#통계학과 수학 관련 면접 Q&A

#1 고유값(eigen value)와 고유벡터(eigen vector)이 무엇이고 왜 중요한지 설명해주세요.

정방행렬 $(n ₩times n)$인 A는 임의의 벡터 $(n ₩times 1)$인 x의 방향과 크기를 변화시킬 수 있다.

수많은 벡터 x중 어떤 벡터들은 A에 의해 선형 변환되었을 때에도 원래 벡터와 평행한 경우가 있다. **이렇듯 "Ax = λx의 형태로 표현되며, 이 때 x를 고유벡터, λ를 고유값이라 한다."**

고유값과 고유벡터를 통해 A를 고유값과 고유벡터들로 분해하는 고**유값 분해(eigen decomposition)**, 정방행렬 뿐만 아닌 $m ₩times n$행렬도 분해할 수 있는 **특이값 분해(SVD)**, 데이터들을 차원 축소시킬 때 가장 원래 의미를 잘 보존시키는 **주성분 분석(PCA)**등에 활용할 수 있으므로 중요하다.

#2 샘플링(Sampling)과 리샘플링(Resampling)이 무엇이고 리샘플링의 장점을 말씀해주세요.

샘플링이란 **표본추출**을 의미하는 것으로, 모집단 전체에 대한 추정치(estimate)를 얻기 위해 임의의 sample을 뽑아내는 것이다. 모집단 전체에 대한 조사는 불가능하기 때문에 sample을 이용하여 모집단에 대한 추론(inference)을 하게되는 것이다. 하지만 표본은 모집단을 닮은 모집단의 mirror image 같은 존재이지만, 모집단 그 자체일수는 없다. 따라서 표본에는 반드시 모집단의 원래 패턴에서 놓친 부분, 즉 **noise가 존재할 수 밖에 없다.**

리샘플링은 **모집단의 분포 형태를 알 수 없을 때 주로 사용하는 방법이다.** 즉, 모분포를 알 수 없으므로 일반적인 통계적 공식들을 사용하기 힘들 때, 현재 갖고 있는 데이터를 이용하여 모분포와 비슷할 것으로 추정되는 분포를 만들어 보자는 것이다. **리샘플링은 표본에서 데이터를 다시 추출하여 다양한 통계적 분석을 가능하게 합니다.** 즉, 같은 샘플을 여러 번 사용해서 성능을 측정하는 방식이다. 가장 많이 사용되는 방법이며 종류로는 K-fold 교차 검증, 부트스트래핑이 있다.

리샘플링은 표본을 추출하면서 원래 데이터 셋을 복원하기 때문에 이를 통해서 모집단의 분포에 어떤 가정도 필요 없이 표본만으로 추론이 가능하다는 장점이 있다.

#3 확률 모형과 확률 변수는 무엇인가요?

확률변수는 표본 공간의 사건을 실수 값으로 매핑하는 함수로, 대문자 X로 표기됩니다. 무작위(Random) 실험을 했을 때, 특정 확률로 발생하는 각각의 결과를 수치적 값으로 표현하는 변수라고 할 수 있다. 또한 확률 변수에는 이산확률변수, 연속확률변수두가지 경우가 있다. 이산확률변수는 확률변수 X가 취할 수 있는 값이 유한하기 때문에 셀 수 있는 확률변수이다. 반면에 연속확률변수는 어떠한 두 수 사이에 반드시 다른 수가 존재하는, 셀 수 없는 범위의 확률변수를 가지는 경우에 사용된다.

주사위 굴리기 예제를 생각해보자.

일단 주사위를 굴리는 상황은 어떤 수가 나올지 모르므로, 확률상황이다. "주사위를 굴렸을 때 나오는 값"을 확률변수 X라고 할 수 있다. 1~6이 표본공간이 되고, 셀 수 있으므로 이산확률변수가 된다. P(X=1)와 같은 식으로 표현하고, 이는 "주사위를 굴렸을 때, 1이라는 값이 나올 확률"로 해석할 수 있다.

확률모형(Probability Model)이란 확률변수를 이용하여 데이터의 분포를 수학적으로 정의한 모형이다. 데이터 분포를 묘사하기 위해서 사용된다. 보통 **확률 분포 함수(probability distribution function)**또는 **확률 밀도 함수(probability density function)**를 주로 사용하며, 이때 함수의 계수를 분포의 모수(parameter)라고 부른다. **확률분포(Probability Distribution)**란 표본공간에 정의된 확률을 이용하여 확률변수의 값 또는 영역에 대한 확률을 표현한 것이다.

