체들끼리 상호작용하게 한 모델을 다체계(many-body system) 모델 또는 행위자 기반 모델(agent-based model)이라고 한다.

각 요소의 미크로적인 현상과 전체적으로 발생하는 매크로적인 현상 사이의 차이를 보정하거나 미크로적인 행동은 알고 있는 상황에서 매크로적인 현상이 어떻게 움직이는지 예상하고 싶을 때 이런 모델링을 이용한다.

**2) 다체계 모델을 구성하는 요소**

다체계 모델의 본질은 요소끼리 상호작용하는 모습을 구현하는 것이다. 대표적인 상호작용 방식은 다음과 같다.

**① 각각의 요소가 동일한 강도로 상호작용한다.**

이 방식으로 상호작용하는 집단에서는 현상을 수리적으로 해석하기가 비교적 쉬운 경우가 많다.

**② 요소가 이차원이나 삼차원 공간의 ‘위치정보’를 가지고 있고 가까운 요소끼리만 상호작용한다.**

모델에 공간적인 정보가 포함돼 있을 때 (각각의 요소가 ‘존재하는 위치’가 중요한 경우) 멀리 있는 요소와는 상호작용하지 않는다고 가정한다.

**③ 요소들 사이에 특정한 연결(네트워크)이 정해져 있고 연결된 상대끼리만 상호작용한다.**

시스템 안에 있는 요소와 그 요소끼리 연결되는 방식을 모두 합쳐 네트워크라고 말한다. 그리고 각각의 연결을 링크(link)라고 하고 연결되는 각각의 요소를 노드(node)라고 한다.

**3) 시간과 공간의 이산화**

다체계 모델에서는 많은 요소에 대한 수리 모델을 동시에 움직이므로 이론적으로 해석하기 어려울 뿐만 아니라 시뮬레이션할 때도 막대한 경비가 소요될 때가 있다. 그래서 일반적으로 시간이나 공간을 이산적으로 구분해서 단순화한 후에 모델화하는데 이 작업을 이산화(discetization)라고 한다. 이산화된 시간을 시간 스텝(time step)이라고 하고 이산화된 공간의 한 칸을 셀(cell) 또는 사이트(site)라고 한다.

**4) 다체계 모델을 분석하는 방식**

① 이론 해석 방법

상호작용이나 각 요소의 움직임이 완전히 동일한 규칙을 따를 때 이론 해석이 가능하다

② 시뮬레이션 분석 방법

실제로 모델을 움직여보고 그 결과를 통해 중요한 내용을 파악하거나 현상을 예측한다.

**5) 다양한 집단 현상 모델**

**① 집단 모델**

각각의 개체가 운동하는 모습을 보고 무리가 발생하는 원인을 이해하는데 사용되는 다체계 모델이다. (ex. Vicsek 모델)

**② 동기 현상 모델**

우리의 심장은 심근세포라는 작은 세포들의 모임으로 구성돼 있다. 이 세포들이 모두 함께 동일한 타이밍에 박동함으로써 심장이 크게 박동할 수 있다. 이처럼 어떤 타이밍에서 일제히 움직이는 것을 동기(synchronization)현상이라고 한다. (ex. 쿠라모토 모델)

**③ 인간의 행동과 의사결정 모델**

‘죄수의 딜레마 게임’은 가위바위보와 같은 방식으로 플레이어가 ‘협력’ 또는 ‘배신’중 하나를 선택한다. 상대 플레이어가 하는 행동에 따라 점수를 얻을 수 있는데 이러한 게임을 반복해서 플레이했을 때의 모습을 분석하고 이 분석을 통해 인간사회에서 발생하는 협력 현상을 모델화 하는 방법이다.

**6) 상호작용 네트워크 (네트워크 해석의 기본 방법)**

**① 다른 노드와 얼마나 연결됐나**

네트워크에서 어떤 노드가 연결된 다른 노드의 개수를 *차수(degree)*라고 한다. 모든 노드에 대해 이 차수를 계산하면 차수의 출현 분포를 구할 수 있다. 이것을 *차수분포(degree distribution)*라고 한다. 이 차수분포는 네트워크의 형태를 특징짓는 중요한 요소가 된다.

같은 차수인 노드끼리 연결되기 쉬운 정도를 나타내는 지표를 *동류성(assortativity)*라고 한다.

