통계학

- **통계적 모델링**
  - 적절한 가정 위에서 확률분포를 추정(inference)하는 것이 목표
  - 기계학습과 통계학이 공통적으로 추구하는 목표
  - 유한한 개수의 데이터만 관찰해서 모집단의 분포를 정확하게 알아낸다는 것은 불가능

    => 근사적으로 확률분포 추정
    - 예측모형의 목적: 분포를 정확하게 맞추는 것보다 데이터와 추정 방법의 불확실성을 고려해서 위험을 최소하하는 것
- **모수적(parametric) 방법론**

  데이터가 특정 확률분포를 따른다는 선험적으로(a priori) 가정한 후, 그 분포를 결정하는 모수(parameter)를 측정하는 방법
- **비모수적(nonparametric) 방법론**

  특정 확률분포를 가정하지 않고 데이터에 따라 모델의 구조 및 모수의 개수가 유연하게 바뀜

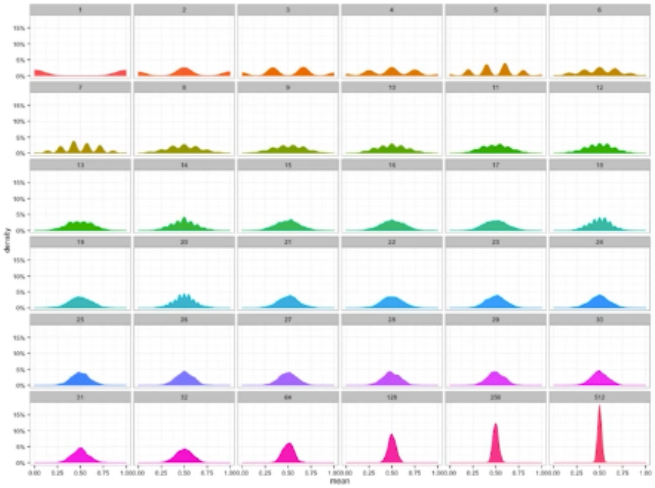
  기계학습의 많은 방법론이 비모수적 방법론에 속함
- 확률분포 가정하는 방법: 히스토그램을 통해 우선 모양 관찰
  - **베르누이분포**: 데이터가 2개의 값(0, 1)만 가지는 경우
  - **카테고리분포**: 데이터가 N개의 이산적인 값을 가지는 경우
  - **베타분포**: 데이터가 $[0, 1]$ 사이에서 값을 가지는 경우
  - **감마분포, 로그정규분포 등**: 데이터가 0 이상의 값을 가지는 경우
  - **정규분포, 라플라스분포 등**: 데이터가 $\mathbb{R}$ 전체에서 값을 가지는 경우

  기계적으로 확률분포 가정 X, 데이터 생성 원리를 먼저 고려하는 것이 원칙

  각 분포마다 검정하는 방법 있음 => 모수를 추정한 후에는 반드시 검정 필요
- 데이터의 확률분포 가정 => 모수 추정 가능
- 정규분포의 모수는 평균 과 분산 σ^2으로 이를 추정하는 통계량(statistic)은 아래와 같음

	표본 통계량	(추정된) 모수
평균	$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$	$\mathbb{E}[\overline{X}] = \mu$
분산	$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$	$\mathbb{E}[S^2] = \sigma^2$

  표본분산을 구할 때 N이 $N-1$로 나누는 이유는 불편(unbiased) 추정량을 구하기 위함
- **표집분포(sampling distribution)**
  - **표본분포(sample distribution)**: 단순 표분들의 분포, N이 커져도 정규분포를 따르지 않음
  - **표집분포(sampling distribution)**: 통계량(표본평균과 표본분산)의 확률 분포
  - 표본평균의 표집분포는 N이 커질수록 정규분포 $N(\mu, \sigma^2 N)$를 따름 => **중심극한정리(Central Limit Theorem)**
  - 중심극한정리는 모집단의 분포가 정규분포를 따르지 않아도 성립



- 표본평균이나 표본분산은 중요한 통계량이지만 확률분포마다 사용하는 모수 다름 => 적절한 통계량이 달라지게 됨
- **최대가능도 추정법(maximum likelihood esstimation, MLE)**

  이론적으로 가장 가능성이 높은 모수를 추정하는 방법

  $\hat{\theta}_{MLE} = \underset{\theta}{argmax} L(\theta; X) = \underset{\theta}{argmax} P(X|\theta)$

  - **가능도(likelihood) 함수**
    - 가능도 함수 $L(\theta; X)$는 모수 θ 를 따르는 분포가 x를 관찰할 가능성
    - 확률로 해석하면 X

      주어진 데이터 X에 대해 (데이터가 주어진 상황에서) 모수 θ를 변수로 둔 함수, 대소비교 가능한 함수
    - 데이터 집합 X가 독립적으로 추출되었을 경우 로그가능도 최적화
    - $L(\theta; X) = \prod_{i=1}^{n} P(x_i|\theta) \Rightarrow logL(\theta; X) = \sum_{i=1}^{n} logP(x_i|\theta)$
- 왜 로그가능도를 사용하는지
  - 로그가능도를 최적화하는 모수 θ는 가능도를 최적화하는 MLE가 됨
  - 데이터의 숫자가 적으면 상관없지만 **만일 데이터의 숫자가 수억 단위가 된다면**