다음과 같은 함수들이 확률모형에 포함될 수 있다.

· 확률질량함수(PMF, Probability Mass Function) - 이산형
· 확률밀도함수(PDF, Probability Density Function) - 연속형
· 누적분포함수(CDF, Cumulative Distribution Function)

추가적으로 **확률 통계의 기초 용어**를 정리하면 다음과 같다. (주사위 굴리기 예제 사용)

- 실험(Experiment)은 하나의 행위가 하나 이상의 결과를 도출하는 것에 대한 과정 혹은 절차를 나타낸다. - 예시) 주사위를 던진다. - 결과(Outcome)는 어떤 실험에 의해 발생 가능한 결과이다. 특정 실험의 가능한 결과들은 각각 유일(unique)하다. 한번의 실험을 시행했을 때, 단 하나의 outcome만을 나타낸다. - 예시) 주사위의 눈 (ex. 3, 4, 6) - 표본 공간(Sample space)은 확률 실험에서 발생할 수 있

는 모든 결과로 구성된 집합(set)이다. 발생할 수 있는 모든 결과의 집합이므로, 중복된 원소를 가질 수 있다. - 예시) 가능한 주사위의 모든 눈 집합 (ex. Ω = {1, 2, 3, 4, 5, 6}) - 사건(Event)은 우리가 관심있는 Sample space의 부분집합이다. - 예시) 주사위 눈이 3이 나온다, 짝수/홀수가 나온다.

#4 누적 분포 함수와 확률 밀도 함수는 무엇인가요? 수식과 함께 표현해주세요.

확률 변수 X가 임의의 실수 집합 B에 포함되는 사건의 확률이 다음과 같이 어떤 음이 아닌 함수 f의 적분으로 주어진다고 하자.

이 때의 X를 연속확률변수라고 하며, 함수 $f(x)$를 **확률 밀도 함수(Probability Density Function, PDF)**라고 한다. 단, 실수 집합 B가 실수 전체일 경우 실수 전체에 대한 확률밀도함수의 적분은 1을 만족해야 한다.

누적 분포 함수(Cumulative Distribution Function, CDF)는 확률변수가 특정 값보다 작거나 같을 확률을 나타내는 함수이다.

확률 밀도 함수와 누적 분포 함수는 **미분과 적분의 관계**를 갖는다. 확률 밀도 함수를 음의 무한대에서 특정값 a까지 적분을 하면, a에 대한 누적 분포 함수를 얻을 수 있다. 반대로 누적 분포 함수를 미분하면 확률 밀도 함수를 얻을 수 있다.

#5 조건부 확률은 무엇인가요?

조건부 확률은 사건 A가 일어났다는 전제 하에 사건 B가 일어날 확률이다. 조건부 확률은 **베이즈 정리**와도 이어지며, 조건부 확률은 $P(B|A) = P(A \cap B) / P(A)$로 정의된다.

베이즈 정리를 통해 가능도(Likelihood)와 증거(Evidence)를 바탕으로 사전확률을 사후확률로 업데이트한다.

- **D**: 새로 관찰되는 데이터
- **$₩theta$**: 모델에서 계산하고 싶어하는 모수 (가설)
- 사후확률(Posterior): 데이터를 관찰했을 때, 이 가설이 성립할 확률 (데이터 관찰 이후 측정하기 때문에 사후확률)
- 사전확률(Prior): 가설에 대해 사전에 세운 확률 (데이터 관측 이후 사후확률이 사전확률이 된다.)
- 가능도(Likelihood): 현재 주어진 모수 (가정) 에서 이 데이터가 관찰될 가능성
- 증거(Evidence): 데이터 전체의 분포

#6 공분산과 상관계수는 무엇일까요? 수식과 함께 표현해주세요.

공분산은 확률변수 X의 편차(평균으로부터 얼마나 떨어져 있는지)와 확률변수 Y의 편차를 곱한 것의 평균값이다.

공분산은 두 변수 간에 양의 상관관계가 있는지, 음의 상관관계가 있는지 정도를 알려준다. 하지만 상관관계가 얼마나 큰지는 제대로 반영하지 못한다.

