데이터

독립성

[독립성의 중요성]

(확률)변수들 간의 독립성을 의미하며, 하나의 변수가 다른 변수에 영향을 미치지 않는 상태를 말한다.

독립성 가정은 많은 통계적 검정에서 기본 가정으로 사용된다.

통계적 검정

데이터의 전수 조사는 실제로 불가능한 경우가 많으므로, 통계학적 추정과 검정을 통해 의사 결정을 내린다. 이는 데이터가 가설을 잘 따르는지 판단할 수 있게 한다.

[통계적 검정의 기본]

귀무 가설과 **대립 가설**을 먼저 설정한다. 각 가설은 서로 대립하는 주장을 펼친다. 각 가설의 기각 여부는 일반적인 방법으로 **검정통계량**을 바탕으로 한다.

검정통계량을 바탕으로 귀무 가설을 기각한다., 또는 귀무 가설을 기각하지 않는다. 이렇게 두 가지의 의사 결정을 내릴 수 있다. 이 때 **신뢰 구간**과 **기각역**을 근거로 결정을 내리게 된다.

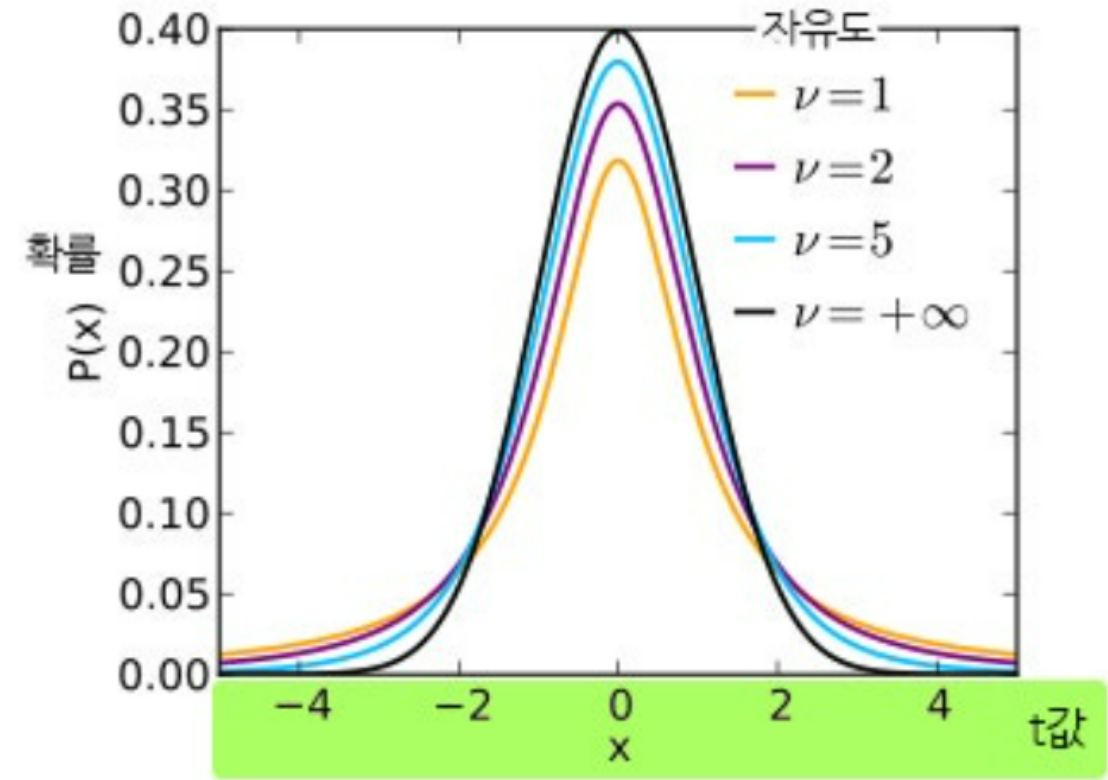
[주요 기법]:

- **카이제곱 검정** : 범주형 데이터에서 두 변수 간 관계를 분석하는 방법이다.
  - *적합도* : 범주형 데이터에서의 그룹의 비율, 그룹의 분포 등을 비교 검정.
  - *독립성* : 두 범주형 데이터 간 관계를 검정.
  - *동질성* : 데이터 그룹에 대응하는 분포를 검정.

	비교 대상		
적합도 검정 (goodness of fit test)	[범주형 변수] Group 1 Group 2 ... Group r	VS.	[알려진 사실] $p_1 : p_2 : \cdots : p_r$
독립성 검정 (Test of Independence)	[범주형 변수 A] Group 1 Group 2 ... Group r	VS.	[범주형 변수 B] Group 1 Group 2 ... Group r
동질성 검정 (Test of Homogeneity)	[부 모집단] Population 1 Population 2 ... Population r	VS.	[범주형 변수] Group 1 Group 2 ... Group r

---

- **t-검정** : 두 데이터 그룹 간의 평균 차이를 비교하는 방법.



