밑바닥부터 시작하는 데이터 과학

3회차

- 가설 검정
- 최적화
- 차원축소
- Pandas

1. 가설 검정

> **검정(testing)**
>
> 데이터 뒤에 숨어있는 확률 변수의 분포와 모수에 대한 가설의 진위를 정량적(quantitatively)으로 증명하는 작업.

1. 가설 검정

가설 검정으로 접근할 수 있는 문제들..

문제 1.

> 어떤 동전을 15번 던졌더니 12번이 앞면이 나왔다.
> 이 동전은 휘어지지 않은 공정한 동전(fair coin)인가?

문제 2.

> 어떤 트레이더의 일주일 수익률은 다음과 같다.
> -2.5%, -5%, 4.3%, -3.7% -5.6%
> 이 트레이더는 계속해서 돈을 잃을 사람인가?

1. 가설 검정

가설 검정의 기본 논리

> **1.** 데이터가 어떤 고정된(fixed) 확률 분포를 가지는 **확률 변수**라고 가정한다.

> **2.** 이 확률 분포의 모수값이 특정한 값을 가진다고 가정한다. 이 때 모수가 가지는 특정한 값은 우리가 검증하고자 하는 사실과 관련이 있어야 한다. 이러한 가정을 **귀무 가설(null hypothesis)**이라고 한다.

> **3.** 만약 데이터가 주어진 귀무 가설에 따른 표본이고 이 표본 데이터를 특정한 수식에 따라 계산한 숫자는 특정한 확률 분포를 따르게 된다. 이 숫자를 **검정 통계량(test statistics)**라고 하며 검정 통계량의 확률 분포를 **검정 통계 분포(test statistics distribution)**라고 한다. 검정 통계 분포의 종류 및 모수의 값은 처음에 정한 가설 및 수식에 의해 결정된다.

1. 가설 검정

가설 검정의 기본 논리

> **4.** 주어진 귀무 가설이 맞으면서도 표본 데이터에 의해서 실제로 계산된 검정통계량의 값과 같은 혹은 그보다 더 극단적인(extreme) 또는 더 희귀한(rare) 값이 나올 수 있는 확률을 계산한다. 이를 **유의 확률(p-value)**이라고 한다.

> **5.** 만약 유의 확률이 미리 정한 특정한 기준값보다 작은 경우를 생각하자. 이 기준값을 **유의 수준(significance level)**이라고 하는 데 보통 1% 혹은 5% 정도의 작은 값을 지정한다. 유의 확률이 유의 수준으로 정한 값(예 1%)보다도 작다는 말은 해당 검정 통계 분포에서 이 검정 통계치(혹은 더 극단적인 경우)가 나올 수 있는 확률이 아주 작다는 의미이므로 가장 근본이 되는 가설 즉, 귀무 가설이 틀렸다는 의미이다. 따라서 이 경우에는 귀무 가설을 **기각(reject)**한다.

> **6.** 만약 유의 확률이 유의 수준보다 크다면 해당 검정 통계 분포에서 이 검정 통계치가 나오는 것이 불가능하지만은 않다는 의미이므로 귀무 가설을 기각할 수 없다. 따라서 이 경우에는 귀무 가설을 **채택(accept)**한다.

2. 최적화

최적화(optimization)

주어진 함수(오차함수)의 최댓값 혹은 최솟값을 찾는 문제

2. 최적화

최적화 문제를 풀기 위한 기본 개념

미분공식

(1) 상수

> - 상수를 미분하면 0이 된다.
> - $\frac{d}{dx}(c) = 0$

(2) 거듭제곱

> - x의 n 제곱을 미분하면 $n-1$ 제곱으로 제곱수가 1씩 감소한다.
> - 이 공식은 n이 자연수이거나 음의 정수일 때 성립한다. $n = 0$일 때는 성립하지 않는다.
> - $\frac{d}{dx}(x^n) = nx^{n-1}$

(3) 로그

> - 로그함수를 미분하면 x^{-1} 이 된다.
> - $\frac{d}{dx}(\log x) = \frac{1}{x}$

(4) 지수

> - 밑이 오일러 수인 지수함수는 미분해도 변하지 않는다.
> - $\frac{d}{dx}(e^x) = e^x$

2. 최적화

최적화 문제를 풀기 위한 기본 개념

편미분 (partial differentiation)

다변수 함수에서 특정 변수를 제외한 나머지 변수를 상수로 생각하여 미분하는 방식.

