04_Statistics

Introduction

이번 장에서는 데이터에서 이러한 매개변수를 학습하는 방법에 대해 설명합니다.
확률 모델의 모든 매개변수 θ가 알려져 있다고 가정했습니다.
D로부터 θ를 추정하는 과정을 모델 피팅 또는 트레이닝이라고 하며, 머신 러닝의 핵심입니다.
이러한 추정치를 생성하는 방법에는 여러 가지가 있지만, 대부분은 다음과 같은 형태의 최적화 문제로 귀결됩니다.

$$\hat{\theta} = argmin\ L(\theta)$$

여기서 L(θ)는 일종의 손실 함수 또는 목적 함수입니다.
이 장에서는 몇 가지 다른 손실함수에 대해 설명합니다.

Maximum likelihood estimation(MLE)

- ### Definition
  1. **정의:**
     - MLE은 데이터셋 D에 대한 likelihood 함수 $p(D|\theta)$를 최대화하는 매개변수 θ를 찾는 것으로 정의됩니다.
       $\theta_{mle} = argmax_\theta\ p(D|\theta)$ 데이터가 임의의 파라미터 θ에 의존하는 확률분포를 따르고 있는 표본 데이터가 주어졌을 때,
       $v_1, v_2, \cdots, v_n$ 데이터가 발생할 확률을 구하고자 합니다.
       $p(v_1, \cdots, v_n|\theta)$ 하지만 현재 θ를 모르기에 베이즈 규칙을 이용하여 θ가 발생할 우도 (주어진 표본들에 비추어봤을 때 모집단의 모수 θ에 대한 추정값)으로 바꾸어 생각할 수 있다.
       $L(\theta|v_1, \cdots, v_n)$
  2. **로그 우도:**
     - 로그 우도 $L(\theta)$는 각 예제에 대한 로그 확률의 합으로 표현됩니다.
     - $L(\theta) = \sum_{n=1}^{N}\ log\ p(y_n|x_n, \theta)$
     - 로그 우도 함수를 사용하는 이유
       - 우도 함수는 0에서 1사이에 값을 취함
         우도 함수에 로그값을 취해 $-\infty < ln\ L <= 0$
         - 우도 함수의 값은 사례수가 많은 경우에는 극히 작은 값
  3. **MLE 목적 함수:**

- MLE는 로그 우도를 최대화하는 것으로 정의되며, 최적화 문제로 표현됩니다:
  최적화 알고리즘을 사용하여 이 목적 함수를 최대화하면, 모델의 파라미터를 찾을 수 있다.
  - $\hat{\theta}_{mle} = argmax_\theta \, L(\theta)$

4. **음의 로그 우도 (NLL : Negative Log Likelihood):**
   - 목적에 따라 최적화를 위해 (비용 함수를 최소화하기 위해) 음의 로그 우도가 일반적으로 사용됩니다.
   - $NLL(\theta) = -\sum_{n=1}^{N} \, log \, p(y_n|x_n, \theta)$
5. **무조건적 (비지도) MLE:**
   - 무조건적 (비지도) 모델의 경우, 출력 y_n은 있지만 입력 x_n이 없을 때 음의 로그 우도를 최소화하여 MLE를 찾습니다.
   - $\hat{\theta}_{mle} = argmin_\theta \, -\sum_{n=1}^{N} \, log \, p(y_n|\theta)$
6. **입출력의 결합 우도:**
   - 어떤 경우에는 입력과 출력의 결합 우도를 최대화하고 싶을 수 있습니다.
     입력과 출력 간의 상관 관계를 모델로써 잘 파악하고자 하는 목적을 나타내는 표현
   - $\hat{\theta}_{mle} = argmin_\theta \, -\sum_{n=1}^{N} \, log \, p(y_n|\theta)$

요약하면, MLE는 통계 모델의 매개변수를 likelihood 함수를 최대화하여 추정하는 방법으로, 딥러닝에서는 입력을 주고 출력을 예측하는 작업에 사용됩니다. 음의 로그 우도는 일반적으로 최적화의 목적 함수로 사용되며, 조건부와 결합 우도는 모델링 작업에 따라 선택됩니다.

- # Justification for MLE

1. **Bayesian 해석:**
   - MLE를 베이지안 관점에서 바라볼 수 있습니다.
     베이지안 관점에서 MLE는 베이지안 사후 확률인 $p(\theta|D)$를 균일한 사전 분포(uniform prior)를 사용한 델타 함수(delta function)로 근사화한 것으로 볼 수 있습니다. 베이지안 통계에서 사후 확률은 베이즈 정리를 사용하여 다음과 같이 표현됩니다:
     $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \; p(\theta|D) = \frac{p(D|\theta) \cdot p(\theta)}{p(D)}$
     여기서 $p(D \mid \theta)$는 likelihood, $p(\theta)$는 prior, $p(D)$는 evidence(데이터의 확률)입니다. 베이지안 추론에서는 posterior(θ에 대한 확률분포)를 구하기 위해 prior와 likelihood의 곱을 evidence로 나누어야 합니다. 만약 균일한 사전 분포를 사용한다면, prior $p(\theta)$는 상수가 되어 posterior를 likelihood로만 구하는 것과 같아집니다. $p(\theta \mid D) \propto p(D \mid \theta)$
     $\theta_{map} = argmax_\theta \, log \, p(\theta|D) = argmax_\theta \, log \, p(D|\theta) + log \, p(\theta)$
     $log \, p(\theta)$는 상수이기 때문에 최적화에 영향을 미치지 않습니다. 따라서 균일한 사전 분포를 사용하면, MLE를 구하는 것과 동일한 결과를 얻게 됩니다.
     $\theta_{map} = \theta_{mle}$
2. **예측 분포의 근사:**

- 또 다른 MLE를 정당화하는 방법은 얻어진 예측 분포 $p(y|\theta_{mle})$가 데이터의 경험적 분포(empirical distribution)와 가능한 비슷하도록 만드는 것입니다. 예측 분포는 모델이 주어진 파라미터 MLEθMLE를 기반으로 얻은 결과물인데, 이 분포가 실제 데이터의 경험적 분포와 유사하면 모델이 데이터를 잘 설명하고 있다고 할 수 있습니다.
- KL(Kullback-Leibler) 다이버전스를 사용하여 두 분포 간의 유사성을 측정합니다. KL 다이버전스를 최소화하는 것은 MLE를 수행하는 것과 동일하며, 이는 Negative Log Likelihood(NLL)를 최소화하는 것과도 동일합니다.

이를 통해 모델이 주어진 데이터의 경험적 분포와 가능한 가까워지도록 모델을 학습하는 것이 MLE를 정당화하는 한 가지 방법임을 보여줍니다.

MLE는 베이지안 관점에서는 균일한 사전 분포를 사용한 베이지안 추론의 특별한 경우로 해석되며, 예측 분포의 경험적 분포와의 유사성을 최대화함으로써 정당화될 수 있습니다.

Example

Bernoulli 분포에 대한 MLE

- **베르누이 분포:** 동전 던지기의 결과를 나타내는 확률변수 Y를 고려하고, 이때 이벤트 $Y = 1$은 앞면을 나타내고 $Y = 0$은 뒷면을 나타냅니다. 확률 θ는 앞면이 나올 확률입니다.
- **Negative Log Likelihood (NLL):** 베르누이 분포의 NLL은 다음과 같이 정의됩니다.

  $NLL(\theta) = -\sum_{n=1}^{N} [I(y_n = 1) log\, \theta + I(y_n = 0) log(1 - \theta)]$ 여기서 I는 지시함수입니다.
- **MLE 계산:** NLL를 최소화하기 위해 편미분하여 미분이 0이 되는 값을 찾습니다. 결과적으로 MLE를 사용하여 θ는 다음과 같이 계산됩니다:

  $\theta_{mle} = \frac{N_1}{N_0+N_1}$
  여기서 N_1은 1의 개수(앞면), N_0은 0의 개수(뒷면), N은 전체 샘플 수입니다.
- **해석:** MLE는 단순히 앞면의 경험적 비율로 계산됩니다.

categorical 분포에 대한 MLE

- **범주 분포:**
  K면체 주사위를 N번 던질 때의 결과를 나타내는 확률변수 Y_n을 고려합니다. 예를 들어, 주사위를 던져 나오는 눈의 수, 각 주사위 면에 대한 확률을 범주 분포로 나타낼 수 있습니다. 각각의 범주 k에 대한 확률은 θ_k입니다.
- **NLL 및 MLE 계산:**
  범주 분포에 대한 NLL을 정의하고, MLE를 구하기 위해 Lagrange multipliers를 사용합니다.

- **Lagrange multipliers** (라그랑주 승수)
  등식 제약이 있는 최적화 문제에서 사용되는 방법 중 하나입니다. 예를 들어, 범주 분포의 경우 다음과 같이 등식 제약을 가지면서 NLL을 최소화하는 문제를 풀게 됩니다.
  - 예제
    목적 함수 f 와 등식 제한조건 g 이 다음과 같은 경우를 생각하자.
    $f(x,\ y) = x^2 + y^2$
    $g(x,\ y) = x + y = 1$

    이를 만족하는 f의 최적해(최대 or 최솟값)를 찾아라.

    라그랑주 승수법의 보조함수를 아래와 같이 정의합니다.
    $L(x,\ y,\ \lambda) = x^2 + y^2 - \lambda(x + y - 1)$

    위의 함수 L의 gradient vector가 0벡터인 위치를 구합니다.
    $\frac{\partial L}{\partial x} = 2x - \lambda = 0$
    $\frac{\partial L}{\partial y} = 2y - \lambda = 0$
    $\frac{\partial L}{\partial \lambda} = x + y - 1 = 0$