특징으로는 위 그림과 같은 t-분포를 기반으로 한다는 것인데, 이는 데이터의 수가 30이 넘어가지 않을 때 적용된다.

특성 핸들링

데이터의 특성을 만져 원하는 목적을 이룰 수 있게 하는 방법이다.

- **기법**:
  - **차원 축소**: **PCA**(**주성분 분석**), **LDA**(**선형 판별 분석**) 등을 통해 차원을 줄여 데이터의 복잡성을 낮추는 방법이다. 핵심은 차원을 줄이기 이전의 주요 특성들을 차원을 줄

인 이후에도 유지하는 것이다.
    - **특성 필터링**: 무의미한 데이터는 빼거나, 정제되지 않은 데이터를 변형하는 등의 특성을 만지는 방법이다.

특성 분석

주어진 데이터에서 각 특성의 중요도를 평가하고 분석하는 과정이다. 중요한 특성은 이후의 여러 작업들에 큰 영향을 미친다.

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

분류기

베이즈 정리 & 나이브 베이즈

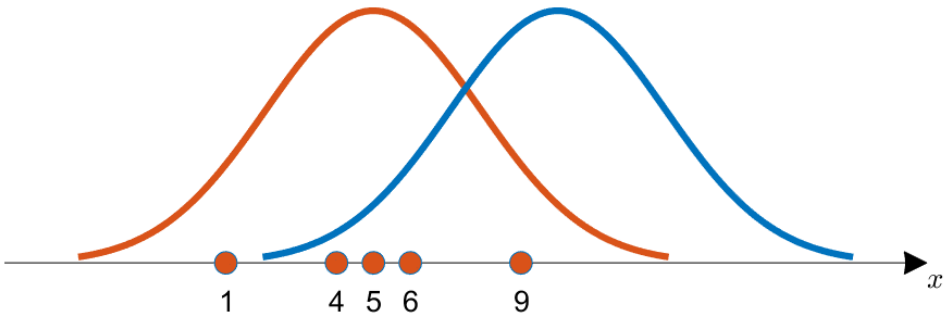
- **베이즈 정리**: 조건부 확률을 기반으로 사전 확률을 갱신하는 방법이다.
- **나이브 베이즈**: 모든 특성 간의 독립성을 가정하여 베이즈 정리를 적용하는 간단하면서도 강력한 분류 알고리즘이다.

최대우도법(최대가능도법)

- **정의**: 주어진 데이터에 대해, 일어날 가능성을 최대화하는 매개변수를 찾는 방법이다.

확률 변수의 모임 확률 변수에 대한 확률

주어진 확률 변수들의 확률을 최대화하는 분포를 찾아낸다.



최대우도법은 회귀 분석, 통계 모델링 등에서 사용되며, 확률 모델의 매개변수를 추정하는 데 자주 사용된다.

로지스틱 회귀(Logistic Regression)

로지스틱 회귀는 **이진 분류 문제**에서 사용되는 **선형 모델**의 일종으로, 데이터가 특정 클래스에 속할 확률을 추정한다. 주로 종속 변수가 두 가지 범주로 나뉘는 경우에 사용되며, 이를 통해 특정 사건이 발생할 확률을 예측할 수 있다.

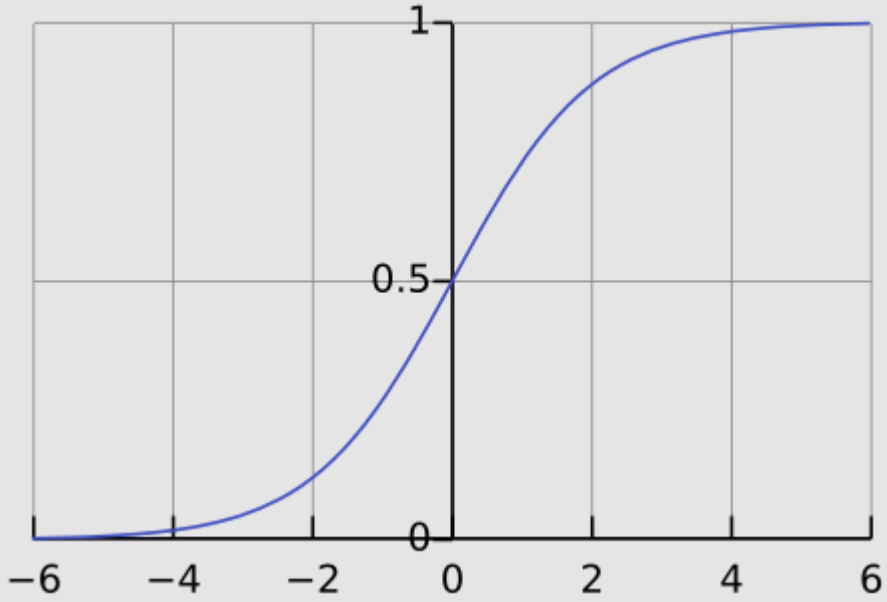
예 : *스팸 메일 분류, 암 진단 여부, 고객 이탈 예측* 등...