공분산의 문제는 확률변수의 단위 크기에 영향을 많이 받는다는 것이다. 이를 보완할 수 있는 것이 바로 상관계수이다.

상관계수는 양의 상관관계가 있는지 음의 상관관계가 있는지 알려줄 뿐만 아니라, 그 상관성이 얼마나 큰지도 알려준다. 1 또는 -1에 가까울수록 상관성이 큰 것이고, 0에 가까울수록 상관성이 작은 것이다.

#7 신뢰 구간의 정의는 무엇인가요?

구간 추정에서 모수가 **a 에서 b 사이에 있을 것으로 추정(신뢰구간)**하고 **그 확률(%, 신뢰수준)**을 구한다.

신뢰구간(Confidence Interval)은 모집단의 모수(parameter)가 위치해 있을 것으로 신뢰할 수 있는 구간이다. 모수가 어느 범위 안에 있는지를 확률적으로 보여주는 방법이라고 할 수 있다. 신뢰구간을 구하는 이유는 모수의 신뢰성을 가늠하기 위함이다.

추가적으로, 신뢰구간에 대한 정확한 해석은 모평균을 포함할 확률이 95%가 되는 구간이 아닌, 같은 방법으로 100번 표본을 추출했을 때, 함께 계산되는 100개의 신뢰구간 중 모평균을 포함한 신뢰구간들의 숫자가 95개정도 된다라고 해야한다. 왜냐면, 모평균은 이미 정해져 있는 값이므로 전자의 해석을 사용할 수 없기 때문이다.

신뢰수준은 방법의 정확도, 참값을 구하기 위한 작업을 많이 반복했을 때, 참값이 특정 범위에 있는 비율이다.

모수(Parameter)는 모집단의 특성을 보여주는 값이다. 예를들어, 평균, 분산 등의 고정인 값이 있을 수 있다.

#8 p-value를 모르는 사람에게 설명한다면 어떻게 설명하실 건가요?

p-value는 귀무가설이 참일 때, 관찰된 데이터와 같거나 더 극단적인 데이터가 나올 확률을 나타낸다. 귀무가설이란 기존의 주장을 말하며, 이와 반대로 새로운 주장을 대립가설이라고 한다.

예를 들어, 어느 제약회사에서 치료약 A를 개발했다. 기존에는 치료약 A가 없었으므로 귀무가설은 "치료약 A가 효과가 없다"라고 설정한다. 반대로 대립가설은 "치료약 A는 효과가 있다"로 설정한다. 회사에서는 검정을 한 결과, 귀무가설을 기각하고 대립가설을 채택했다. 치료약 A는 판매되었고 높은 매출을 기록했다. 그런데 알고보니 치료약 A가 효과가 없다는 것이 밝혀졌다. 참인 귀무가설을 기각했기에 이는 1종 오류가 일어났다고 볼 수 있다.

다시 돌아와서 p-value는 **1종 오류를 범할 확률**을 말한다. 예를 들어, p-value가 5%라면, 100번 중 5번 1종 오류가 발생한다는 말이다. 검정을 할 때는 유의 수준 $₩alpha$를 정하는데, 이것이 1종 오류의 상한선이 된다. 그래서 유의 수준보다 p-value가 작다면 실험의 오류가 상한선보다 작으므로 귀무가설을 기각하고 대립가설을 채택한다. 만약 크다면 상한선을 넘었으므로 귀무가설을 채택한다.

#9 R square의 의미는 무엇인가요?

결정계수(R square)는 선형 회귀 모델에서 데이터에 대해 회귀선이 얼마나 잘 설명하는지에 대한 설명력을 의미한다. 결정계수는 0~1 의 값을 가질 수 있고, 만약 값이 1 이라면 회귀선으로 모든 데이터를 다 설명할 수 있다고 이해할 수 있다.

관측값은 실제 데이터의 값을 말하며, 추정값은 회귀 모델을 통해 나온 값을 말한다. 회귀 모델의 성능을 평가하는 방법은 결정계수 외에도 MAE, MSE, RMSE 가 있다.

#10

평균(mean)과 중앙값(median)중에 어떤 케이스에서 뭐를 써야할까요?

· 평균(mean): 모든 관측값의 합을 자료의 개수로 나눈 것
· 중앙값(median): 전체 관측값을 크기 순서로 배열했을 때 가운데 위치하는 값

평균은 전체 관측값이 골고루 반영되므로 대표값으로서 가치가 있다. 평균 근처에 **표본이 몰려 있는 상황에서 대표값으로 유용**하지만 극단적인 값에 영향을 많이 받는다.