    위의 연립방정식을 풀면 다음과 같은 최적해의 위치를 구할 수 있습니다.
    $x = y = \frac{1}{2},\ \lambda = 1$

  $\theta_{mle,k} = \frac{N_k}{N}$
  여기서 N_k는 $Y = k$인 경우의 개수이고, N은 전체 샘플 수입니다.
- **해석:**
  각 범주의 경험적 비율로 MLE를 계산합니다. 즉, 각 범주에 대한 확률은 해당 범주가 관측된 빈도로 나눈 것으로 추정됩니다. 이는 주어진 데이터에서 각 범주가 나타날 확률을 경험적으로 파악하는 것을 의미

univariate Gaussian 분포에 대한 MLE

- **단변량 가우시안 분포의 MLE:** 단변량 가우시안에서는 평균과 분산을 계산합니다.
  $\mu_{mle} = \frac{1}{N}\sum_{n-1}^{N} y_n$
  $\sigma^2_{mle} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \mu_{mle})^2$

multivariate Gaussian 분포에 대한 MLE

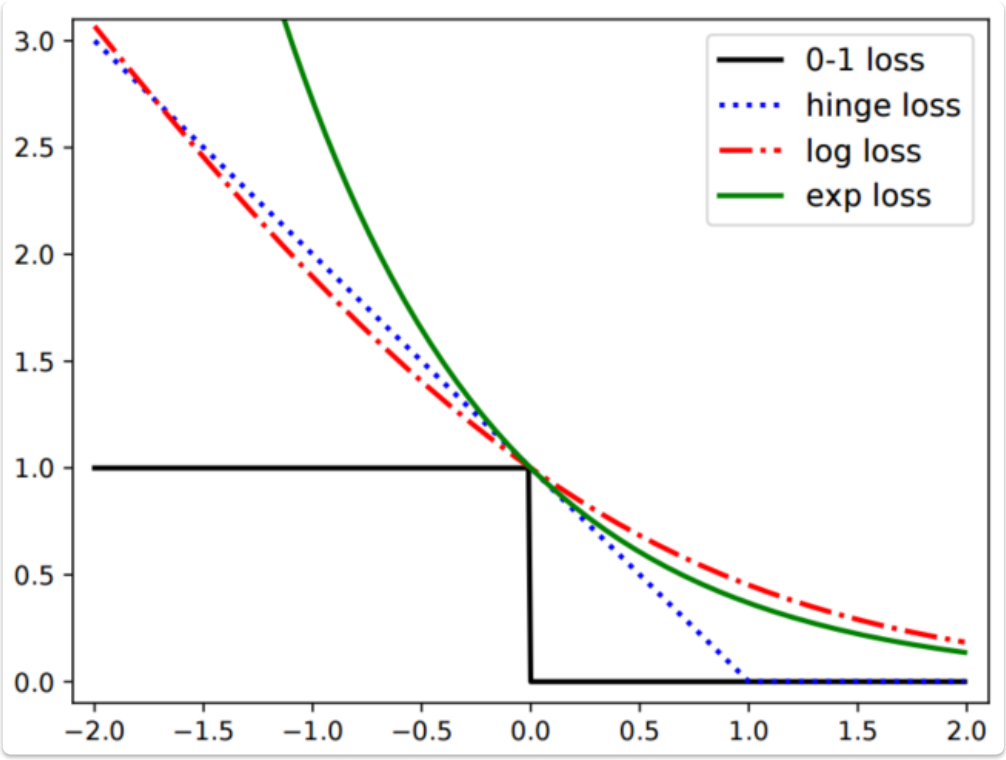
- **다변량 가우시안 분포의 MLE:** 다변량 가우시안에서는 평균 벡터와 공분산 행렬을 계산합니다.
  $\mu_{mle} = \frac{1}{N}\sum_{n-1}^{N} y_n$
  $\Sigma_{mle} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \mu_{mle})(y_n - \mu_{mle})^T$
- **해석:**
  여러 변수가 상호 작용하는 경우, 각 변수 간의 공분산(두 변수 간의 선형관계)도 중요하게 작용합니다. 따라서 평균 벡터와 공분산 행렬을 사용하여 다변량 분포

를 묘사합니다. 이것은 변수들 간의 상관관계 및 분산-공분산 특성을 고려하는 더 복잡한 분포를 나타냅니다.

- ### linear regression 분포에 대한 MLE
  - **선형 회귀의 RSS 및 MLE 계산:**
    $RSS(w) = (Xw - y)^T(Xw - y)$
    $w_{mle} = (X^TX)^{-1}X^Ty$
  - **해석:**
    - RSS는 회귀 모델의 예측값과 실제 관측값 간의 차이를 나타내는 잔차의 합을 나타내며, 이를 최소화하는 것이 회귀 모델의 적합도를 높이는 방향입니다.
    - MLE는 주어진 데이터에 대해 회귀 모델이 가장 확률적으로 적합하도록 회귀 계수를 추정합니다. 역행렬 계산을 통해 입력과 출력 간의 관계를 최적화하여 회귀 모델을 구성합니다.

Empirical Risk Minimization (ERM)

- **정의:** MLE를 일반화하여 어떤 다른 손실 함수를 사용할 수 있도록 하는 개념입니다.
- ## Example
  0-1 손실 함수는 다음과 같이 정의됩니다. $l_{01}(y_n, \theta; x_n) = \begin{cases} 0, \ if \ y_n = f(x_n; \theta) \\ 1, \ if \ y_n \neq f(x_n; \theta) \end{cases}$
  여기서 $f(x; \theta)$는 예측기입니다.
  이는 데이터 포인트가 올바른 범주에 할당되면 0이고, 그렇지 않으면 1입니다. 이러한 0-1 손실을 사용하여 에측 모델을 최적화하면 훈련 데이터셋에서의 미스클래스피케이션 비율을 최소화하는 것이 됩니다.
  이때, 경험적 위험(Empirical risk)는 다음과 같이 표현됩니다. $L(\theta) = \frac{1}{N}\sum_{n=1}^{N} l_{01}(y_n, \theta; x_n)$
  이것은 훈련 세트에서의 경험적 미스클래스피케이션 비율을 나타냅니다.
- ## Surrogate loss
  - **문제상황:**
    0-1 손실은 비선형 및 미분 불가능한 특성을 가지고 있어 최적화가 어려울 수 있습니다. 이를 극복하기 위해 대체 손실 함수를 사용합니다. 대표적인 대체 손실 함수로는 log loss와 hinge loss가 있습니다.



이진 분류를 위한 다양한 손실 함수의 그림입니다. 가로축은 마진이고 세로축은 손실입니다.

- Log Loss:
  확률적인 이진 분류 문제를 고려해보겠습니다. 모델이 만들어내는 Log odds(η)를 사용하여 다음과 같이 확률을 정의할 수 있습니다:

  $p(\tilde{y}|x,\theta) = \sigma(\tilde{y}\eta) = \frac{1}{1+e^{-\tilde{y}\eta}}$
  여기서 σ는 시그모이드 함수이며, $\eta = f(x;\theta)$는 Log-odds입니다.

  로그 손실은 다음과 같이 정의됩니다:

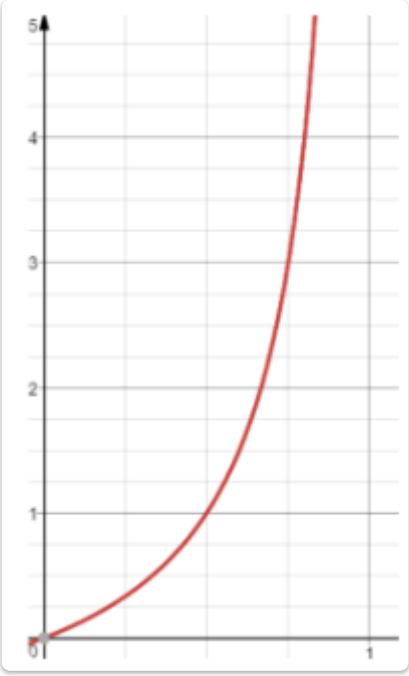
  $l_{ll}(\tilde{y},\eta) = log(1+e^{-\tilde{y}\eta})$
  이 손실 함수는 미스클래스피케이션 비율에 대한 최대한 타이트한 상한(Tight Upper Bound)을 제공하며, 그림에서 그래프로 확인할 수 있습니다.

  - Tight Upper Bound?
    타이트한 상한은 어떤 값이나 함수가 다른 값이나 함수를 너무 크게 벗어나지 않도록 하는 상한을 나타냅니다. 로그 손실은 미스클래스피케이션 비율에 대한 최대한 타이트한 상한을 제공합니다. 즉, 로그 손실은 모든 확률에 대해 높은 불확실성을 가진 경우에도 손실을 제한하는 특성이 있습니다. 이는 모델이 불확실성을 과도하게 추정하는 경우에도 로그 손실이 크게 증가하지 않도록 합니다.

    로지스틱 회귀에서 로그 손실을 최소화하는 것은 모델을 미스클래스피케이션 비율에 대해 더 견고하게 만들어, 불확실성이 높은 예측에서도 안정적인 학습을 가능케 합니다.

- **Log-odds?**
  오즈는 사건이 발생할 확률을 사건이 발생하지 않을 확률로 나눈 비율입니다.
  이를 수식으로 나타내면 다음과 같습니다:

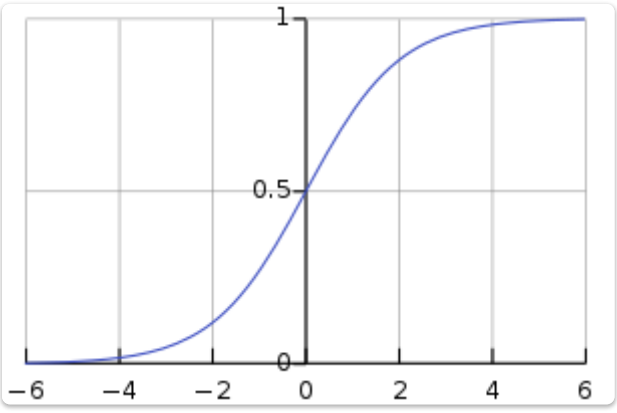
  $odds = \frac{p(y=1|x)}{1-p(y=1|x)}$
  $p(y=1|x)$ 확률 값이 다음 그래프와 같이 1에 가까워질수록 odds 값은 엄청나게 상승합니다.



  - **한계점:**
    $0 < odds < \infty$의 범위에 속하기에 , 범위에 제약이 있다.
    확률값과 odds값은 비대칭성(Asymmetric)을 띈다.

    이러한 한계를 극복하기 위해 다음과 같이 로그함수를 취한다:

    $logit(p) = log\frac{p}{1-p}$
    이를 통해 $-\infty < log(odds) < \infty$의 범위를 가지며,
    대칭성 또한 가지게 된다.



- **Hinge Loss:**
  힌지 손실은 다음과 같이 정의됩니다:

  $l_{hinge}(\tilde{y}, \eta) = max(0, 1 - \tilde{y}\eta)$
  힌지 손실은 0-1 손실 함수의 볼록한 상한을 제공합니다. 이는 미분 가능한 손실 함수보다 최적화에 더 용이합니다. 그러나 힌지 손실은 미분 가능하지 않은 부분이 있기 때문에 미분 가능한 손실 함수와 달리 모든 최적화 알고리즘에서 사용할 수는 없습니다. 힌지 손실은 SVM에서 주로 사용되며, 모델을 효과적으로 학습시키면서도 미분 가능하지 않은 부분을 다루기 위해 대안적인 손실 함수를 사용할 때 유용합니다.
  - **Convex Upper Bound?**
    볼록한 상한(Convex Upper Bound)"은 함수나 손실 함수의 상한(upper bound)이 볼록 함수(convex function)의 형태를 띄는 것을 의미합니다. 여기서 "볼록한"이라는 용어는 함수의 곡선이 아래로 볼록하게 휘어지는 형태를 나타냅니다.

이러한 대체 손실 함수는 0-1 손실을 근사화하는 데 사용될 수 있으며, 그라디언트 기반 최적화를 통해 모델을 효과적으로 학습할 수 있도록 도와줍니다.

Other Estimation Methods

The metod of moments

- **모멘트 방법(Method of Moments, MOM):**
  통계 분포의 매개변수를 추정하기 위한 최대우도추정(MLE)의 대안입니다. MOM은 분포의 이론적 모멘트를 경험적 모멘트에 대입하고 이로부터 얻은 연립 방정식을 푸는 방법입니다.

  - **모멘트란?**
    확률 분포의 특성을 설명하기 위한 통계적인 개념 중 하나로, 모멘트는 데이터 집합의 각 원소를 사용하여 계산되는 측도입니다. 모멘트는 분포의 형태와 특징을 파악하는 데 사용됩니다. n차 모멘트는 확률 변수의 n승을 사용하여 계산됩니다.

