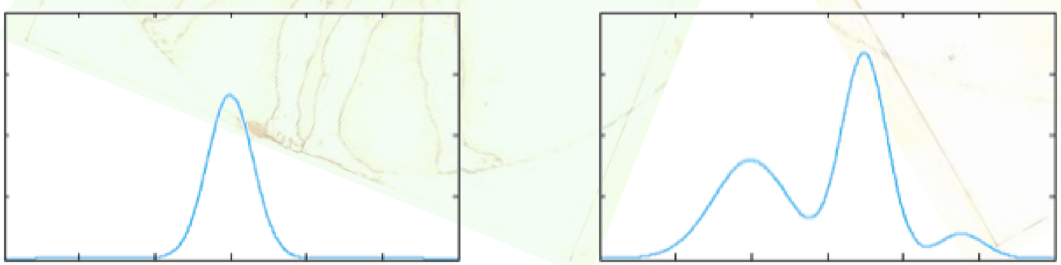
Ch. 3 확률 분포 추정

Pattern Recognition & Machine Learning Laboratory
Ha-na Jo

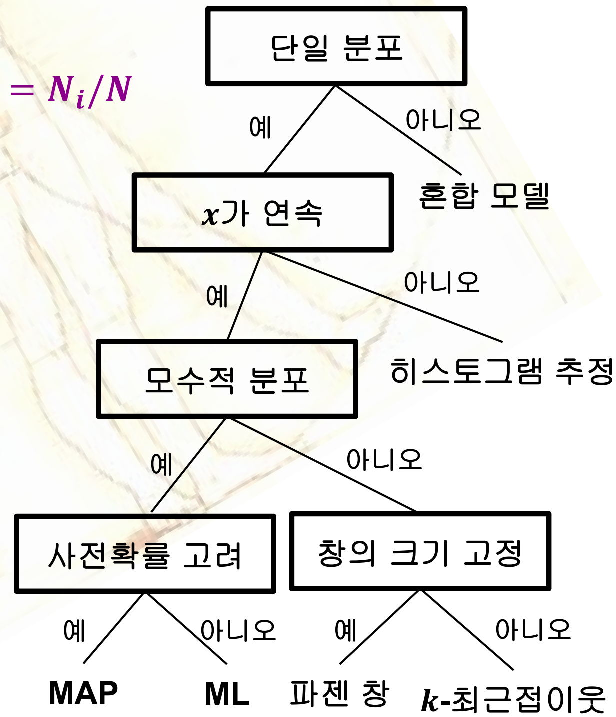
July 7, 2021

Introduction

- 단원 목표
  - 2장 베이시언 분류기에서 알고 있다고 가정한 사전 확률과 우도를 추정하는 것이 목표
- 사전 확률의 추정
  - $P(\omega_i)$
  - X의 크기가 충분히 클 때, 사전 확률은 실제 값에 근접
    - Ex) X의 크기가 N, ω_i에 속하는 샘플 수가 N_i일 때, $P(\omega_i) = N_i/N$
- 우도의 추정
  - $p(x|\omega_i)$
  - 정규 분포의 형태라면, 정규 분포 매개 변수 추정 문제
    - 임의의 형태라면 히스토그램 추정 사용
  - 다른 부류의 샘플은 서로 영향을 미치지 않는다고 가정
    - X_i로 $p(x|\omega_i)$를 추정하는 문제를 $X, p(x)$로 표기