네트워크에 있는 어떤 노드에서 다른 노드로 최단경로로 이동할 때 지나가야 하는 링크 수를 최단경로 길이라고 한다. 이 최단경로 길이를 모든 노드의 쌍에 대해 계산해서 평균 낸 값을 평균경로 길이라고 말하며 이 평균경로 길이는 네트워크상에서 이동하기 쉬운 정도를 나타내는 지표로 이용된다.

**② ‘중심성’으로 중요한 노드를 특징 짓는다**

모든 노드 쌍의 최단경로를 구했을 때 그 노드 쌍의 경로가 현재 주목하고 있는 노드를 통과한 비율을 *매개 중심성(betweenness centrality)*라고 하며 이 매개 중심성은 네트워크상에 있는 정보를 전파하거나 운송현상의 검정을 특징지을 때 자주 사용된다.

그 외에도 단순히 차수의 크기만을 보는 *차수 중심성(degree centrality)*, 다른 노드와의 거리를 보는 *근접 중섬성(closeness centrality)*이 있다.

**③ 네트워크 모델**

사람의 네트워크에서는 친구끼리 친구 관계인 경우가 있는데 이 상황을 노드와 링크로 나타내면 인간관계의 삼각형이 만들어진다. 이 삼각형이 네트워크 안에 얼마나 있는지를 계산한 값을 *클러스터 계수(clustering coefficient)*라고 한다. 클러스터 계수가 높을수록 노드가 그룹 지어진다고 말할 수 있다. 그리고 클러스터 계수가 높은(링크가 빽빽히 존재하는) 그룹끼리 적은 수의 링크로 연결된 구조를 커뮤니티 구조라고 한다.

**11장 모델을 결정하기 위해 필요한 요소**

수리 모델을 적용할 대상에 따라 접근법이 달라진다. 이해 지향형 모델링은 대상이 되는 현상이 어떻게 발생하는지를 조사하기 위한 방법이고, 응용 지향형 모델링은 현상은 중요하게 생각하지 않고 응용하는 성능을 추구하는 방법이다.

**1) 결정론적 모델 vs. 확률 모델, 통계 모델**

수리 모델은 ‘결정론적 모델인지 아닌지’를 기준으로 크게 두 가지로 나뉜다. 결정론적인 수리 모델이란 확률이라는 개념이 들어있지 않은 모델을 말하며 대표적인 것으로 방정식 모델, 상미분 방정식으로 구현되는 모델 등이 있다.

데이터 분석 분야에서는 기본적으로는 확률 통계적인 요소를 포함한 모델을 이용해야 한다. 그러나 숫자로 나타낼 수 없는 데이터의 동작이나 평균적인 동작에 관해 깊이 이해하려고 할 때 노이즈를 무시할 수 있다면 결정론적 모델 접근법도 유용하다.

**2) 이해 지향형 모델링의 요점**

이해 지향형 모델링의 최종적인 목적은 현상을 이해하고 데이터가 생성되는 규칙을 이해하는 것이므로 이해하기 쉬운 수리 모델을 만들어야 한다. 이해 지향형 모델링에서는 ‘데이터를 얼마나 설명할 수 있는지’와 ‘모델의 복잡도’가 항상 서로 타협해야 하는 관계가 된다. 이해하기 쉬운 모델을 만들려면 다음과 같은 조건을 만족해야한다.

① 매개변수의 개수가 적다

② 사용하는 함수가 간단하다

③ 모델에 포함된 각각의 요소(수리 구조, 변수, 매개변수)를 직관적으로 이해할 수 있다

④ 수리적으로 해석할 수 있다

**3) 응용 지향형 모델링의 요점**

응용 지향형 모델링에서는 달성해야 할 목표를 제대로 평가지표에 적용시키는 일이 중요하다.

**① 문제를 정의한다**

응용 지향형 모델링에서는 애당초 무엇이 ‘목적’인지를 확실히 정의하는 일이 중요하다. 문제를 정식화할 때는 일단 자신이 달성하려고 하는 목표와 문제 설정을 숫자로 나타내는 방식으로 표현할 수 있어야 한다.