  이론적 모멘트(μ_k)는 랜덤 변수 Y의 특정 함수들의 기대값을 나타냅니다. 예를 들어, $\mu 1 = E[Y], \mu 2 = E[Y^2]$ 등입니다. 경험적 모멘트는 데이터에서 추정되며 μ^k로 표기됩니다.
  이론적 모멘트와 경험적 모멘트는 다음과 같이 표현됩니다.

  - 이론적 모멘트: $\mu_k = E[Y^k]$
  - 경험적 모멘트: $\mu^k = \frac{1}{N} \sum_{n=1} N y_n^k$
  - 모멘트 방법 추정식: $\mu_k = \mu^k$

  MOM 추정은 이론적 모멘트를 그들의 경험적 상응물에 대입하고 매개변수를 푸는 것을 포함합니다. MOM은 계산상 간단하지만 MLE만큼 효율적으로 데이터를 활용하지

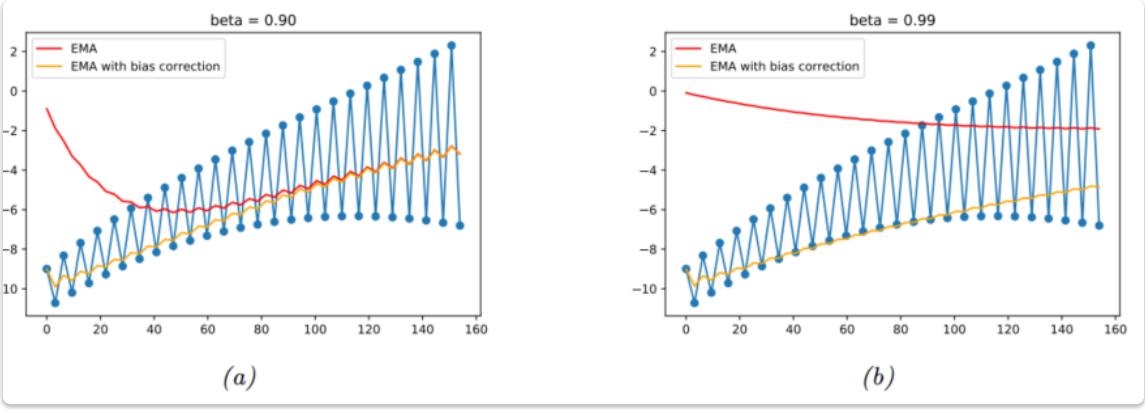
못할 수 있으며 특정 경우에는 일관성 없는 결과를 초래할 수 있습니다.

- Why?
  해당 특징을 설명하는 모멘트를 선택하지 않으면 추정된 파라미터가 실제와 불일치할 수 있습니다. 또한 선택한 모멘트가 모집단에서 존재하지 않는 경우에도 불일치성이 발생할 수 있습니다.

Online (recursive) estimation

- 순차적으로 도착하는 데이터를 처리할 때 온라인 학습 또는 재귀 추정이 사용됩니다. 새로운 데이터 포인트가 이용 가능해짐에 따라 매개 변수 추정치를 업데이트하는 개념입니다. 재귀 업데이트는 다음과 같은 형태를 가집니다:
  $\theta_t = f(\hat{\theta}_{t-1}, y_t)$
  여기서 θ_t는 업데이트된 매개 변수, $\hat{\theta}_{t-1}$은 이전 추정치, y_t는 새로운 데이터 포인트입니다.
- 지수 가중 이동 평균(EWMA): 온라인 학습에서 최근 예제에 더 많은 가중치를 부여하기 위해 지수 가중 이동 평균(EWMA)가 도입됩니다. 업데이트 규칙은 다음과 같다.

  - $\hat{\mu}_t = \beta\mu_{t-1} + (1-\beta)y_t$ = $\beta^2\mu_{t-2} + \beta(1-\beta)y_{t-1} + (1-\beta)y_t$
    - μ_t는 현재 시점 t에서의 EWMA 추정치입니다.
    - μ_{t-1}은 이전 시점 $t-1$에서의 EWMA 추정치입니다.
    - y_t는 현재 시점 t에서의 새로운 데이터 포인트입니다.
  - $(1-\beta)\sum_{k=0}^{t}\beta^k = (1-\beta)\frac{1-\beta^{t+1}}{1-\beta} = 1-\beta^{t+1}$
  - $\tilde{\mu}_t = \frac{\hat{\mu}_t}{1-\beta^t}$
    여기서 0 < β < 1입니다. 따라서, 우리는 $t \to \infty$일 때 $\beta^{t+1} \to 0$은 이므로, β가 작을수록 과거를 더 빨리 잊고 최근 데이터에 보다 최근 데이터에 더 빠르게 적응합니다. 이는 그림에 설명되어 있습니다.



1. beta = 0.9인 경우:
   - 빨간색 선은 일반적인 EWMA를 나타냅니다.
   - 오렌지색 선은 편향 보정이 적용된 EWMA를 나타냅니다.
   - EWMA는 현재 데이터 포인트에 대한 가중치를 높게 둔 채로 이전 데이터의 영향을 서서히 감소시키는 경향이 있습니다.

- 편향 보정된 EWMA는 초기에 특히 빨간색 선에 비해 작은 초기 편향을 가지며 점진적으로 증가합니다.

2. **beta = 0.99인 경우:**
   - 높은 β 값은 현재 데이터 포인트에 대한 가중치를 훨씬 높게 두기 때문에 최신 데이터에 더 민감하게 반응합니다.
   - 편향 보정된 EWMA는 초기에 상대적으로 큰 초기 편향을 가지며 더 빠르게 상승합니다.

이 규칙에 따라 EWMA는 이전 추정치와 새로운 데이터 포인트 간의 가중 평균을 계산하여 새로운 추정치를 얻습니다. 이 때, β값이 낮을수록 새로운 데이터에 대한 영향이 커지며, β 값이 높을수록 이전 추정치의 영향이 큽니다.

EWMA는 새로운 데이터에 대한 적응성과 추정치의 부드러움 사이의 균형을 제공합니다. β 값에 따라 가중치가 조절되어 최신 데이터와 이전 데이터 간의 균형이 조절됩니다.
편향 보정은 초기 편향을 완화하기 위해 도입되었으며, 초기에 편향이 크게 줄어들 수 있도록 합니다.

Regularization

- **문제 정의:**
  MLE(최대우도추정) 및 ERM(최소 경험적 위험)의 근본적인 문제는 훈련 세트에서 손실을 최소화하려고 시도할 것이지만, 이로 인해 미래 데이터에서 손실이 낮은 모델이 되지 않을 수 있다는 것입니다. 이를 과적합이라고 합니다.

  문제의 핵심은 모델이 관측된 훈련 데이터를 완벽하게 맞출 수 있는 충분한 매개변수를 가지고 있다는 것입니다. 따라서 이 모델은 경험적 분포를 완벽하게 일치시킬 수 있습니다. 그러나 대부분의 경우 경험적 분포는 실제 분포와 동일하지 않으므로 N개의 예제 집합에 모든 확률 질량을 할당하면 미래의 새로운 데이터를 위한 확률이 남아 있지 않게 됩니다. 즉, 모델이 일반화되지 않을 수 있습니다.
- **해결 방안:**
  과적합에 대한 주요 해결책은 정규화를 사용하는 것이며, 이는 NLL(또는 경험적 위험)에 패널티 항을 추가하는 것을 의미합니다. 따라서 다음과 같은 형태의 목적 함수를 최적화합니다.

  $L(\theta;\lambda) = \frac{1}{N}\sum_{n=1} N\ell(y_n,\theta;x_n) + \lambda C(\theta)$

  여기서 $\lambda \geq 0$는 정규화 매개변수이며, $C(\theta)$는 어떤 형태의 복잡성 페널티입니다.일반적인 복잡성 페널티로는 $C(\theta) = -log\,p(\theta)$를 사용하는 것이 흔합니다. 여기서 $p(\theta)$는 θ에 대한 사전 분포(prior)입니다.

이것을 최소화하는 것은 다음을 최대화하는 것과 동일합니다.

$\hat{\theta} = argmax_\theta \log p(\theta|D) = argmax_\theta [\log p(D|\theta) + \log p(\theta) - const]$

이를 최대 사후 추정이라고 합니다(MAP 추정).

Examples

MAP estimation for the Bernoulli distribution

- **문제정의:**
  전 던지기 시나리오가 있으며 베르누이 분포에서 앞면이 나올 확률인 θ를 추정하는 것이 목표입니다. 최대 우도 추정(MLE)은 특히 제한된 데이터가 있는 경우(예: 하나의 앞면만 관찰된 경우) 과적합을 초래할 수 있습니다.
- **해결방안:**
  과적합을 완화하고 θ의 극단적인 값(예: $\theta = 0$ 또는 $\theta = 1$)을 피하기 위해 $Beta(\theta \mid a, b)$ 형태의 베타 분포를 사용하여 정규화 항 또는 사전을 도입합니다. 여기서 $a, b > 1$은 θ의 값을 $a/(a+b)$ 근처로 유도하기 위해 사용됩니다. 로그 우도와 로그 사전(log prior)을 더하면 다음과 같습니다. $l(\theta) = log\, p(D|\theta) + log\, p(\theta)$
  $= [N_1 log\theta + N_0 log(1-\theta)] + [(a-1)log\,\theta + (b-1)log(1-\theta)]$
  - Add-one smoothing:
    - **정의:**
      Add-One Smoothing은 확률 모델에서 확률을 계산할 때 발생할 수 있는 0의 확률 문제를 해결하기 위한 방법 중 하나입니다. 특히, 이 방법은 카운트 기반 모델에서 0이 아닌 모든 카운트에 1을 더하여 확률을 계산하는 기술입니다.

    $\theta_{MAP} = \frac{N_1+a-1}{N_1+N_0+a+b-2}$
    최대 사후확률(Maximum A Posteriori, MAP) 추정을 사용하여 베르누이 분포의 모수 θ를 추정 $a = b = 2$로 설정한다면 θ값은 0.5 근처의 값을 약간 선호하는 것으로 파악됩니다. 그 경우 추정치는 다음과 같습니다. $\theta_{MAP} = \frac{N_1+1}{N_1+N_0+2}$
    0 카운트 문제를 피하기 위해 간단하게 1을 각각의 카운트에 더하고, 총 카운트에는 2를 더한 값으로 정규화한 것입니다. 이렇게 함으로써 0이나 1과 같은 극단적인 추정치를 피하고 모수 θ의 추정치를 부드럽게 업데이트할 수 있습니다.
- **흑조 박쥐 패러독스:**
  이 예제에서의 0 카운트 문제와 과적합은 철학에서 "흑조 박쥐 패러독스"라는 문제와 유사하며, 이는 모든 백조가 흰색이라는 고대 서양의 개념에서 파생되었습니다. 이 맥락에서 흑조는 존재할 수 없는 것의 상징이었습니다. 이러한 고전적인 사고 방식을 피하려면 귀납의 문제로 알려진 문제를 해결해야 합니다. 이 패러독스의 해결책은 귀납이 일반적으로 불가능하며, 우리가 할 수 있는 최선은 경험적 데이터를 사전 지식과 결합하여 미래에 대한 타당한 추측을 만드는 것입니다.