중앙값에서는 관측값을 크기 순서로 배열할 때 관측값의 위치가 중요하고, 가운데 위치한 관측값 이외의 관측값들의 크기는 중요하지 않다. 따라서 평균과는 달리 중

앙값은 관측값들의 변화에 민감하지 않고 특히 아주 큰 관측값이나 아주 작은 관측값(즉, outlier)에 영향을 받지 않는다. 중앙값이 유용한 경우는 **표본의 편차, 혹은 왜곡이 심하게 나타나는 경우**이다.

#11 중심극한정리는 왜 유용한걸까요?

중심극한정리란 크기가 n인 표본 크기가 충분히 클 경우, 표본 평균의 분포가 정규분포에 수렴한다는 것이다. 중심극한정리가 유용한 이유는 **모집단의 형태가 어떻든지 간에 상관없이 표본 평균의 분포가 정규분포를 따르기 때문**이다.

#12 엔트로피(Entropy)에 대해 설명해주세요. 가능하면 정보이득(Information Gain)도요.

엔트로피는 주어진 데이터의 혼잡도를 의미하며, 엔트로피는 다음과 같이 데이터가 어떤 클래스에 속할 확률에 대한 기댓값으로 표현할 수 있다.

엔트로피는 데이터가 서로 다른 클래스에 속하면 높고, 같은 클래스에 속하면 낮다. 다시 말하면 각각의 데이터가 특정 클래스에 속할 확률이 높고 나머지 클래스에 속할 확률이 낮다면 엔트로피가 낮고, 모든 각각의 클래스에 속할 확률이 비슷하다면 엔트로피는 높다.

정보이득은 데이터가 어떤 클래스에 속할 확률이 커짐에 따라 정보를 잘 얻게되는 것을 말하며, 감소되는 엔트로피 양을 의미한다. 수식으로는 기존 시스템의 엔트로피에서 현재 엔트로피를 뺀 값으로 표현된다. 의사결정트리는 가지를 칠 때 이 값을 사용하여 가지를 친다. 이 때 어떤 데이터를 두 집합으로 나누었을 때 두 집합의 정보이득이 크도록, 엔트로피는 작아지도록 분할을 한다.

#13 어떨 때 모수적 방법론을 쓸 수 있고, 어떨 때 비모수적 방법론을 쓸 수 있나요?

표본의 통계량(평균, 표준편차 등)을 통해 모집단의 모수(모평균, 모표준편차 등)를 추정하는 방법을 통계적 추론이라고 한다.

모집단이 어떤 분포를 따른다는 가정 하에 통계적 추론을 하는 방법을 모수적 방법이라 하는데, 표본의 수가 30개 이상일 때 중심극한 정리에 의해 정규분포를 따르므로 **모수적 방법론**을 사용한다.

반대로, 모집단의 분포를 가정하지 않는 비모수적 방법은, 표본의 수가 30개 미만이거나 정규성 검정에서 정규 분포를 따르지 않는다고 증명되는 경우 **비모수적 방**

법론을 사용한다.

#14 “likelihood”와 “probability”의 차이는 무엇일까요?

확률(Probability)은 어떤 시행(trial)에서 특정 결과(sample)가 나올 가능성을 말한다. 즉, 시행 전 모든 경우의 수의 가능성은 정해져 있으며 그 총합은 1(100%)이다.

가능도(Likelihood)은 어떤 시행(trial)을 충분히 수행한 뒤 그 결과(sample)를 토대로 경우의 수의 가능성을 도출하는 것을 말한다. 아무리 충분히 수행해도 어디까지나 추론(inference)이기 때문에 가능성의 합이 1이 되지 않을수도 있다.

PDF(probability density function)에서는 **확률변수**를 변수로 보기 때문에 총합이 1이지만, likelihood function에서는 **분포의 모수**를 변수로 보기 때문에 총합이 1이 되지 않을수도 있다.

#15 통계에서 사용되는 bootstrap의 의미는 무엇인가요.