- **시그모이드 함수 사용**: 로지스틱 회귀는 시그모이드 함수를 사용하여 예측 결과를 0과 1 사이의 확률 값으로 변환한다.
- **선형 회귀와의 차이**: 선형 회귀와 달리, 로지스틱 회귀는 종속 변수의 값이 0과 1로 제한되어야 하는 경우에 적합하다.
- **해석 용이성**: 모델의 계수를 해석하여 각 독립 변수가 종속 변수에 미치는 영향을 이해할 수 있다.

$$ h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} $$

여기서, $h_\theta(x)$는 **예측 확률**을 의미하며, $\theta^T x$는 **독립 변수의 선형 결합**이다.

함수 $h_\theta(x)$의 출력은 0과 1 사이의 값을 가진다.



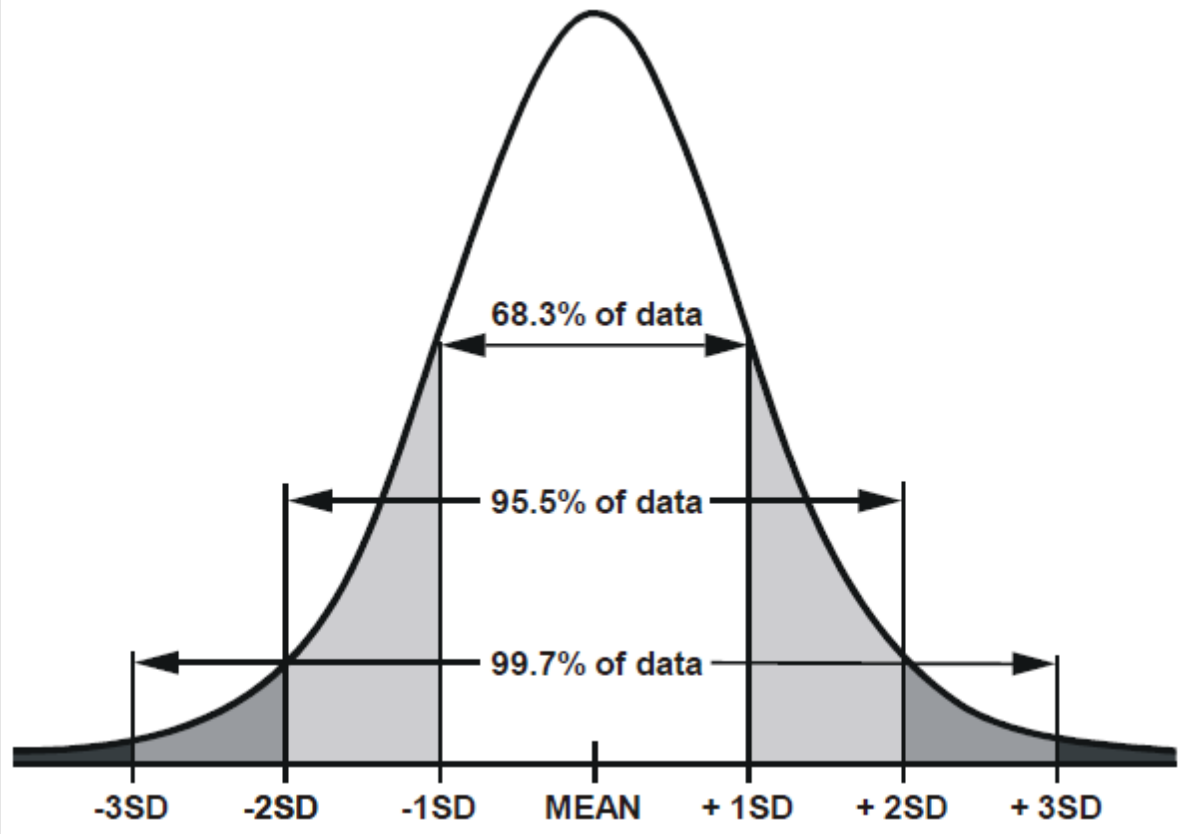
평가

신뢰 구간(Confidence Interval)

신뢰 구간은 **모집단의 특정 모수**(예: 평균)가 포함될 것으로 예상되는 범위를 나타낸다. 이는 모델 예측의 불확실성을 정량화하며, 결과의 **신뢰성**을 평가할 수 있는 지표이다.