(1) 편미분 표기 방법

$$f_x(x,y) = \frac{\partial f}{\partial x}$$

$$f_y(x,y) = \frac{\partial f}{\partial y}$$

(2) 편미분 간단 예시

$$f(x,y) = x^2 + 4xy + 4y^2$$

$$f_x(x,y) = \frac{\partial f}{\partial x} = 2x + 4y$$

$$f_y(x,y) = \frac{\partial f}{\partial y} = 4x + 8y$$

2. 최적화

최적화 문제를 풀기 위한 기본 개념

그레디언트 (Gradient)

다변수 함수의 편미분 벡터

(1) 그레디언트 표기 방법

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \cdots \frac{\partial f}{\partial x_n} \right)$$

- 각 변수의 1차 편미분 값들로 구성된 벡터

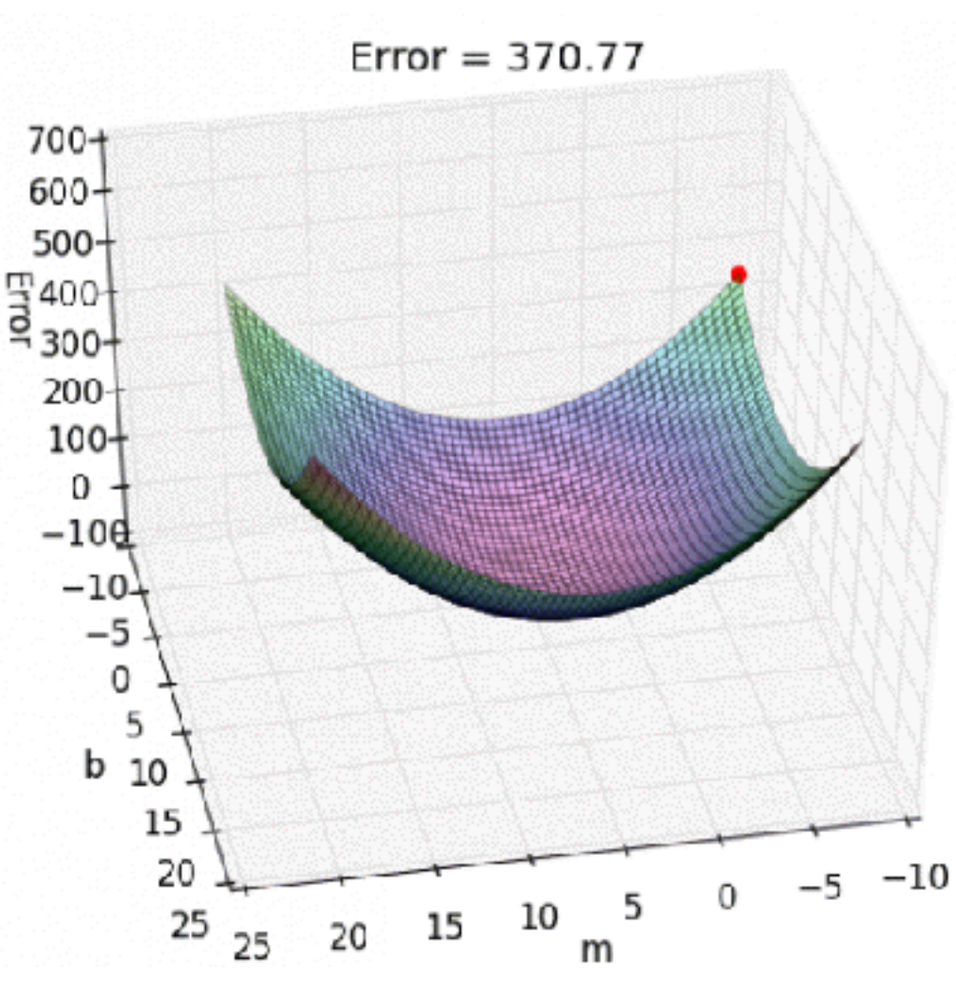
(2) 그레디언트 의미

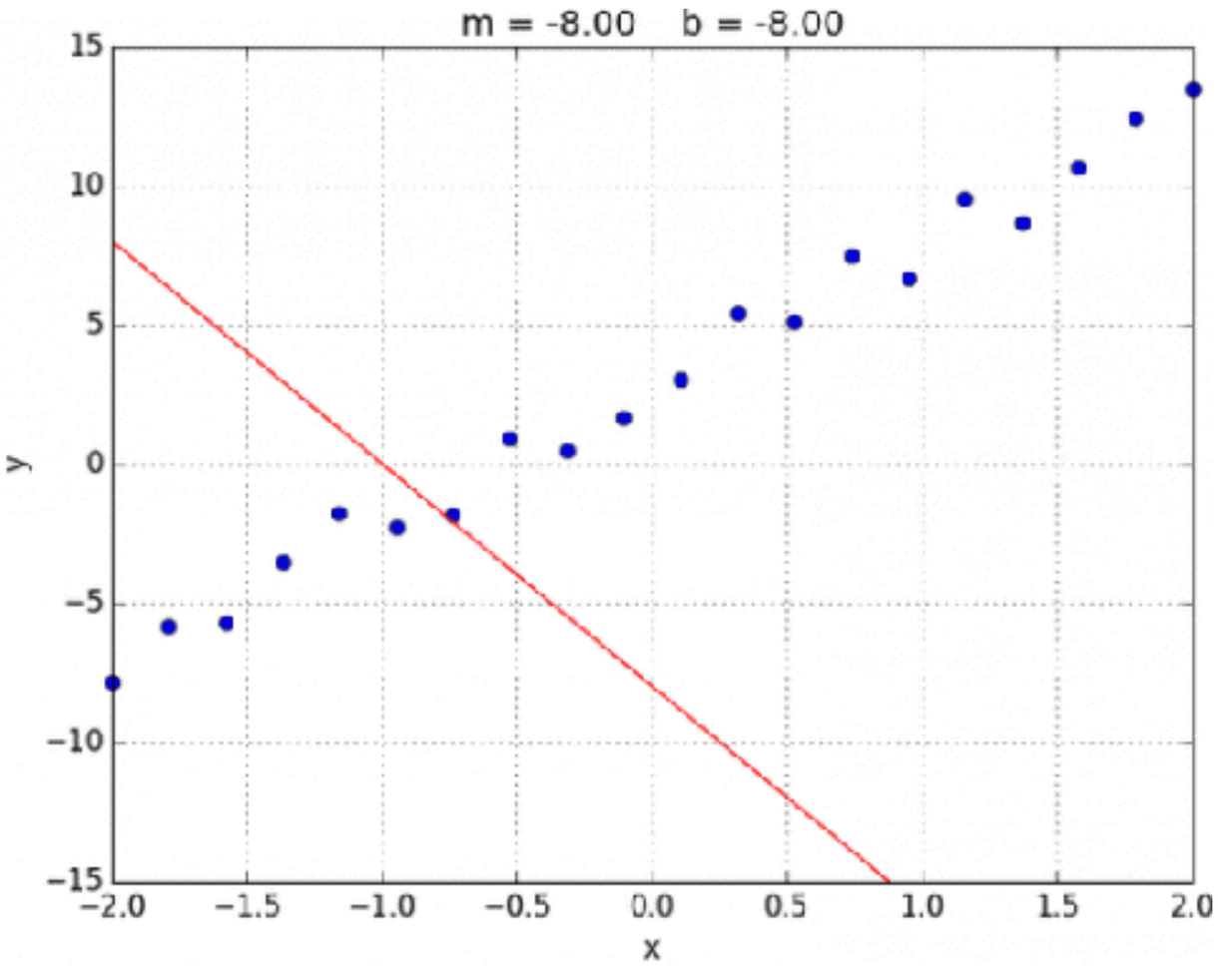
- 스칼라장의 최대의 증가율을 나타내는 벡터장을 뜻한다.
- **산이나 언덕**을 가정해보자. **어떤 지점(x,y)**에서의 **높이를 H(x,y)**로 표현하는 경우, 그레디언트는 가장 (위를 바라보는)경사가 가파른 방향과 그 경사의 크기를 나타낸다.
- **어떤 지점 : 스칼라, 가파른 경사의 방향과 크기 : 그레디언트**