**② 성능을 중시해서 모델을 선택한다**

응용 지향형 모델링에서는 이해 지향형 모델링과 달리 사용한 모델을 결정할 때 성능의 좋고 나쁨을 기준으로 한다. 성능의 우열을 비교하는 지표를 잘 살펴보는 것이 중요하다.

**③ 데이터의 성질**

수리 모델은 현실 세계에서 발생하는 현상을 직접 재현하는 것이 아니라 주어진 데이터가 생성되는 규칙을 재현하는 것이다. 사용 할 수 있는 데이터가 한 방향으로 치우치거나 오차나 결손값이 많이 포함돼 있으면 이런 요소들이 모델의 성능에 그대로 반영된다.

또한 사용할 수 있는 데이터의 차원이나 샘플의 크기에 따라서도 사용할 수 있는 모델이 달라진다. 문제나 데이터의 양에 따라 자신에게 적합한 모델을 선택해야 한다.

**12장 모델을 설계한다**

어떤 문제에 수리 모델을 사용하기로 결정했다면 수리 모델을 설계할 때 어떤 변수를 포함시킬지, 어떤 요소를 수리 구조안에 배치할지, 어디에 매개변수를 마련할지를 해당 문제에 따라 결정해야 한다.

**1) 변수를 선택한다**

수리 모델의 성능이 변하지 않는다면 포함할 변수의 개수는 적을수록 좋다. 그렇기에 모델이 투입할 변수를 선택하는 일은 매우 중요하다.

① 무엇에 대응되는지 설명할 수 없는 변수는 최대한 제거한다.

② 동일한 정보를 나타낸다고 생각되는 변수들은 대표적인 것을 남기고 제외하거나 적당한 차원축소를 실행해서 그 개수를 줄인다.

**2) 데이터 수집과 실험계획**

데이터의 퀄리티는 수리 모델의 성능에 크게 좌우한다. 대상에 관해 다양한 요인을 생각해 볼 수 있는 상황에서 데이터를 어떻게 수집할지에 관해서 실험계획법이라는 이미 확립된 방법이 있는데 실험계획법에서는 전제로 둘 수 있는 조건들의 조합 중에 어느 조건을 몇 번, 어떤 순서로 정리해서 실시할지 검토한다. 그리고 분산분석(ANOVA)라는 통계적인 기법을 사용해서 각각의 요인이 주는 영향을 평가한다.

**① 피셔의 삼원칙**

이 개념은 현재 주목하는 요인 이외의 요인에 의해 발생되는 데이터의 치우침을 조정할 수 있다.

a) 반복(replication)

동일한 조건에서 여러 번 반복해서 관측하는 방법이다. 신뢰할 수 있는 평균값을 구할 수 있을 뿐만 아니라 측정오차의 크기를 미리 가늠해 볼 수 있으므로 각각의 요인이 어떻게 영향을 주는지를 통계적으로 해석할 때 중요한 실마리가 된다.

b) 무작위법(randomization)

관측하는 순서나 장소를 무작위로 결정하고 대상을 할당하는 방식도 무작위로 결정하는 방법이다.

c) 국소 관리(local control)

현재 주목하고 있지 않은 비본질적인 영향은 측정작업 전체에서 균일해야 이상적이라고 말할 수 있는데, 그렇지 않은 경우에도 일부분 정도는 균일하다고 간주할 때가 있다. 이런 경우에는 균일하다고 간주할 수 있는 몇 개의 블록으로 데이터를 나눠서 관측한다.

**3) 수리 구조와 매개변수 선택**

**① 목적변수의 분산을 무시할 수 없는 경우**

목적변수의 움직임에서 보이는 확률적인 분산이 본질적인지 또는 무시할 수 있는지 생각해 본다. 분산이 본질적인 경우에 수리 모델은 목적변수의 확률적인 움직임을 재현하는 것을 목표로 하고 이 경우에도 분산된 정도가 다른 변수의 영향보다 그다지 크지 않을 때는 시계열 모델링을 통해 확률적으로 일정한 정확도를 가지고 값을 예측할 수 있다.

**② 분산을 생각하지 않아도 되는 경우**

분산을 무시해도 되는 경우에는 목적변수를 설명변수로 나타낸 함수를 구하는 것을 목표로 한다.