    => **컴퓨터의 정확도로 가능도를 계산하는 것은 불가능**
  - 데이터가 독립일 경우

    => 로그를 사용하면 가능도의 곱셈을 로그가능도의 덧셈으로 바꿀 있음

    => 컴퓨터 연산 가능
  - 경사하강법으로 가능도를 최적화할 때, 미분 연산 사용, 로그가능도 사용 시

    => **연산량 $O(n^2)$에서 $O(n)$으로 줄여줌**
  - 대게의 손실함수의 경우, 경사하강법 사용 => **음의 로그가능도(negative log-likelihood)를 최적화**하게 됨
- **정규분포에서 MLE**
  - 추정된 모평균: $\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i$
  - 추정된 모분산: $\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$
  - MLE는 불편추정량 보장 X

    but, 통계적 일관성(consistency)는 보장 가능
- **카테고리 분포(Multinoulli)에서 MLE**

Multinoulli($x; p_1, \ldots, p_d$)를 따르는 확률변수 X로부터 독립적인 표본 $x_1, \ldots, x_n$을 얻었을 때 최대가능도 추정법을 이용하여 모수 추정하면

$\hat{\theta}_{MLE} = \underset{p_1,\ldots,p_d}{argmax}\ log\ P(x_i|\theta) = \underset{p_1,\ldots,p_d}{argmax}\ log(\prod_{i=1}^{n}\prod_{k=1}^{d} p_k^{x+i,k})$

- 목적식: $log(\prod_{i=1}^{n}\prod_{k=1}^{d} P_k^{x_i,k}) = \sum_{i=1}^{d}(\sum_{i=1}^{n} X_i, k) log\ p_k = \sum_{k=1}^{d} n_k log\ p_k$
- 카테고리 분포의 모수는 $\sum_{k=1}^{d} P_k = 1$ 제약식을 만족해야 함
- 제약식을 만족하면서 목적식을 최대화하는 것
- **라그랑주 승수법**을 통해 최적화 문제 풀 수 있음
  - 라그랑주 승수법(Lagrange Multiplier Method)

    제약 조건이 있는 최적화 문제를 풀기 위한 고안한 방법

    어떠한 문제의 최적점을 찾는 것이 아니라, 최적점이 되기 위한 조건을 찾는 방법

    => 최적해의 필요조건 찾는 방법

  $\Rightarrow \mathcal{L}(p_1, \ldots, p_k, \lambda) = \sum_{k=1}^{d} n_k log p_k + \lambda(1 - \sum_k p_k)$

  - 목적식 미분: $0 = \frac{\partial \mathcal{L}}{\partial p_k} = \frac{n_k}{p_k} - \lambda$
  - 제약식 미분: $0 = \frac{\partial \mathcal{L}}{\partial \lambda} = 1 - \sum_{k=1}^{d} p_k$
  - 각각 미부해서 나오는 이 두 식을 합하면: $p_k = \frac{n_k}{\sum_{k=1}^{d} n_k}$
- 카테고리 분포의 MLE: **경우의 수를 세어서 비율을 구하는 것**

- **딥러닝에서 최대가능도 추정법**
  - 최대가능도 추정법을 이용해서 기계학습 모델 학습 가능
  - 딥러닝 모델의 가중치를 $\theta = (W^{(1)}, \ldots, W^{(L)})$라 표기했을 때 분류 문제에서 소프트맥스 벡터는 카테고리분포의 모수 $(p_1, \ldots, p_k)$를 모델링함
  - one-hot vector로 표현한 정답 레이블 $\mathrm{y} = (y_1, \ldots, y_k)$을 관찰데이터로 이용해 확률분포인 softmax vector의 로그가능도(log likelihood)를 최적화할 수 있음

    $\hat{\theta}_{\mathrm{MLE}} = \underset{\theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} y_{i,k} \log \left(\mathrm{MLP}_\theta\left(\mathbf{x}_i\right)_k\right)$

- **확률분포의 거리**
  - 기계학습에서 사용되는 손실함수

    => 모델이 학습하는 확률붙포와 데이터에서 관찰되는 확률분포의 거리 통해 유도
  - 데이터공간에 2개의 확률분포 $P(x), Q(x)$가 있을 경우, **두 확률분포 사이의 거리(distance)**를 계산할 때 아래의 함수 이용
    - 총변동 거리 (Total Variation Distance, TV)
    - 쿨백-라이블러 발산 (Kullback-Leibler Divergence, KL)
      - 이산확률변수: $\mathbb{KL}(P\|Q) = \sum_{\mathbf{x}\in\mathcal{X}} P(\mathbf{x}) \log\left(\frac{P(\mathbf{x})}{Q(\mathbf{x})}\right)$
      - 연속확률변수: $\mathbb{KL}(P\|Q) = \int_X P(\mathbf{x}) \log\left(\frac{P(\mathbf{x})}{Q(\mathbf{x})}\right) \mathrm{d}\mathbf{x}$
      - 분해: $\mathbb{KL}(P\|Q) = \underbrace{-\mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[\log Q(\mathbf{x})]}_{\text{크로스 엔트로피}} + \underbrace{\mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[\log P(\mathbf{x})]}_{\text{엔트로피}}$

        분류 문제에서 정답레이블을 P, 모델 예측을 Q라 두면 **최대가능도 추정법은 쿨백-라이블러 발산을 최소화**하는 것과 같음

- 바슈타인 거리 (Wasserstein Distance)