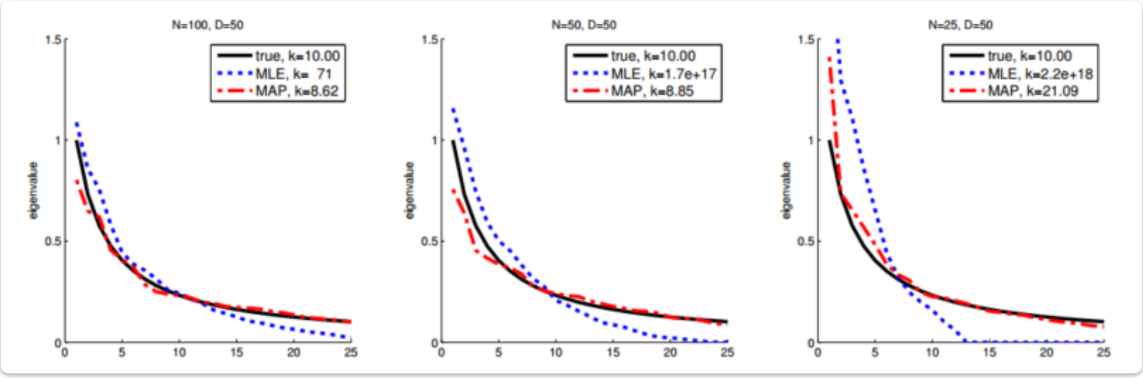
MAP estimation for the multivariate Gaussian

- **문제상황:**
  다변수 가우시안(MVN)의 MAP(Minimum A Posteriori) 추정에서 높은 차원에서 MLE(Maximum Likelihood Estimation)로 공분산을 추정하는 것은 어렵다.
- **해결방안: Shrinkage estimate**
  Shrinkage 추정은 MAP 추정의 한 형태로, 역 Wishart 사전 분포를 사용합니다. 역 Wishart 사전은 고유값의 대각 성분은 MLE와 같게 유지하면서, 그 외의 비대각 성분을 0에 가깝게 수축시키는 정규화를 수행합니다. 여기서 대각선 요소들은 각 변수의 분산을 나타내며, 비대각 요소들은 각 변수 간의 공분산을 나타냅니다.

  이렇게 함으로써, 변수 간의 공분산을 추정할 때 적은 데이터가 주어졌을 때에도 안정적이고 일반화된 추정을 할 수 있습니다. 특히, 변수 간의 관계가 거의 없다고 가정할 때 이러한 수축(Shrinkage)은 효과적입니다.

  1. **MLE for Covariance:**
     - MLE로 추정한 공분산 행렬은 다음과 같습니다.
       $\sum_{MLE} = \frac{1}{N} \sum i = 1N(y_i i - \mu_{MLE})(y_i - \mu_{MLE})^T$
  2. **Prior for Covariance:**
     - 역 Wishart 사전 분포를 사용하며, 사전 산포 행렬은 다음과 같습니다.
       $\sum_{prior} = \lambda \sum_0 + (1 - \lambda) \sum_{MLE}$
       여기서 λ는 정규화 정도를 제어합니다.
  3. **MAP Estimate for Covariance:**
     - 위에서 정의한 사전 분포와 MLE에 대한 식을 사용하여 MAP 추정치를 계산합니다.

```
$\sum_{MAP}{(i,j)}=\begin{cases}\sum_{MLE}(i,j)\qquad
if\,i=j\\ (1-\lambda)\sum_{MLE}\quad otherwise\end{cases}$
```



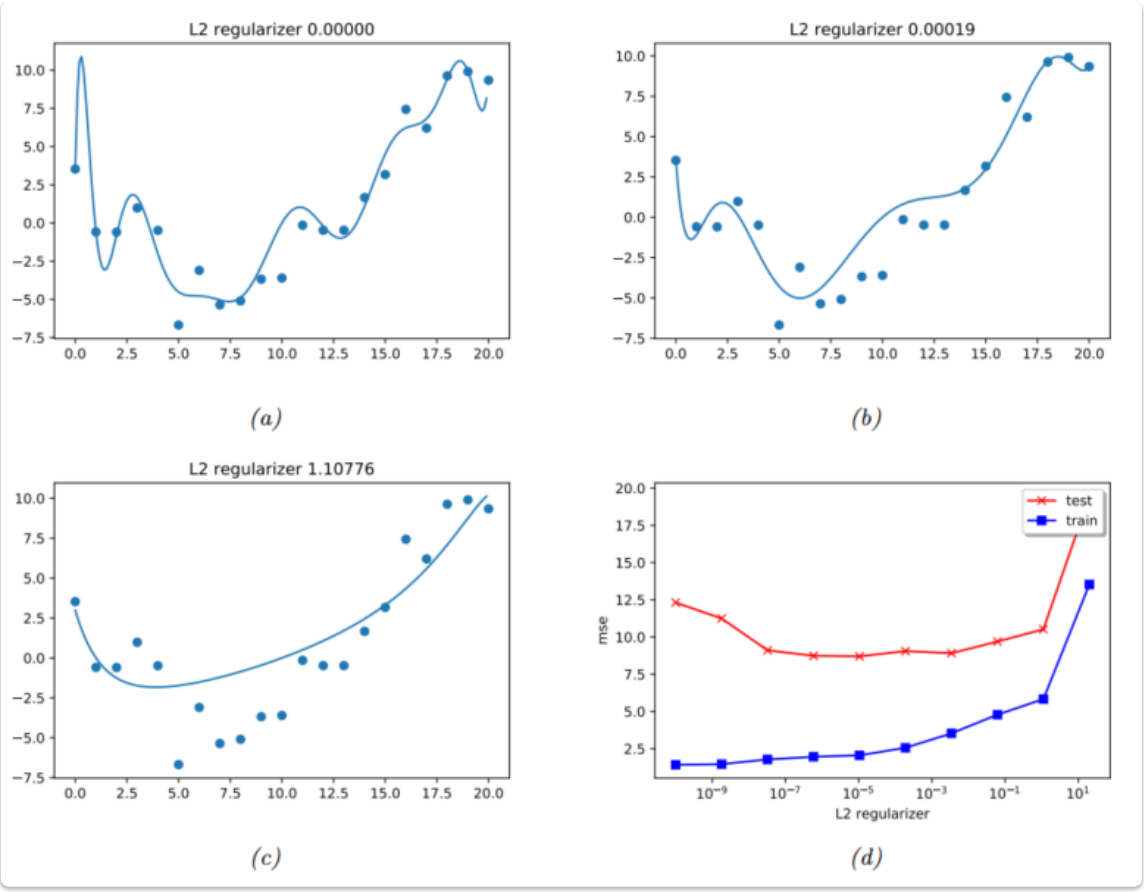
위의 그림은 D=50 차원에서 샘플 수 N이 100, 50, 25일 때 공분산 행렬을 추정하는 실험을 다룹니다. 여기서 사용된 추정 방법은 식 (4.98)에 따른 MAP 추정이며, 정규화 파라미터 λ는 0.9로 설정되었습니다.

그림에 대한 설명은 다음과 같습니다:

1. 실제 공분산 행렬의 고유값(eigenvalues)을 내림차순으로 나열한 실선 그래프 (solid black).
2. MLE로 추정한 공분산 행렬의 고유값을 내림차순으로 나열한 점선 그래프 (dotted blue).
3. MAP로 추정한 공분산 행렬의 고유값을 내림차순으로 나열한 대시 그래프 (dashed red).

그 결과로, MLE 추정치는 종종 나쁜 조건을 가지고 있음을 볼 수 있지만, MAP 추정치는 수치적으로 잘 행동하는 것으로 나타났습니다. 종종 MLE는 공분산 행렬의 역행렬을 계산하는 동안 수치적인 불안정성을 보일 수 있으나, MAP 추정치는 이러한 문제를 완화하는 경향이 있습니다. 이는 안정적인 다변수 가우시안 추정을 위해 Shrinkage 방법이 MLE의 불안정성을 줄이고, 안정적인 추정치를 얻을 수 있도록 도와준다는 것을 보여줍니다.

- ## weight decay
  - **문제정의:**
    다항 회귀에서 너무 높은 차수를 사용하는 것이 오버피팅을 일으킬 수 있음을 보여줍니다.
  - **해결방법:**
    이 경우 다항식의 차수를 감소시키는 것이 하나의 해결책이 될 수 있지만, 더 일반적인 해결책은 가중치(회귀 계수)의 크기에 패널티를 주는 것입니다. 이는 제로 평균 가우시안 사전분포 $p(w)$를 사용하여 수행됩니다. 결과적인 MAP 추정치는 다음과 같이 주어집니다. $\hat{w}_{MAP} = argmin_w NLL(w) + \lambda ||w||_2^2$
    여기서 $||w||_2^2 = \sum_{d=1}^{D} w_d^2$이며, λ가 클수록 매개변수들은 **큰** 것에 대해 (제로 평균 사전분포에서 벗어나는 것에 대해) 페널티를 받게 되어 모델이 덜 유연해집니다.
  - **선형회귀 때 해결방법:** Ridge regression
    선형 회귀의 경우, 이러한 패널티 방식을 리지 회귀라고 합니다. 다항 회귀 예제를 고려해 보겠습니다. 여기서 예측 변수는 다음과 같은 형태를 가지고 있습니다.
    $f(x; w) = \sum d = 0D w_d x_d = w^T[1, x, x^2, \ldots, x^D]$



고차 다항식을 사용한다고 가정하면 (예를 들어 D = 14), 데이터 포인트가 매우 적은 경우에도 MLE는 가중치를 조절하여 데이터를 아주 잘 맞출 수 있지만, 결과적인 함수는 매우 "불규칙"하여 오버피팅이 발생합니다. 그림에서는 λ를 증가시킴에 따라 오버피팅이 감소하는 것을 보여줍니다. 왼쪽에서 오른쪽으로 갈수록 정규화 정도가 증가하므로 모델 복잡성이 감소합니다.

- # Picking the regularizer using a validation set
  - **문제정의:**
    정규화 강도(λ)를 선택하는 방법에 대한 핵심 질문은 어떻게 정규화 강도를 선택할 것인가입니다. 작은 값은 우리가 경험적 위험을 최소화하도록 중점을 두게 할 것이며, 이는 오버피팅을 초래할 수 있습니다. 반면에 큰 값은 사전과 가깝게 유지하려고 노력하게 될 것이며, 이는 언더피팅을 초래할 수 있습니다.
  - **해결방안:**
    이 섹션에서는 λ의 값을 선택하는 데에 광범위하게 사용되는 간단한 방법을 설명합니다. 기본 아이디어는 데이터를 두 개의 상호 배타적인 세트로 나누는 것입니다.

    하나는 훈련 세트 D_{train}이고 다른 하나는 검증 세트 D_{valid}입니다(때로는 개발 세트로도 불립니다). (일반적으로 데이터의 약 80%를 훈련 세트로 사용하고 나머지 20%를 검증 세트로 사용합니다.) 우리는 D_train에서 모델을 맞추고(각 λ 설정에 대해), 그 후에 D_{valid}에서 성능을 평가합니다. 그런 다음 검증 성능이 가장 우수한 λ의 값을 선택합니다.

데이터셋에 대한 정규화된 경험적 위험을 다음과 같이 정의해 봅시다:

$R_\lambda(\theta, D) = \frac{1}{|D|} \sum_{(x,y)} \in D^\ell(y, f(x;\theta)) + \lambda C(\theta)$
여기서 $\ell(y, f(x;\theta))$는 손실 함수, $C(\theta)$는 정규화 항입니다.