부트스트랩(Bootstrap)은 가설검증을 하거나 metric을 계산하기 전에 random sampling을 적용하는 방법이다. 모수의 분포를 추정하는 방법 중 하나는, 현재 가진 표본에서 추가적으로 표본을 복원추출하고 각 표본에 대한 통계량을 다시 계산하는 것이다. 부트스트랩이 여기에 해당하며, 여러번의 무작위 추출을 통해, 평균의 신뢰구간을 구할 수 있다.

200개로만 통계량을 구하는 것이 아니라 200개를 기준으로 복원 추출하여 새로운 통계량을 구하는 것을 예시로 들 수 있다.

머신러닝에서 부트스트램의 의미

머신러닝에서 부트스트랩은 아래와 같이 해석될 수 있다.

· 랜덤 샘플링을 통해 학습 데이터를 늘리는 방법
· 여러 모델을 학습시켜 추론 결과의 평균을 사용하는 방법(=앙상블)

복원추출이란?

복원추출(Sampling with replacement)이란 확률을 구할 때, 추출했던 것을 원래대로 돌려놓고 다시 추출하는 방법을 말한다.

#16 모수가 매우 적은 (수십개 이하) 케이스의 경우 어떤 방식으로 예측 모델을 수립할 수 있을까요?

모수는 모집단의 수가 아닌, 평균, 표준편차 등의 모집단의 특징을 말합니다. 여기서는 모집단의 수로 잘못 쓰인 것으로 보이며, 데이터가 적은 경우라 가정하고 답변을 작성하였습니다.

표본이 매우 작은 경우 표본평균의 분포가 정규분포를 따른다고 가정할 수 없으므로 **비모수적 방법**을 채택하여 예측 모델을 수립할 수 있다. 하지만 중심극한정리에 의해 표본의 크기가 30보다 클 경우 표본평균이 정규분포를 따른다고 가정할 수 있으므로, 이 경우에는 모수적 방법을 사용한다.

#17 베이지안과 프리퀀티스트 간의 입장차이를 설명해주실 수 있나요?

베이지안은 사건의 확률을 바라볼 때, 사전 확률을 미리 염두해두고 사건의 발생에 따라 베이즈 정리로 사후 확률을 구해 다시 사전 확률을 업데이트시킨다. 즉, 베이지안은 **과거의 사건이 현재 사건에 영향을 끼친다는 입장**을 가지고 있다.

반면, 프리퀀티스트는 확률을 무한번 실험한 결과, 객관적으로 발생하는 현상의 빈도수로 바라본다. 즉, 프리퀀티스트는 **현재의 객관적인 확률에 의해서만 사건이 발생한다는 입장**을 가지고 있다.

#18 검정력(statistical power)은 무엇일까요?

검정력은 대립가설 H1이 참인 경우 귀무가설 H0를 기각(대립가설 H1을 채택)할 확률이다.

#19 missing value가 있을 경우 채워야 할까요? 그 이유는 무엇인가요?

missing value를 처리하는 방법에는 크게 4가지가 있다.

1. 그대로 놔두기: 누락된 데이터를 그대로 놔두는 방법이다.
2. 삭제하기: 누락된 데이터를 제거하는 방법이다. 그러나 중요한 정보를 가진 데이터를 잃을 위험이 있다.
3. 특정 값으로 채우기: 0, 빈번한 값, 지정한 상수값으로 채우기
4. 예측하여 채우기: K-means, 평균값, 중앙값으로 대체하는 것

1번 방법을 사용하여, 데이터가 누락된 채로 놔둔다고 가정하자. 일부 xgboost같은

알고리즘은 결측값을 고려하여 잘 학습한다. 그러나 **결측치를 처리하는 로직이 없는 알고리즘(ex. sklearn의 LinearRegression)은 누락된 데이터 때문에 엉망이 될 수 있다.** 따라서 결측치를 처리해주어야한다.

2번 방법을 사용하여, 누락된 데이터를 제거한다고 해보자. 제거하는 방법은 가장 쉬운 방법이다. 그러나 만약 100명 중 한명의 특징(feature)이 누락된 상태이므로, 해당 특징을 전부 삭제한다면 중요한 특성을 잃어버리는 결과를 초래하게 된다.

3번, 4번 방법을 사용하여 결측치를 채운다고 해보자. 결측치를 채움으로서, 중요한 정보를 잃지않고 특성을 유지할 수 있다. 그러나 만약 100명 중 99명의 특징이 누락된 상태라고 한다면, 해당 특징을 어떠한 값으로 채우는 행위가 무의미할 것이다.