이 모델은 95% 신뢰 수준에서, 예측된 확률이 실제 확률과 ±5% 이내에 있을 것이다.

: 신뢰 구간이 좁을수록 모델의 예측이 더 정확하다는 것을 의미한다.



혼동 행렬(Confusion Matrix)

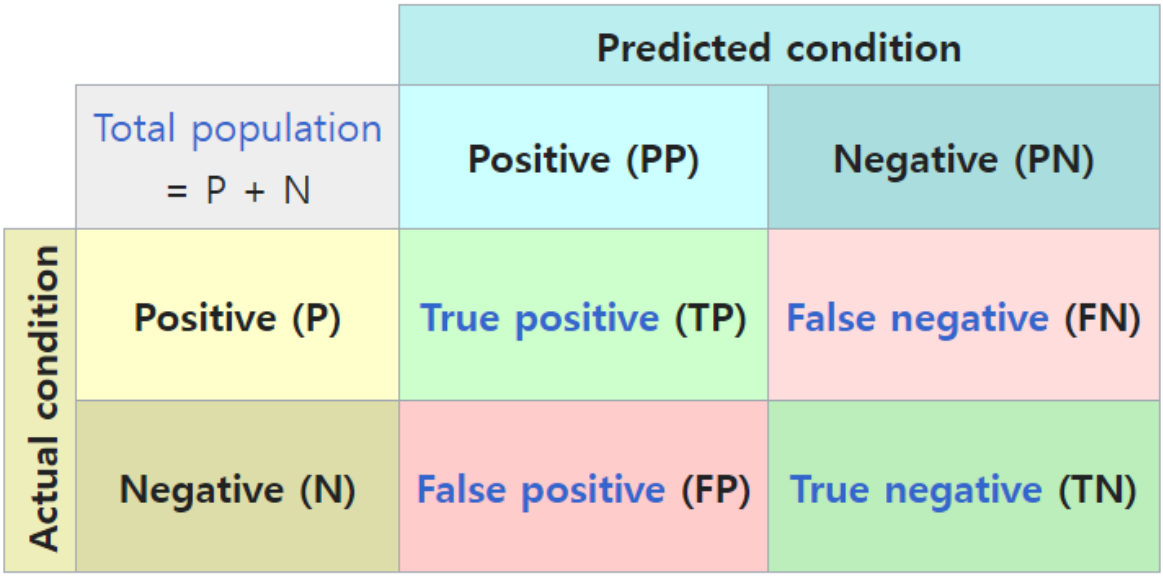
혼동 행렬은 **분류 모델의 성능**을 평가하는 데 사용되는 도구로, 모델이 예측한 값과 실제 값을 비교하여 성능을 정량화한다.

구성 요소

- **TP(참 긍정, True Positive)**: 실제로 긍정인 데이터를 모델이 긍정으로 예측한 경우.
- **FP(거짓 긍정, False Positive)**: 실제로 부정인 데이터를 모델이 긍정으로 예측한 경우.
- **TN(참 부정, True Negative)**: 실제로 부정인 데이터를 모델이 부정으로 예측한 경우.
- **FN(거짓 부정, False Negative)**: 실제로 긍정인 데이터를 모델이 부정으로 예측한 경우.

예 :

- **암 진단 모델**:
  - TP: 실제로 암이 있는 환자를 암이 있다고 예측!
  - FP: 실제로 암이 없는 환자를 암이 있다고 예측!
  - TN: 실제로 암이 없는 환자를 암이 없다고 예측!
  - FN: 실제로 암이 있는 환자를 암이 없다고 예측!



ROC & AUC

ROC(Receiver Operating Characteristic) 곡선

ROC 곡선은 분류기의 **민감도(True Positive Rate, TPR)** 와 **특이도(False Positive Rate, FPR)** 간의 관계를 나타내는 곡선이다.

- **민감도 (TPR)** : 실제로 긍정인 데이터를 긍정으로 예측한 비율.
- **특이도 (FPR)** : 실제로 부정인 데이터를 긍정으로 잘못 예측한 비율.

AUC(Area Under the Curve)

AUC는 **ROC 곡선 아래의 면적**을 의미하며, 모델의 전반적인 성능을 평가하는 지표로 사용된다. **AUC** 값이 1에 가까울수록 모델의 성능이 우수함을 의미한다.

- **AUC** $=$ 0.5 : 모델이 무작위로 예측하는 수준
- **AUC** $>$ 0.7 : 꽤 좋은 모델
- **AUC** $>$ 0.9 : 매우 우수한 모델