각 λ에 대해 매개변수 추정치를 계산합니다:

$\theta^\lambda(D_{train}) = argmin_\theta R_\lambda(\theta, D_{train})$

그런 다음 검증 위험을 계산합니다:

$R_{val,\lambda} = R_0(\hat{\theta}_\lambda(D_{train}), D_{valid})$
계산된 매개변수 추정치를 사용하여 검증 세트 D_{valid}에서 검증 위험을 계산합니다
이는 모집단 위험의 추정값이며, 실제 분포 $p^*(x, y)$ 하에서의 기대 손실입니다.

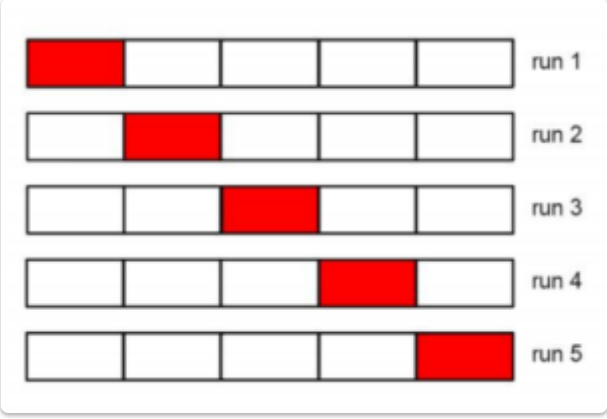
최적 λ 선택:

$\lambda^* = argmin_{\lambda \in S} R_{val,\lambda}$

(이는 S 내의 각 λ 값에 대해 모델을 한 번씩 맞추어야 하지만, 어떤 경우에는 이를 더 효율적으로 수행할 수 있습니다.) λ^*를 선택한 후 모델을 전체 데이터 집합인 $D = D_{train} \cup D_{valid}$에 재적합(다시 훈련)하여 다음과 같은 결과를 얻을 수 있습니다.

$\hat{\theta}^* = argmin_\theta \, R_{\lambda^*}(\theta, D)$

- # Cross-validation
  - **문제상황:**
    훈련 세트의 크기가 작은 경우에는 20%를 검증 세트로 남겨두면 모델 매개변수의 불안정한 추정치를 얻을 수 있습니다.
  - **해결방안:**
    이에 대한 간단하면서도 널리 사용되는 해결책은 교차 검증(CV)을 사용하는 것입니다. 아이디어는 다음과 같습니다. 훈련 데이터를 K 개의 폴드로 나눈 다음 각 폴드 $k \in 1, \ldots, K$에 대해 나머지 모든 폴드를 사용하여 모델을 훈련하고 k'번째 폴드에서 테스트하는 것입니다.



이러한 경우 다음과 같이 교차 검증 위험을 정의할 수 있습니다.

$R_{cv,\lambda} = \frac{1}{K}\sum_{k=1}^{K} R_0(\hat{\theta}_\lambda(D-k), D_k)$

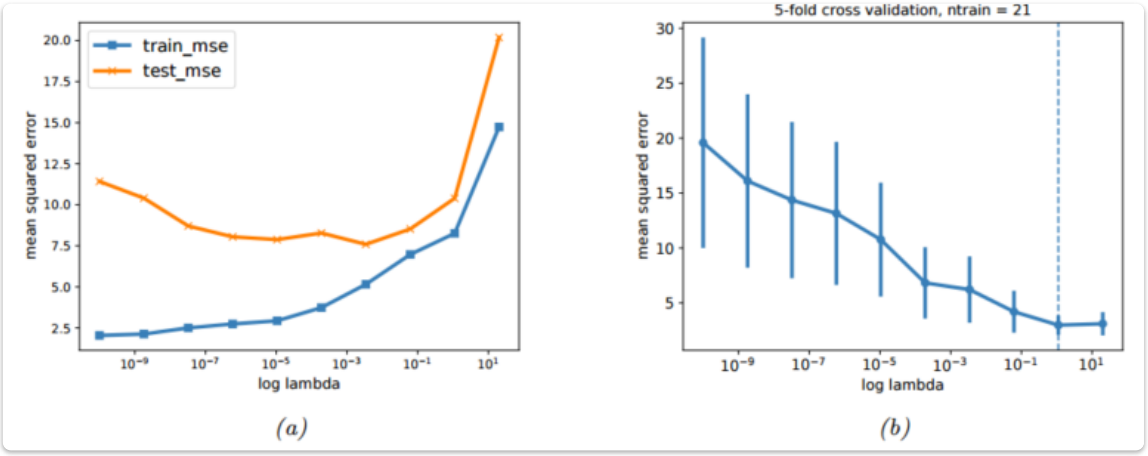
여기서 D_k는 k번째 폴드의 데이터이고, $D-k$는 다른 모든 데이터입니다. 이를 교차-검증된 위험이라고 합니다. 위의 그림에서는 K = 5에 대한 이 절차를 설명하고 있습니다. K = N으로 설정하면 한 개를 제외한 모든 항목에 대해 학습하고 나머지 하나를 테스트하는 방식으로 불리는 leave-one-out 교차 검증 방법을 얻을 수 있습니다.

- ## The one standard error rule

  모델의 평균 손실과 그 불확실성을 추정하고, 표준 오차를 고려하여 최적의 정규화 강도를 선택하는 단계입니다.

  다음으로, 각 예제 n에 대한 손실을 $L_n = l(y_n, f(x_n; \hat{\theta}_\lambda(D-n)))$로 정의하고, $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} L_n$를 추정된 평균으로 정의합니다. 그리고 $\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1} N(L_n - \hat{\mu})^2$를 추정된 분산으로 정의합니다. 이때, $\hat{\sigma}$는 L_n의 표본 간 고유한 가변성을 측정하고, $se(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{N}}$는 평균 $\hat{\mu}$에 대한 불확실성을 나타냅니다.

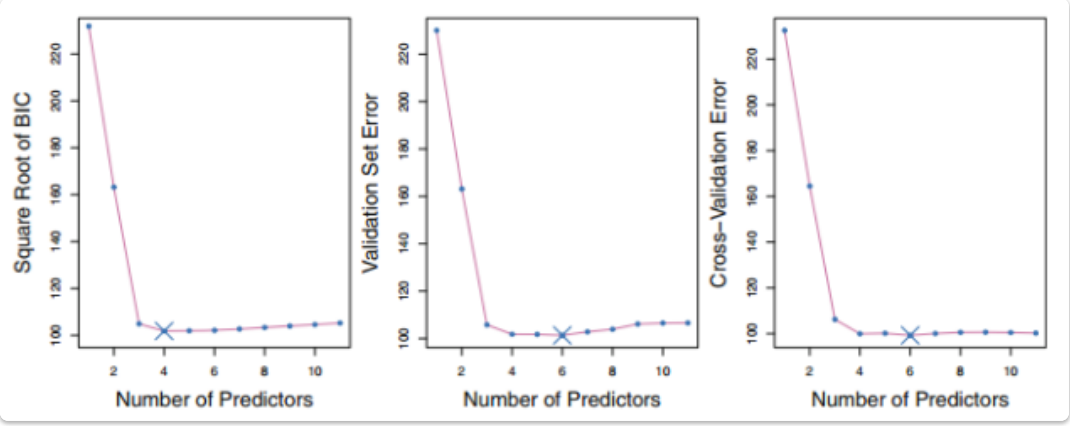
- ## Example: ridge regression

  릿지 회귀(Ridge Regression)가 21개의 데이터 포인트에 대한 14차 다항식을 적합시킬 때, 정규화 매개변수(Regularizer λ)의 다양한 값에 대해 어떻게 적용되었는지를 설명하고 있습니다. Ridge Regression 모델의 정규화 매개변수(Regularizer

λ)의 다양한 값에 대한 훈련과 테스트 세트에서의 성능 변화를 시각적으로 분석해 보겠습니다:

1. **MSE on Train and Test vs log(λ):**
   - 그림 (a)에서는 로그 스케일의 정규화 매개변수(λ)에 따른 훈련 세트 및 테스트 세트에서의 평균 제곱 오차(MSE)가 표시됩니다.
   - 왼쪽에서 오른쪽으로 가면서 정규화가 강해지고(λ 값이 증가), 모델 복잡성이 줄어들게 됩니다(MSE증가).
   - 훈련 세트와 테스트 세트 간의 MSE를 시각화하여 모델의 적합성과 일반화 성능 간의 균형을 확인할 수 있습니다.
2. **5-fold Cross-validation Estimate of Test MSE:**
   - 그림 (b)에서는 5-폴드 교차 검증을 사용하여 테스트 MSE의 추정치를 표시합니다.
   - 오류 막대는 평균의 표준 오차를 나타냅니다. 각 λ 값에서 5개의 폴드에 대한 교차 검증 결과의 흩어진 정도를 보여줍니다.
   - 여러 정규화 매개변수에 대한 모델의 교차 검증 성능을 비교하여 어떤 정규화 강도가 더 적절한지를 판단할 수 있습니다.
3. **Vertical Line and One Standard Error Rule:**



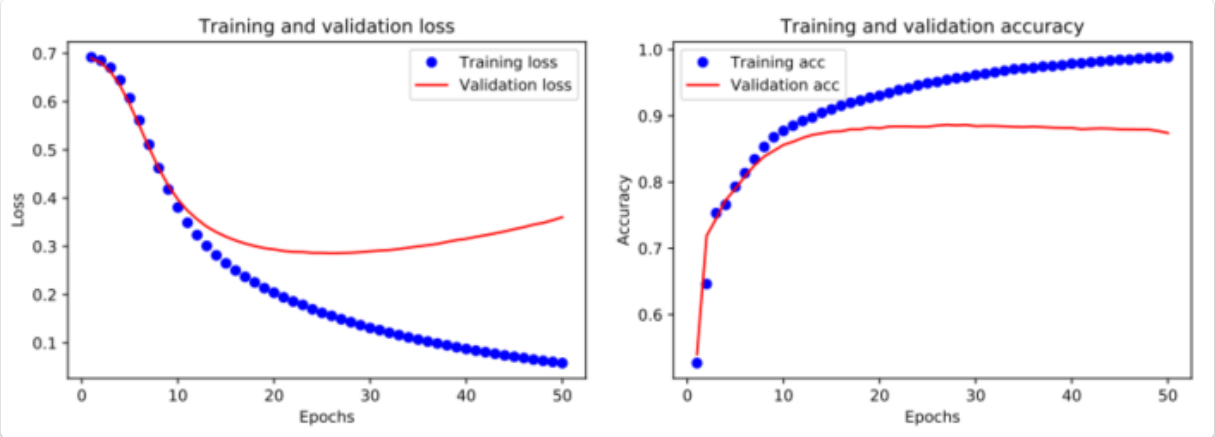
Gareth James의 An Introduction to Statistical Learning

   - 오류막대는 1 표준 오차 규칙(One Standard Error Rule)에서 선택한 정규화 매개변수를 나타냅니다.???
   - 테스트 오류가 최소인 경우에서의 정규화 매개변수보다 1 표준 오차 이상 크지 않으면서 가장 단순한 모델을 선택합니다.