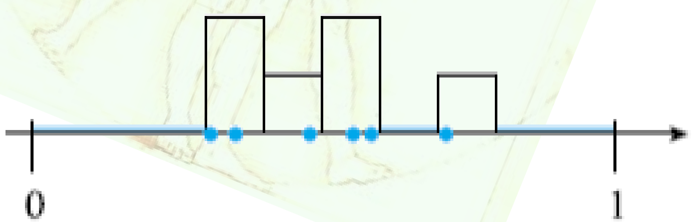
우도의 분포 형태 예



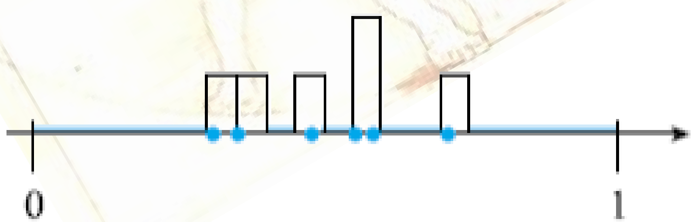
우도 추정의 여러 방법을 구별하는 트리

히스토그램 추정

- 방법
  - 우도의 분포가 임의의 모양일 때 사용
  - 샘플의 범위를 N개의 구간으로 나누고 구간 별로 그 안의 샘플의 수를 확인
    - 나눈 구간은 **bin (빈)** 이라고 명칭
  - 확률 분포로 사용하기 위해서는 각 빈의 값을 N으로 나누어 정규화
  - 표현과 연산이 단순하면서 분포의 특성을 잘 표현
- 한계
  - 이산 확률 분포
  - 현실적인 적용에는 확률 분포가 정의되는 공간이 낮은 차원, 충분히 큰 X 필요
    - X의 크기가 작을 때 구간이 많으면 적은 구간만 확률 값 존재
      - 의미 없는 희소한 히스토그램
    - 차원이 증가함에 따라 빈의 개수는 지수적으로 증가
  - 동적 구간을 각각 **10**개, **20**개로 나눈 **1**차원 예



동적 구간을 **10**개로 나눔



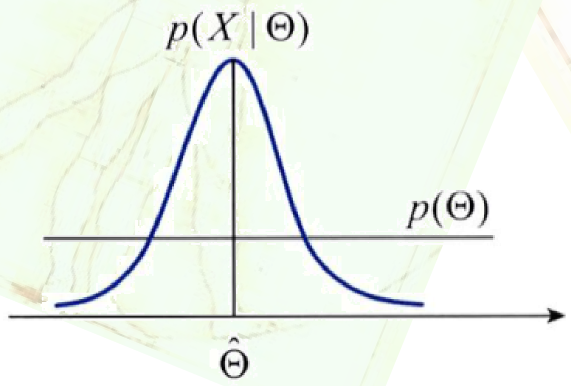
동적 구간을 **20**개로 나눔

최대 우도 (1/2)

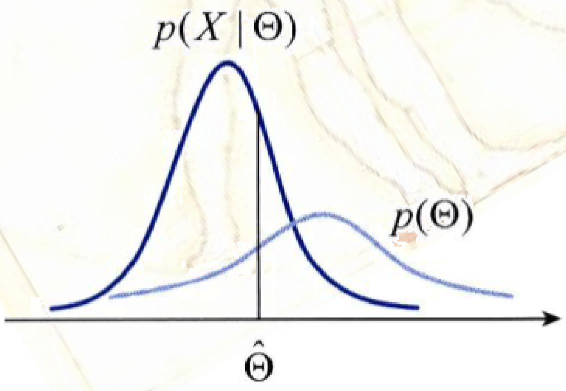
- 목표
  - 우도를 최대로 만드는 $\boldsymbol{\Theta}$를 찾는 것이 목적
    - $\hat{\boldsymbol{\Theta}} = \boldsymbol{arg\ max\ p(X|\Theta)}$
  - 개념적으로 어떠한 형태의 분포에도 적용 가능
  - 현실적으로는 정규 분포와 같이 매개 변수로 표현되는 경우만 가능
    - 가장 큰 값을 실제 계산으로 찾아야 하기 때문에 매개 변수 필요
  - 모든 샘플은 독립적으로 추출되었다고 가정
    - $\boldsymbol{p(X|\Theta) = p(x_1|\Theta)p(x_2|\Theta)\cdots p(x_N|\Theta) = \prod_{i=1}^{N} p(x_i|\Theta)}$
      - $\boldsymbol{X}$는 훈련 집합으로 $\boldsymbol{X = \{x_1, x_2, \cdots, x_N\}}$
- 로그 우도
  - 우도에 단조 증가 함수인 $\boldsymbol{ln}$을 취한 것
    - $\boldsymbol{f(.)}$가 단조 증가 함수이면 $\boldsymbol{arg\ max\ p(X|\Theta) = arg\ max\ f(p(X|\Theta))}$
    - $\hat{\boldsymbol{\Theta}} = \boldsymbol{arg\ max \sum_{i=1}^{N} ln\ p(x_i|\Theta)}$
      - $\boldsymbol{log}$함수의 장점 : 곱셈, 나눗셈을 덧셈, 뺄셈으로 변환 가능, 지수 제거 가능
- 최적화 문제
  - 미분을 이용한 최적화 알고리즘을 사용
  - 미분의 성질에 따르면 $\hat{\boldsymbol{\Theta}}$는 미분한 값이 **0**
    - 만족하는 해가 여러 개이면 그 중 가장 큰 값을 선택

최대 우도 (2/2)

- **Maximum a posterior (MAP)**
  - 사전 확률을 고려하여 최적의 매개 변수를 찾는 방법
    - $\hat{\Theta} = arg\ max\ p(\Theta) \sum_{i=1}^{N} ln\ p(x_i|\Theta)$
  - 최대 우도법 (ML)의 경우, 사전 확률이 균일하다는 가정이 존재
    - 그러나, 사전 확률이 균일하지 않는 경우도 존재
  - MAP의 경우, 균일하지 않은 사전 확률
    - 사전 확률에 따라 최적해에 영향
  - ML과 MAP의 비교 예
    - MAP의 경우 최적해 ($\hat{\Theta}$)가 ML의 최적해 ($\hat{\Theta}$)보다 오른쪽으로 치우친 형태