따라서 **결측치 상태나 비율, 어떤 모델을 사용할 것인지**에 따라서 결측치 대응 방법이 달라질 수 있다.

#20 아웃라이어의 판단하는 기준은 무엇인가요?

이상치(outlier)는 전체 데이터의 패턴에서 벗어난 이상한 값을 가진 데이터를 말한다. 이상치는 모델의 성능에 영향을 미치므로 이를 탐지하는 것은 정말 중요하다.

이상치를 탐지하는 방법 중 하나로 **IQR(Inter Quantile Range) 기법**이 있다. IQR 기법을 사용하기 위해서는 우선 데이터를 오름차순으로 정렬하고 25%, 50%, 75%, 100%로 4등분을 한다. 이 75% 지점과 25% 지점의 값의 차이를 IQR이라고 한다. 이 IQR에 1.5를 곱한 값을 75% 지점의 값에 더하여 최대값을, 25% 지점의 값에서 빼서 최소값을 계산한다. 이 때 최소값보다 작거나 최대값보다 큰 값을 이상치라고 판단한다.

또 다른 탐지 방법으로는 **Z-score를 계산하는 방식**이 있다. Z-score는 데이터가 평균에서 얼마나 떨어져 있는지를 나타내는 지표로, 임계값을 설정하여 Z-score이 이 값보다 크다면 이상치로 판단한다. 하지만 Z-score는 데이터가 평균에서 얼마나 벗어났는지를 측정하며, 가우시안 분포 가정이 중요하다.

#21 필요한 표본의 크기를 어떻게 계산합니까?

먼저 **모집단의 크기 N을 구하고, 신뢰수준 z와 오차범위 e**를 얼마로 할지 선정하여 표본의 크기를 구할 수 있다.

참고로 신뢰수준은 표본추출을 반복했을 때 얼마나 그 결과를 신뢰할 수 있는지에 대한 정도로 95% 를 주로 사용한다.

오차범위는 작을 수록 모집단의 특성에 대한 유용한 정보를 제공하지만 모집단에 대한 추론이 틀릴 가능성도 높아지므로 10% 를 넘지 않게 한다.

#22 Bias를 통제하는 방법은 무엇입니까?

편향(Bias)는 데이터 내에 있는 모든 정보를 고려하지 않음으로 인해, 지속적으로 잘못된 것들을 학습하는 경향을 의미한다. 이는 언더피팅(Underfitting)과 관계되어 있다.

반대로 분산(Variance)는 데이터 내에 있는 에러나 노이즈까지 잘 잡아내는 highly flexible models에 데이터를 피팅시킴으로써, 실제 현상과 관계 없는 랜덤한 것들까지 학습하는 알고리즘의 경향을 의미한다. 이는 오버피팅(Overfitting)과 관계되어 있다.

편향(Bias)과 분산(Variance)은 한 쪽이 증가하면 다른 한 쪽이 감소하고, 한쪽이 감소하면 다른 한쪽이 증가하는 tradeoff 관계를 가진다.

Bias를 통제하기 위한 방법으로는 뉴런이나 계층의 개수가 같은 모델의 크기 증가, 오류평가시 얻은 지식을 기반으로 입력 특성 수정, 정규화, 모델 구조를 수정, 학습 데이터 추가 등의 방법이 있다.

#23 로그 함수는 어떤 경우 유용합니까? 사례를 들어 설명해주세요.

우선 **단위 수가 너무 큰 값들을 바로 회귀분석 할 경우 결과를 왜곡할 우려가 있으므로 이를 방지하기 위해 사용**된다.

예를들어, 나이와 재산보유액의 관계를 회귀분석으로 푼다고 했을 때, 재산보유액의 숫자가 굉장히 클 수 있다. 재산보유액에 로그를 취할 경우, 데이터의 왜도와 첨도를 줄일 수 있어 정규성이 높아지는 효과를 얻는다.

또한 **비선형관계의 데이터를 선형으로 만들기 위해** 사용된다.

예를들어, 기하급수적으로 늘어나는 제곱 형식의 그래프에 자연로그를 취하면 그 관계가 직선(선형)이 된다.

|로그함수 주의사항

로그 함수는 0~1 사이에서는 음수값을 가지므로, $log(1+x)$와 같은 방법으로 처

리해주어야한다.

| 왜도(skewness)와 첨도(Kurtosis)

· 왜도는 데이터가 한쪽으로 치우친 정도이다.
· 첨도는 분포가 얼마나 뾰족한지를 나타내는 정도이다.