- # Early stopping
  - **정의:**

    Early Stopping은 매우 간단하지만 효과적인 정규화 방법 중 하나로, 주로 복잡한 모델에서 효과가 있습니다. 이 방법은 최적화 알고리즘이 반복적으로 실행되는 특성을 활용합니다. 모델이 훈련 세트에 대한 정보를 지나치게 기억하는 것을 방지하기 위해, 검증 세트에서의 성능을 모니터링하다가 과적합의 징후를 감지하면 최적화 프로세스를 중지합니다.

- **과정:**
  1. **초기화:**
     - 최적화 과정을 시작할 때, 모델의 파라미터는 초기 추정값에서 출발합니다.
  2. **최적화 진행:**
     - 최적화 알고리즘은 반복적으로 훈련 데이터를 사용하여 파라미터를 조정하고 최적화를 시도합니다.
     - 파라미터는 훈련 세트에 대한 적합을 향상시키기 위해 조정됩니다.
  3. **검증 세트 성능 모니터링:**
     - 최적화 도중, 일정 간격으로 검증 세트에서 모델의 성능을 평가합니다.
     - 과적합이 시작되면, 검증 세트에서의 성능이 감소할 것으로 예상됩니다.
  4. **과적합 감지:**
     - 검증 세트에서의 성능 감소가 감지되면, 모델이 훈련 데이터에 과적합되고 있을 가능성이 높습니다.
     - 이를 확인하기 위해, 검증 세트에서의 성능 향상이 멈추는 지점을 찾습니다.
  5. **최적화 중단:**
     - 과적합이 감지되면, 최적화 프로세스를 중단하고 파라미터를 해당 지점에서 사용합니다.
     - 이 지점에서의 모델은 훈련 데이터와 검증 데이터 간의 균형을 유지하면서 학습됩니다.
- **이점:**
  - Early Stopping은 모델의 일반화 성능을 향상시키고, 훈련 데이터에 지나치게 적합되는 것을 방지하여 더 간단하면서도 효과적인 모델을 얻을 수 있도록 도와줍니다.
  - 특히, 복잡한 모델이나 데이터가 부족한 상황에서 사용되는 경우가 많습니다.

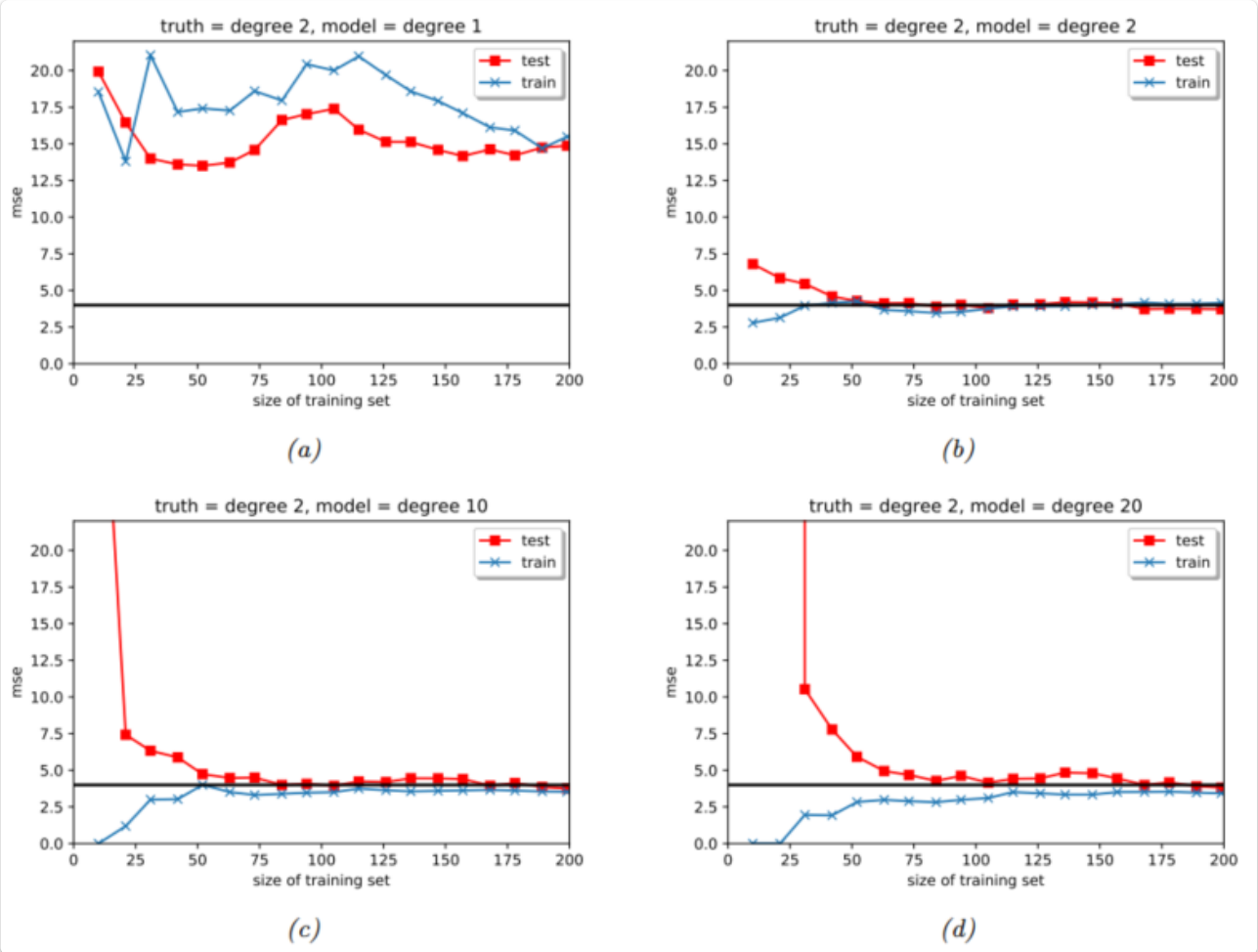


이 그림은 IMDB 영화 감성 데이터셋에서 사용된 텍스트 분류기의 퍼포먼스를 보여줍니다. 신경망 모델은 단어 임베딩의 평균 풀링을 활용하여 감성을 분류하며, 훈련 및 검증 에러를 에포크에 따라 시각적으로 나타냅니다.

- (a) 크로스 엔트로피 손실:
  - 훈련 세트(파란색) 및 검증 세트(빨간색)에 대한 크로스 엔트로피 손실을 보여줍니다.
  - 훈련 초기에 손실이 감소하다가, 검증 세트에서의 손실이 다시 증가하는 지점에서 조기 종료가 발생합니다.
- (b) 분류 정확도:
  - 훈련 세트(파란색) 및 검증 세트(빨간색)에 대한 분류 정확도를 보여줍니다.
  - 초기에 정확도가 증가하다가 일정 지점에서 더 이상 향상되지 않고 감소하기 시작하며, 조기 종료는 이러한 상황에서 훈련을 중단합니다.

이 그림은 조기 종료를 통해 모델이 훈련 세트에 지나치게 적합되어 검증 세트에서의 성능이 저하되기 전에 효과적으로 훈련을 중지하는 효과를 시각적으로 보여줍니다.

- # Using more data



- 개요:
  이 그림은 데이터 양이 증가함에 따라 (고정된 복잡도의 모델을 가정할 때) 과적합의 가능성이 감소하는 것을 보여줍니다. 특히, 훈련 세트 크기 N에 대한 다양한 모델의 훈련 및 테스트 세트에 대한 평균 제곱 오차(MSE)를 보여주는 학습 곡선(learning curve)입니다.
- 설명:

1. **축 설명:**
   - x축은 훈련 세트의 크기 N을 나타냅니다.
   - y축은 MSE(평균 제곱 오차)로, 작을수록 더 나은 성능을 나타냅니다.
2. **모델의 복잡도:**
   - 그림은 네 가지 다른 모델(다항식의 차수가 증가함에 따라)에 대한 결과를 보여줍니다.
3. **결과 관찰:**
   - **Degree 1 모델 (a):**
     - 모델이 너무 간단하여 테스트 에러가 계속 높은 수준을 유지합니다. 이를 '과소적합(underfitting)'이라고 합니다.
   - **Degree 2 모델 (b):**
     - 훈련 세트 크기가 증가함에 따라 테스트 에러가 최적 수준(잡음 바닥)으로 빠르게 수렴합니다.
   - **Degree 10 및 20 모델 (c, d):**
     - 더 복잡한 모델일수록 훈련 에러는 초기에 상승하며, 훈련 세트 크기가 커짐에 따라 테스트 에러와 수렴합니다.
     - 더 복잡한 모델은 훈련 세트에서 훈련할 때 초기에 데이터의 다양한 패턴을 처리하는 데 어려움을 겪습니다.
4. **Bayes Error 및 잡음 바닥:**
   - 수평 검은 선은 베이즈 에러를 나타내며, 이는 최적 예측자(현실 세계에서의 실제 데이터와 가장 잘 일치하는 이상적인 예측자 또는 모델)의 오차로서 내재된 잡음으로 인한 것입니다.
   - 이 예제에서는 차수 2의 다항식이 진짜 모델이며, 잡음의 분산이 $\sigma^2 = 4$입니다.
5. **결론:**
   - 모델이 더 복잡할수록 훈련 에러와 테스트 에러 간의 차이가 크지만, 훈련 세트 크기가 커짐에 따라 감소합니다.
   - 초기에 훈련 에러가 증가하는 이유는 데이터가 더 다양한 입력-출력 패턴 조합을 포함하기 때문입니다. 그러나 결국 훈련 세트는 테스트 세트와 유사해질 것이며 오차율이 수렴하여 해당 모델의 최적 성능을 반영합니다.

   이 그림은 데이터 양이 모델의 복잡도 및 훈련 에러와 테스트 에러 간의 관계에 어떻게 영향을 미치는지를 시각적으로 보여주며, 더 많은 데이터가 과적합을 완화하는 데 어떤 도움이 되는지 설명합니다.

Bayesian Statistics

- **문제상황:**
  지금까지는 데이터로부터 파라미터를 추정하는 여러 방법에 대해 논의해왔습니다. 그러나 이러한 방법은 추정값의 불확실성을 무시합니다. 추정값의 불확실성은 액티브 러닝, 오버피팅 피하기, 또는 어떤 과학적으로 의미 있는 양을 얼마나 신뢰해야 하는지 알아야 하는 등 몇 가지 응용 분야에서 중요할 수 있습니다. 통계학에서는 확률 분포를 사용하여 파라미터에 대한 불확실성을 모델링하는 것을 **추론**(Inference)이라고 합니다.
- **후방 분포 활용:**
  이 섹션에서는 **후방 분포**(Posterior distribution)를 사용하여 불확실성을 표현합니다. 이는 베이지안 통계에서 채택한 접근 방식입니다. 사전 분포(prior distribution)에서 데이터를 관측한 후의 확률 분포를 계산하기 위해 베이즈 규칙을 사용합니다.
- **후방 분포 계산 단계:**
  1. 사전 분포 (p(θ)): 데이터를 관측하기 전의 지식을 나타냅니다.
  2. 우도 함수 (p(D|θ)): 각각의 파라미터 설정에 대한 예상 데이터를 설명합니다.
  3. 주변 우도 (Marginal Likelihood): 주어진 데이터가 발생할 확률의 총 평균을 나타내는 값으로, 파라미터에 대한 적분을 포함합니다.
  4. 베이즈 규칙을 사용하여 후방 분포 계산:
     $p(\theta \mid D) = \frac{p(\theta)p(D|\theta)}{p(D)} = \frac{p(\theta)p(D|\theta)}{\int p(\theta_0)p(D|\theta_0)d\theta_0}$

     여기서 분모 $p(D)$는 주변 우도(marginal likelihood)로, 알려지지 않은 파라미터에 대한 적분을 수행하여 계산됩니다. 이는 데이터의 평균 확률로 해석할 수 있습니다.
- **결과 활용:**
  후방 분포를 계산한 후, 입력에 대한 출력의 후방 예측 분포를 알 수 있습니다. 지도/조건부 모델의 경우 다음과 같이 나타낼 수 있습니다:
  $p(y|x, D) = \int p(y \mid x, \theta)p(\theta \mid D)d\theta$

  이는 베이즈 모델 평균(Bayes Model Averaging, BMA)의 한 형태로 볼 수 있습니다. 우리는 파라미터 값이 가능성에 따라 가중치가 부여된 무한한 모델 세트를 사용하여 예측을 수행합니다. BMA 사용으로 오버피팅의 가능성이 감소합니다. 왜냐하면 최적 모델 하나만 사용하는 것이 아니기 때문입니다.