2. 최적화

경사하강법(Gradient Descent)

> 반복적 시행 착오(trial and error)에 의해 최적화 필요조건을 만족하는 값을 찾는 방법인 수치적 최적화(numerical optimization) 방법 중 하나.
> (최적화 필요조건 : 편미분값 = 0 (최대 혹은 최소가 되는 지점))





3. 차원축소

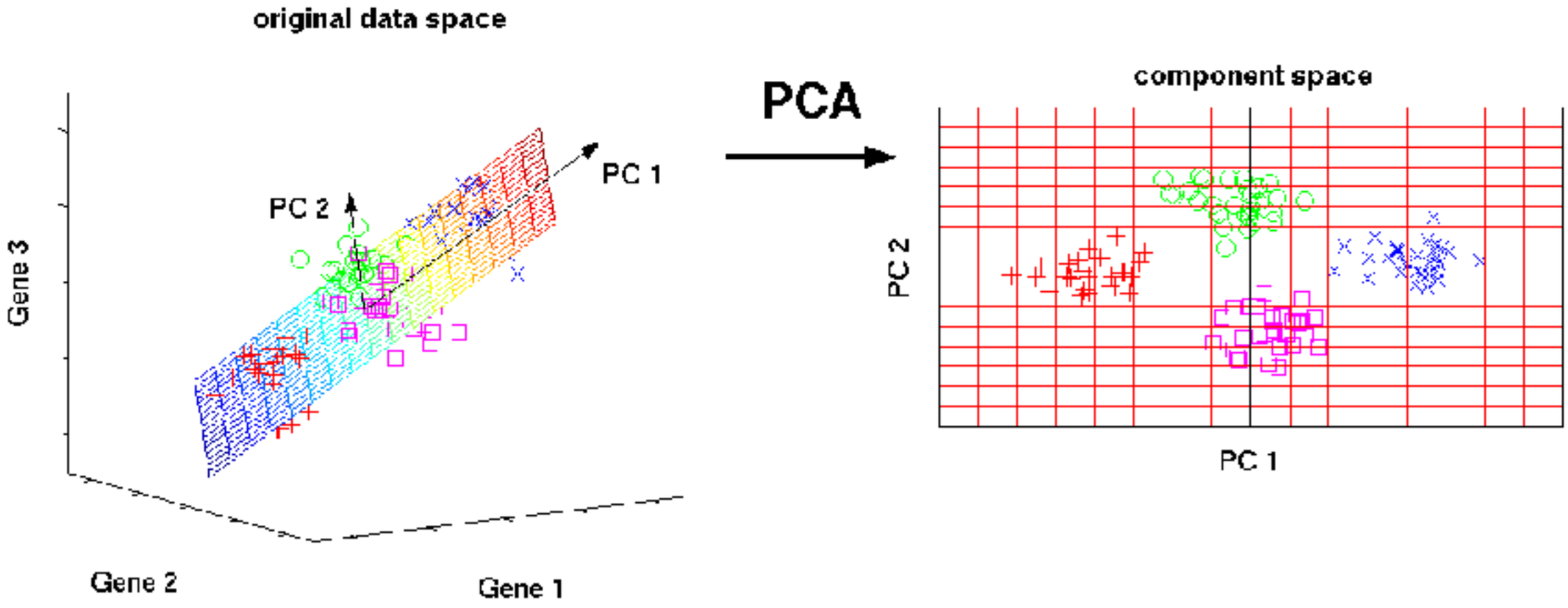
차원 축소 (Dimensionality Reduction)

데이터의 의미를 제대로 표현하는 특징을 추려내는 것

3. 차원축소

주성분 분석 (PCA)

> **주성분분석**은 **데이터의 분포**를 가장 잘 표현하는 성분을 찾아주는 것이다.
> 데이터에서 분산이 가장 큰 방향 벡터를 구한 뒤, 원래 데이터에 projection 해서 차원을 낮춘다.



4. Pandas

Pandas

테이블 형태의 데이터를 다루기 위한 데이터프레임(DataFrame) 자료형을 제공한다.
자료의 탐색이나 정리에 아주 유용하여 데이터 분석에 빠질 수 없는 필수 패키지이다.

과제

차원 축소 구현

- 137p ~ 142p

다음에
또!
같이!
만나요!

learnit Flip your study