**③ 매개변수 값의 범위**

변수와 수리 구조를 결정하면 매개변수가 필요한 곳이 자연스럽게 결정된다. 매개변수는 일반적으로 연속값으로 설정하지만, 처음부터 의미가 부여된 경우에는 매개변수의 범위를 정할 때 주의해야 한다. 추정된 매개변수의 값을 포함해서 데이터와 모델링에 일관성이 있는지를 검사할 수 있게 매개변수가 취할 수 있는 값의 범위를 확인해야 한다.

**4) 모델링에 실패하지 않으려면**

각 분야마다 따라야 할 기존의 체계나 일관성이 있다. 모델을 새로 만들거나 기존 모델을 확장할 경우에는 이런 체계, 법칙과 모델이 일관성을 갖게 만든다.

**13장 매개변수를 추정한다**

수리 모델이 완성된 후에는 매개변수의 값을 데이터에 잘 맞게 조정해야한다. 이 매개변수를 결정하는 방법에는 현재 사용하는 모델이나 문제 설정에 따라 다양한 방법이 있다.

**1) 목적에 맞는 매개변수 추정**

모델을 최대한 정확하게 데이터에 맞추고 싶다면 모델에서 생성되는 값과 실제 데이터와의 차이를 최소화해야 한다. 이 차이를 계산하기 위해 사용되는 지표를 목적함수라고 하며 이것을 최소화해서 매개변수 값을 추정한다. 모델의 형태나 문제 설정에 따라 다양한 목적함수를 사용할 수 있는데, 그중 대표적인 것으로는 평균제곱오차(MSE, mean squared error)와 대수우도(log likelihood)가 있다.

**① 평균제곱오차(MSE, mean squared error)**

**② 대수우도(log likelihood)**

하나의 모델이 주어졌을 때 데이터에 있는 모든 관측값 X가 해당 모델에서 출현할 확률을 계산하고 이 양을 우도(likelihood)라고 한다. 이 우도를 최대화하는 매개변수의 집합이 가장 데이터를 잘 나타내는 모델을 제공한다. 이처럼 우도를 목적함수로 해서 매개변수를 추정하는 방법을 최우추정법(MLE, maximum likelihood estimation)이라고 한다.

**③ 교차 엔트로피(cross entropy)**

정보량이라는 관점에서 두 분포가 얼마나 가까운지를 정량화하는 지표로 교차 엔트로피가 있다. 이는 분류 문제에서 목적함수로 이용된다.

**2) 추정에서 목적함수를 최소화하려면?**

**① 해석적으로 푼다**

매개변수에 대한 목적함수를 간단한 식의 형태로 쓸 수 있다면 수식을 계산해서 매개변수를 구할 수 있다.

**② 매개변수를 스윕한다**

목적함수를 해석적으로 최소화할 수 없는 경우 매개변수의 개수가 적고(열 개 이하) 조사하려는 범위가 크지 않을 경우에는 단순히 모든 경우에 대해 목적함수의 값을 계산하는 방법이 있다.

머신러닝 분야에서 하이퍼매개변수를 이런 방법으로 조정하는 작업을 *그리드 서치(grid search)*라고 한다. 이렇게 매개변수 값을 모두 시험해 보는 방법을 *매개변수 스윕(parameter sweep)*이라고 한다.

**③ 경사하강법**

목적함수를 구체적인 수식으로 계산하고 해당 매개변수로 미분을 계산할 수 있을 때 사용하는 방법이다. 이 방법은 매개변수의 값을 목적함수가 줄어드는 방향으로 조금씩 움직여가는 알고리즘이다.

**3) 베이즈 추정과 베이즈 모델링**

같은 모델이나 대상에서 데이터를 여러 번 가져와 각각의 샘플로 매개변수를 추정하면 추정된 값은 분산될 것이다. 이렇게 매개변수가 분산돼 있다고 생각하고, 즉 매개변수가 확률분포를 따르고 있다고 생각하고 수리 모델을 구축한 후에 매개변수를 추정하는 방법을 베이즈 모델링, 베이즈 추정이라고 한다. 베이즈 추정은 데이터에서 매개변수의 확률분포를 구하는 것을 목표로 한다.

베이즈 모델링에서는 매개변수의 값을 한 개 결정하면 그 매개변수의 값을 가지는 모델이 발생할 확률이 정해진다. 이 확률을 사용해 다양한 매개변수의 값에 대한 모델의 기댓값을 구한 것이 최종적으로 추정되는 예측 모델이 된다.