- # Conjugate priors
  - **정의:**
    우리가 폐쇄된 형태로 후방을 계산할 수 있는 (prior, likelihood) 쌍들을 고려합니다. 특히, 우리는 우도에 대해 짝으로 결합된 사전을 사용할 것입니다. 사전 $p(\theta)$이 우도 함수 $p(D|\theta)$에 대해 "Conjugate"이라고 말하는데, 이는 후방(Posterior)이 사전과 동일한 매개변수화된 패밀리에 속한다는 것을 의미합니다. 다시 말해, F가 베이지안 업데이트에 대해 폐쇄된 것입니다. 만약 패밀리 F가 지수 패밀리(exponential family)에 해당한다면, 계산은 폐쇄된 형태로 수행될 수 있습니다.
  - **공액적:**
    "공액적인(conjugate)"이라는 용어는 사전 확률분포가 주어진 가능도 함수에 대해 특

별한 관계를 가지며, 그 관계로 인해 베이지안 업데이트를 수행한 후의 사후 분포가 여전히 동일한 종류의 확률분포라는 것을 나타냅니다.

- **폐쇄된 상태:**
  우리가 사용하는 확률분포들이 특정한 종류의 관계로 인해 베이지안 업데이트를 통해 계산이 가능하고, 그 업데이트를 거치면 여전히 동일한 종류의 확률분포가 유지된다면, 이러한 확률분포들의 집합을 F라고 합니다. 이때 F가 베이지안 업데이트에 대해 폐쇄되었다고 합니다.

The beta-binomial model

- **개요:**
  베타-이항(Beta-Binomial) 모델은 동전을 N번 던져서 앞면이 나올 확률 θ를 추정하는 확률 모델입니다. 주어진 데이터 D를 통해 확률 θ에 대한 사후 분포 $p(\theta|D)$를 어떻게 계산하는지 설명하겠습니다. 각 동전 던지기의 결과를 y_n으로 나타냅니다. $y_n = 1$은 n번째 시행이 앞면이 나온 사건을 나타내고, $y_n = 0$은 뒷면이 나온 사건을 나타냅니다. 모든 데이터를 $D = y_n : n = 1 : N$로 표기하며, N번의 시행 결과를 모두 담고 있습니다. 각 시행의 결과 y_n은 베르누이 분포를 따르며, 확률 θ로 정의됩니다. 여기서 θ는 $[0, 1]$ 구간에 속하는 확률로, 동전이 앞면이 나올 확률을 나타냅니다.

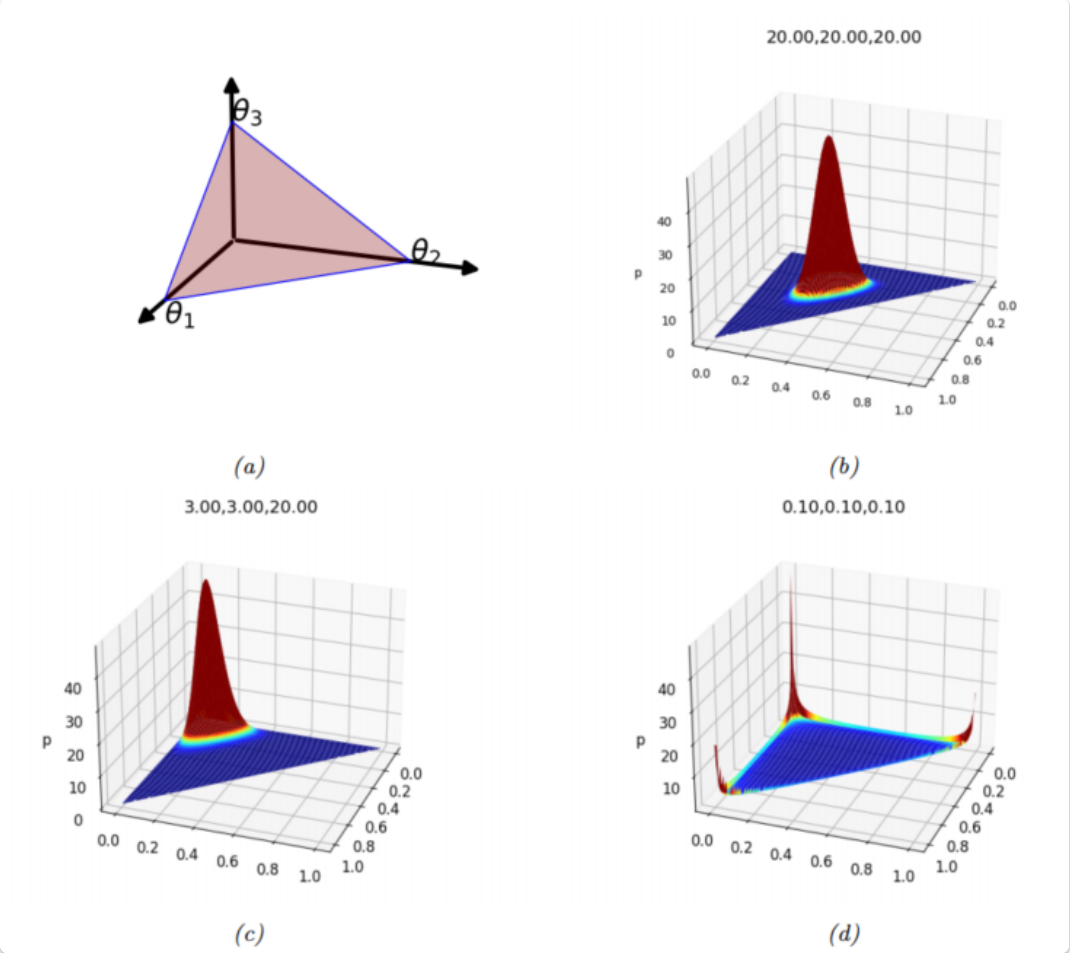
The Dirichlet-multinomial model

- **개요:**
  "디리클레-다항 분포 모델(Dirichlet-multinomial model)"에 관한 내용으로, 섹션 4.6.2에서 이진 변수(동전 던지기와 같은)에서의 결과를 K-ary 변수(주사위 던지기와 같은)로 일반화하는 것을 소개하고 있습니다.
- **정의:**
  다항 분포의 파라미터에 대한 사전 분포로서 디리클레 분포를 사용하는 것을 의미합니다. 디리클레 분포는 K-1 차원의 단순소(simplicial complex) 내의 점에 대한 분포를 정의하며, 다항 분포의 파라미터가 각각의 값을 가질 확률을 나타냅니다.
  1. **기본 설정:**
     - $Y \sim Cat(\theta)$: Y는 매개변수 벡터 θ를 가진 범주형 분포를 따르는 확률변수입니다.
  2. **가능도 함수:**
     - 가능도 함수 $p(D \mid \theta)$는 매개변수 벡터 θ가 주어졌을 때 데이터 집합 D를 관측할 확률을 나타냅니다.
     - $D = y_1, y_2, \ldots, y_N$: N개의 관측치로 이루어진 데이터셋입니다.
  3. **최종형태:**
     - $p(D \mid \theta) = \Pi_{c=1} C\theta_c^{N_c}$: 이는 범주형 분포의 매개변수 벡터 θ가 주어졌을 때 데이터 집합 D를 관측할 확률을 나타내는 최종 형태입니다. N_c는 데이터셋에서 각 범주가 나타나는 횟수를 나타냅니다.

Prior

- **정의:**
  범주형 분포(Categorical distribution)의 공액 사전 분포(Conjugate prior)로서 디리클레 분포(Dirichlet distribution)를 소개하고 있습니다. 디리클레 분포는 베타 분포의 다변량 확장으로, 확률의 단순한 형태를 보다 복잡한 다차원 공간으로 일반화한 분포입니다. 이 분포는 다차원 확률 변수에 대한 사전 분포로서 특히 범주형 분포와의 공액성(conjugacy)이 있습니다.

1. **디리클레 분포의 정의:**
   - 디리클레 분포는 다음과 같은 영역에서 정의됩니다.
     $S_K = \theta : 0 \leq \theta k \leq 1, \sum_{k=1} K\theta_k = 1$
     이는 확률 변수 벡터 θ가 0과 1 사이에 있으며, 모든 성분의 합이 1이 되는 영역입니다.
2. **디리클레 분포의 확률 밀도 함수(PDF):**
   - 디리클레 분포의 확률 밀도 함수는 다음과 같이 정의됩니다.
     $Dir(\theta \mid \alpha) = \frac{1}{B(\alpha)} \Pi_{k=1}^{K} \theta_k^{\alpha_k - 1} I(\theta \in SK)$
     여기서 α는 분포의 매개변수이며, $B(\alpha)$는 다변량 베타 함수로 정의되어 있습니다.
3. **다변량 베타 함수:**
   - 다변량 베타 함수는 다음과 같이 정의됩니다.
     $B(\alpha) = \frac{\Pi_{k=1} K\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$



K=3일 때 디리클레 분포가 3차원 공간에서 정의되며, 이를 표면으로 나타내어 simplex(단순체) 상의 분포를 보여주고 있다고 설명하고 있습니다. 각 지점은 다음의 조건을 만족합니다: $0 \leq \theta k \leq 1$ 그리고 $\sum_{k=1}^{3} \theta_k = 1$. 이것은 삼각형 모양의 표면으로 시각적으로 표현됩니다.