#24

베르누이 분포, 이항 분포, 카테고리 분포, 다항 분포, 가우시안 정규 분포, t 분포, 카이제곱 분포, F 분포, 베타 분포, 감마 분포에 대해 설명해주세요.

##1 베르누이 분포

우선 베르누이 시행이란 결과가 두 가지 중 하나만 나오는 것을 말한다. 베르누이 확률변수는 시행결과가 0 또는 1이 나오므로 이산확률변수이다.

##2 이항 분포

베르누이 시행을 N번 시행한 것을 말한다. 예를 들어, 동전 던지기를 10번 던져서 앞면이 나온 횟수를 확률 변수로 둔다. 마찬가지로 시행 결과가 횟수로 나오므로 이산확률변수이다.

##3 카테고리 분포

카테고리 분포(Categorical distribution)는 베르누이 분포를 확장한 개념이다. 즉 카테고리 시행(여러개의 카테고리 중 하나를 선택하는 실험)의 결과는 카테고리 분포를 따르게 된다. 카테고리 분포를 누적하면 다항분포를 얻게 된다.

카테고리 확률변수는 one-hot vector로 표현할 수 있다. 예를 들어, 주사위의 경우 K=6인 카테고리 분포를 따른다고 표기할 수 있다. 눈이 2인 주사위면이 나왔다고 할때, 이때 카테고리 RV=[0,1,0,0,0,0]이 된다. RV안의 각 원소들은 베르누이 분포를 따르고, 각각 자신들만의 모수를 갖는다. (RV = Random Variable = 확률변수)

##4 다항 분포

성공확률이 Θ인 베르누이 시행을 n번 반복했을 때의 성공횟수가 이항분포를 따르는 것처럼, 성공확률이 Θ=(Θ1...Θk)인 카테고리 시행을 n번 반복했을 때의 각 카테고리별 성공횟수는 다항분포(Multinomial distribution)을 따르게 된다.

(베르누이 분포 -> 이항 분포)≈(카테고리 분포 -> 다항 분포)

##5 가우시안 정규 분포
평균을 중심으로 좌우가 대칭인 종 모양을 그리는 정규분포이다.

정규 분포 식에서 변수는 x이다. σ와 μ는 그래프를 종모양으로 만드는데 사용된다. μ는 확률 변수 X의 평균이고 μ는확률 변수 X의 표준 편차이다. 종 모양의 그래프는 평균을 기준으로 좌우 대칭을 이룬다. 표준 편차가 높을 수록 그래프는 완만한 곡선 형태를 띄게 된다.

##6 t 분포
t 분포는 정규분포와 같이 중심을 기준으로 좌우 대칭이고 종모양 형태를 갖고 중심은 0으로 고정되어 있는 분포이다.

자유도(degree of freedom, df)에 따라 종의 형태가 조금씩 변화한다.

df는 표본수와 관련이 있는 개념으로, 표본이 많아지면 표준정규분포와 거의 동일한 형태를 보인다.

##7 카이제곱 분포
정규 분포의 제곱합은 χ^2 분포를 따른다.

##8 F 분포
F 분포는 독립적인 χ^2 변수의 비가 따르는 분포이다.

##9 감마 분포
감마 분포는 감마 함수를 사용하여 전체 k번의 사건이 일어날 때까지 걸리는 시간을 나타내는 연속 확률분포이다.

Θ와 k는 감마 분포의 모수이다.

감마 분포는 0~∞까지 값을 가질 수 있으며 모수의 베이지안 추정을 위해 사용된다. 감마 함수는 팩토리얼을 함수로 일반화한 것이다.

#25 출장을 위해 비행기를 타려고 합니다. 당신은 우산을 가져가야 하는지 알고 싶어 출장지에 사는 친구 3명에게 무작위로 전화를 하고 비가 오는 경우를 독립적으로 질문했습니다. 각 친구는 2/3로 진실을 말

하고 1/3으로 거짓을 말합니다. 3명의 친구가 모두 “그렇습니다. 비가 내리고 있습니다”라고 말했습니다. 실제로 비가 내릴 확률은 얼마입니까?

만약 출장지에 비가 올 확률이 25%라면 실제로 출장지에 비가 내릴 확률은 약 72.7%이다.