**14장 모델을 평가한다**

수리 모델은 여러 번 시행착오를 겪은 후에 사용할 모델을 최종적으로 완성한다. 그렇게 하려면 ‘좋은’ 모델을 선택하기 위한 지표가 필요하다.

**1) ‘좋은 모델’이란**

수리 모델링에서는 목적에 따라 다른 접근법이 필요하고 ‘좋은’ 모델을 평가하기 위한 개념과 지표도 목적에 따라 다르다.

**① 메커니즘의 이해를 목적으로 한 모델 평가**

메커니즘을 이해하는 것을 목적으로 삼고 수리 모델링을 하는 경우 대상이 되는 데이터가 생성되는 메커니즘을 ‘이해’할 수 있는지가 중요하다.

모델을 해석 할 수 있는가? 모델이 데이터에 잘 맞고 일관성이 있는가?

모델을 복잡하게 만들면 데이터에 잘 맞게 되지만, 해석하기가 어려워진다. 그렇기에 어느 요소까지 고려해야 할지를 데이터에 맞는 정도와 함께 균형을 맞춰 검토한다.

**② 통계적 추론을 실행하기 위한 모델 평가**

통계 모델링을 통해 통계적 추론을 실행할 때는 원리를 이해할 때보다 추론할 때에 데이터의 적합성이 더 중요하다. 따라서 모델이 과대 적합 될 위험성도 고려해야 한다. 그래서 모델을 선택할 때는 같은 타입의 모델을 사용하고 여기에 사용할 변수를 선택하고 사용할 변수들 사이의 관계식을 바꿔가면서 모델의 해석성 자체는 변화하지 않는 범위에서 검토한다.

**③ 응용 지향형 모델링에서 모델을 평가**

응용 지향형 모델링에서는 응용할 때 발휘되는 성능이 좋은지 나쁜지가 중요하다. 즉, 모델에 미지의 데이터가 얼마나 잘 맞는지 아닌지만이 중요하다.

**2) 분류 정확도를 측정하는 지표**

① 적합성과 성능을 측정한다

적합도를 나타내는 지표로는 평균제곱오차, 대수우도 또는 등이 있다.

② 정답률, 재현율, 특이도, 적합률, F값

③ ROC 곡선과 AUC

**3) 정보량 규준**

**① 아카이케 정보량 규준 (AIC)**

가장 많이 사용되는 지표이며 통계 모델링 분야에 자주 이용된다. AIC를 계산하고 그 결과가 작은 모델을 선택한다. 모델이 같은 규칙으로 생성된 미지의 데이터를 얼마나 설명할 수 있는지(예측할 수 있는지)를 기반으로 해서 고안됐다.

**② 베이즈 정보량 규준 (BIC)**

베이즈 통계에서 주변우도를 최대화한다는 아이디어를 기반으로 했다는 점이 AIC와 다르다.

**4) 널 모델과 비교, 우도비 검정**

어떤 모델이 다른 모델에 포함된(중첩된) 경우에 이 모델을 비교하는 방법을 생각해보자. 이 분석에는 더 적합도가 좋은 모델을 찾는다는 목적도 있고 대상이 되는 시스템에 있는 요소가 중요한지를 검증한다는 목적도 있다. 방법으로는 널 모델과 우도비 검정이 있다.

**5) 교차검증**

추정된 모델이 미지의 데이터를 설명하는 능력을 측정하기 위해 데이터를 두 가지로 나누며 이 둘을 완전한 독립 데이터로 간주해야 성능을 공평하게 평가할 수 있다.

**① 홀드아웃 검증(hold-out validation)**

훈련 데이터와 테스트 데이터라는 두 가지로 나눠서 성을 평가하는 방법이다.

**② K-분할 교차검증(K-fold cross validation)**

전체 데이터를 K개의 블록으로 분할하고 그중에서 K-1개의 블록을 사용해 모델을 학습한다. 그리고 남은 미사용 블록 한 개를 모델의 성능을 평가하는 데 사용 하는 방법이다.

**③ leave-one-out 교차검증(LOOCV)**

분할하는 개수를 최대한으로 늘리고 테스트용 데이터를 한 개의 샘플만 남기고 교차검층하는 방법을 말한다.