간단한 설명을 위해 각 경우에 대한 그림을 설명하겠습니다:

1. **삼각형 표면:**
   - $K = 3$일 때, 디리클레 분포는 3차원 simplex 상의 분포를 정의합니다. 이 삼각형은 각 θ_k가 0과 1 사이에 있고, 세 성분의 합이 1이 되는 모든 지점을 나타냅니다.
2. **디리클레 분포: α=(20,20,20):**
   - 이 경우에는 디리클레 분포의 매개변수 α가 모두 20으로 같은 값을 가지는 경우입니다. 이는 (1/3, 1/3, 1/3)을 중심으로 하는 좁은 분포를 보여줍니다.
3. **디리클레 분포: α=(3,3,20):**
   - 이 경우에는 매개변수 α 중 하나가 다른 것들에 비해 상대적으로 큰 값을 가집니다. 결과적으로 이는 한 쪽으로 기울어진 형태의 비대칭한 분포를 보여줍니다.
4. **디리클레 분포: α=(0.1,0.1,0.1):**

- 모든 매개변수가 매우 작은 경우로, 각 코너에서 뾰족한 "스파이크"가 나타나는 형태입니다. 이는 특정 값이 나타날 확률이 높고, 나머지는 거의 나타나지 않는 희소한 분포를 보여줍니다.

이러한 시각적인 표현을 통해 디리클레 분포의 매개변수가 분포의 형태에 어떻게 영향을 미치는지를 이해할 수 있습니다. 뾰족한 정도와 중심의 위치는 α 값에 따라 달라지며, 이는 확률분포의 특성을 조절하는 데 사용됩니다.

The Gaussian-Gaussian model

- **정의:**
  가우시안-가우시안(Gaussian-Gaussian) 모델에 대한 파라미터의 사후 분포를 도출합니다. 이 모델에서는 가우시안 분포의 파라미터에 대한 베이지안 추론을 살펴봅니다. 단순함을 위해 분산이 알려져 있다고 가정합니다.

Likelihood Function:

- 분산이 알려져 있다고 가정하면, 주어진 평균 μ에 대한 데이터 집합 D의 우도 함수는 다음과 같습니다:
  $p(D \mid \mu) \propto exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \mu)^2 \right)$
  여기서 y_n은 데이터 포인트이고, N은 데이터의 개수입니다.

Beyond conjugate priors

1. **Noninformative Priors:**
   - 도메인 특정 지식이 적거나 전혀 없는 경우, 비정보 사전은 데이터에게 더 큰 역할을 할 수 있도록 하는 것이 좋습니다.
   - 예를 들어, 실수 값 매개변수 $\mu \in R$를 추론하는 경우, $p(\mu) \propto 1$과 같은 평평한 사전을 사용할 수 있습니다. 이는 "무한히 넓은" 가우시안으로 해석될 수 있습니다.
   - "비정보"라는 용어는 오해의 소지가 있을 수 있으므로 종종 희석 사전(diffuse prior), 최소한의 정보 사전(minimally informative prior), 또는 기본 사전(default prior)과 같은 용어를 사용하는 것이 좋습니다.
2. **Hierarchical Priors:**
   - 베이지안 모델에서는 매개변수에 대한 사전 $p(\theta)$를 지정해야 합니다. 이러한 매개변수들을 하이퍼매개변수(hyperparameters)라고 하며, 이들에 대한 사전을 설정할 수 있습니다.
   - 이로써 계층적 베이지안 모델 또는 다수계층 모델이 형성되며 이는 다음과 같이 시각화될 수 있습니다: $\xi \rightarrow \theta \rightarrow D$.
   - 하이퍼매개변수에 대한 사전이 고정된 경우(예: 어떤 종류의 최소한의 정보 사전을 사용할 수 있음), 결합 분포는 다음과 같은 형태를 갖습니다:
     $p(\xi, \theta, D) = p(\xi)p(\theta \mid \xi)p(D \mid \theta)$.
   - 하이퍼매개변수를 데이터 포인트로 취급하여 하이퍼매개변수를 학습할 수 있으며, 이는 여러 관련 매개변수를 추정해야 하는 경우(예: 서로 다른 하위 모집단 또

는 여러 작업에서)에 유용합니다.

3. **Empirical Priors:**
    - 계층적 베이지안은 데이터에서 매개변수를 추론하는 방법으로 유용하지만, 이러한 모델에서의 사후 추론은 계산적으로 어려울 수 있습니다.
    - 이 섹션에서는 하이퍼매개변수의 점 추정치를 먼저 계산하고, 그 후에 조건부 사후인 $p(\theta \mid \xi, D)$를 계산하는 근사화를 논의합니다. 이는 결합 사후인 $p(\theta, \xi \mid D)$를 계산하는 대신에 사용됩니다.
    - 하이퍼매개변수를 추정하기 위해 주변 우도(marginal likelihood)를 최대화하는 기술을 사용하며, 이를 2형 최대우도(type II maximum likelihood)라고 합니다: $\hat{\xi}_{mml}(D) = argmax_{\xi} p(D \mid \xi)$.
    - $\hat{\xi}$를 얻은 후에는 일반적인 방법으로 조건부 사후 $p(\theta \mid \hat{\xi}, D)$를 계산합니다.

- ## Credible intervals
    - **정의:**
      포스터리어 분포는 대개 시각화하고 다루기 어려운 고차원 객체입니다. 이 분포를 요약하는 일반적인 방법 중 하나는 포스터리어 평균이나 모드와 같은 포인트 추정치를 계산하고, 그 추정치와 관련된 불확실성을 나타내는 신뢰구간을 계산하는 것입니다.
        - **신뢰구간의 정의:**
          $100(1-\alpha)\%$의 신뢰구간은 포스터리어 확률 질량의 $1-\alpha$를 포함하는 구간입니다.
    - **중심 구간과 모수가 알려진 경우:**
      포스터리어가 알려진 함수 형태를 갖는 경우, 중심 구간은 누적 분포 함수(CDF)와 역누적 분포 함수(inverse CDF)를 사용하여 계산할 수 있습니다.
- ## Bayesian machine learning
    - **정의:**
      지금까지는 $p(y|\theta)$ 형태의 무조건 모델에 중점을 두었습니다. 그러나 지도학습(supervised machine learning)에서는 $p(y|x,\theta)$ 형태의 조건부 모델을 사용합니다. 이때 매개변수에 대한 사후 분포는 이제 $p(\theta|D)$가 되며, 여기서 $D = (xn, yn) : n = 1 : N$입니다. 이러한 사후 분포를 계산하는 것은 이미 논의한 원리를 사용하여 수행할 수 있습니다. 이러한 방식은 베이지안 기계 학습(Bayesian machine learning)이라고 불리며, 모델 매개변수에 대해 '베이지안'이라는 개념을 적용하는 것입니다.
- ## Computational issues

Frequentist Statistics

- **정의:**
  베이지안 통계는 모델 매개변수를 다른 미지의 랜덤 변수와 마찬가지로 취급하고 이를 데이터로부터 추론하기 위해 확률 이론의 규칙을 적용합니다. 그러나 매개변수를 랜덤 변수처럼

취급하지 않고, 따라서 사전과 베이즈 규칙을 사용하지 않는 **통계 추론 접근 방식**이 있습니다. 이 대안적인 접근 방식은 빈도주의 통계, 고전적 통계 또는 정통 통계라고 알려져 있습니다.

- **기본 아이디어:**
  - 빈도주의 통계의 기본 아이디어는 데이터에서 추정된 어떤 양(매개변수 또는 예측된 레이블과 같은)이 데이터가 변경될 때 어떻게 변할지 계산하여 불확실성을 표현하는 것입니다.
  - 이 반복된 시행 간의 변이 개념은 빈도주의 접근에서 불확실성을 모델링하는 기초를 형성합니다.
  - 반면에 베이지안 접근은 확률을 반복된 시행이 아닌 정보 관점에서 본다. 이는 우리가 논의한 대로 베이지안이 일회성 이벤트의 확률을 계산할 수 있게 합니다.
  - 더 중요한 것은 베이지안 접근이 빈도주의 접근에 문제를 일으키는 특정 패러독스를 피할 수 있다는 점입니다.

Sampling distributions

- **개요:**
  빈도주의 통계에서는 불확실성이 랜덤 변수의 사후 분포가 아닌 추정량의 표본 분포에 의해 나타납니다.
- **추정량(estimator):**
  - "추정량"은 의사결정 이론 섹션에서 정의되었으며 간단히 말하면, 관측된 데이터가 주어졌을 때 어떤 행동을 취할지를 지정하는 의사결정 절차입니다.
  - 행동은 클래스 레이블을 예측하거나 다음 관측값을 예측하거나 알려지지 않은 매개변수를 예측하는 것일 수 있습니다.
  - 후자의 경우 추정량은 종종 $\hat{\theta}$로 나타내지만, 이 표기법은 매개변수 벡터를 나타내는 것처럼 보이기 때문에 모호합니다. 대신 우리는 $\hat{\Theta}$ 표기법을 사용할 것입니다. 이 함수는 MLE나 모멘트 방법 추정 등을 계산할 수 있습니다.
  - 이 함수를 특정 크기의 데이터셋에 적용하면 $D = x1, \ldots, xN$이 되며, 이때 함수의 출력은 $\hat{\theta} = \hat{\Theta}(D)$로 나타낼 수 있습니다.
- **표본 분포(sampling distribution):**
  - 빈도주의 통계의 핵심 아이디어는 데이터 D를 랜덤 변수로 보고, 데이터가 뽑힌 매개변수 θ^*를 고정된 but 알려지지 않은 상수로 보는 것입니다.
  - 따라서 $\hat{\theta} = \hat{\Theta}(D)$는 랜덤 변수이며, 이의 분포를 추정량의 표본 분포라고 합니다.
  - 이것이 의미하는 바를 이해하기 위해 S개의 다른 데이터셋을 생성하고, 각각 $D(s) = xn \sim p(xn|\theta*) : n = 1 : N$ 형태로 표현해 봅시다. (이를 간략하게 $D(s) \sim \theta^*$로 나타냅니다.)
  - 이제 각 D(s)에 추정량을 적용하여 추정치 세트 $\hat{\theta}(D(s))$를 얻습니다. $S \to \infty$로 가면 이 세트에서 유발되는 분포가 추정량의 표본 분포입니다.

- 더 정확히는 다음과 같습니다:
  $SamplingDist(\hat{\Theta}, \theta^*) = PushThrough(p(D|\tilde{\theta}^*), \hat{\Theta})$
- 여기서 데이터 분포를 추정 함수를 통과하여 추정치의 분포를 유도합니다. 경우에 따라 우리는 표본 분포를 해석적으로 계산할 수 있지만, 일반적으로는 몬테카를로를 사용하여 근사화해야 합니다.
- **몬테카를로?**
  몬테카를로 방법(Monte Carlo method)은 확률적인 실험을 통해 수학적인 문제를 푸는 계산 방법입니다. 몬테카를로 방법은 난수를 생성하여 무작위 표본을 추출하고 이를 사용하여 함수나 문제의 값을 추정합니다. 이 방법은 확률적이며 통계적인 기법을 사용하여 정확한 해답을 찾기 어려운 문제에 대한 근사해를 찾는 데 사용됩니다.

  명칭은 몬테카를로 카지노에서 주사위를 굴려 승패를 결정하는 과정에서 유래되었습니다. 몬테카를로 방법은 1940년대에 원자폭탄의 폭발 효과를 계산하는데 처음 사용되었으며, 이후 다양한 과학 및 엔지니어링 분야에서 폭넓게 적용되고 있습니다.

표본 분포의 의미:

- 표본 분포를 통해 우리는 추정치가 다양한 데이터셋에서 어떻게 변할지를 이해할 수 있습니다.
- 빈도주의 접근에서는 매개변수가 고정되었다고 가정하고 이를 기반으로 불확실성을 모델링합니다.

Reference

Log-Odds
EWMA