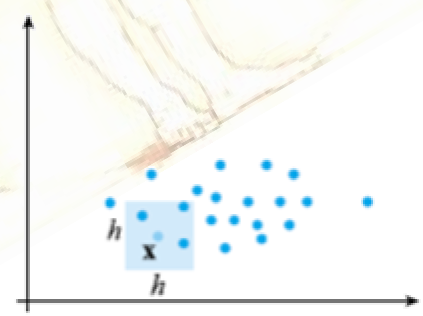
ML 방법



MAP 방법

비모수적 방법 (1/2)

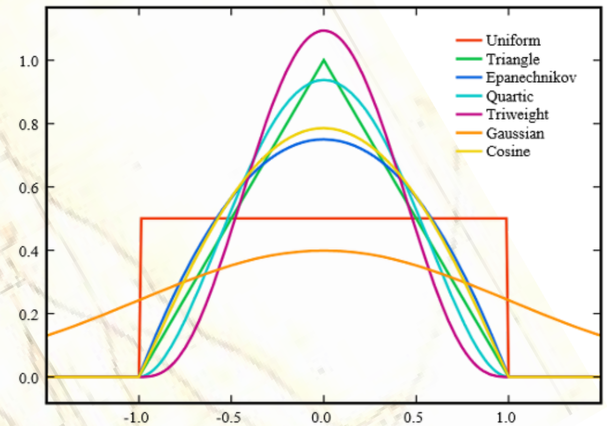
- 모수적 방법
  - **ML, MAP** 방법이 모수적 방법
  - 매개 변수 (모수)로 표현할 수 있는 특정한 확률 분포에서만 적용 가능한 한계 존재
  - 실제로는 특정한 확률 분포를 따르지 않는 경우가 다수
- 비모수적 방법
  - 임의의 확률 분포에서 적용 가능한 방법
    - 예) 파젠 창 방법, 최근접 이웃 방법
- 파젠 창
  - 히스토그램 추정에서 확장
    - 이산 확률 분포가 아닌 확률 밀도 구하기 가능
  - 임의의 점 x을 중점으로 하는 크기 h인 창을 씌우고 그 안의 샘플 개수를 세는 방법
    - 1차원 특징 공간의 경우, $p(x) = \frac{1}{h}\frac{k}{N}$
      - h: 창의 크기, k: 샘플의 개수, N: 전체 샘플의 개수
    - d차원 특징 공간의 경우, $p(x) = \frac{1}{h^d}\frac{k_x}{N}$
      - k의 경우 x에 따라 변하므로 k_x로 표기
  - 계단 모양의 확률 밀도 함수 발생



2차원 특징 공간에서의 파젠 창 방법

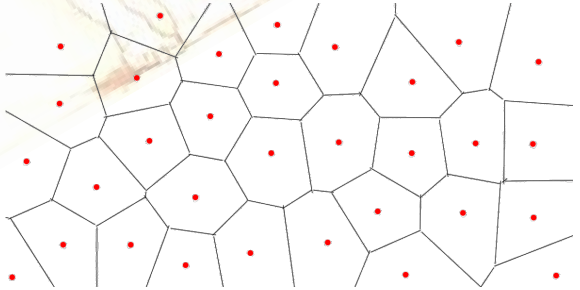
비모수적 방법 (2/2)

- 파젠 창 (Cont.)
  - 커널 함수
    - 파젠 창 방법의 매끄럽지 않은 확률 밀도 함수를 보안
    - x와 더 가까운 샘플에 가중치를 부여
      - $\kappa(x) \geq 0$ 과 $\int_x \kappa(x)dx = 1$ 만족
  - N이 충분히 크고, h가 충분히 작으면 실제와 유사
    - 현실적으로는, N가 고정
    - 최적의 h를 구하는 방법은 실험적으로 파악
  - 취약점
    - 특징 공간의 크기가 작은 경우 거의 모든 점에서 $p(x) = 0$
    - 필요한 샘플이 차원이 커짐에 따라 지수적으로 증가



커널 함수의 예

- k-최근접 이웃 추정
  - x를 중심으로 창을 씌우고 k개의 샘플이 들어올 때까지 창의 크기를 확장하는 방법
    - $p(x) = \frac{1}{h_x{}^d}\frac{k}{N}$
    - 파젠 창은 창 고정, k-최근접 이웃 추정은 창 크기 변동
    - 차원이 높을 때 계산량이 많고 시간 복잡도가 $\Theta(kdN)$
      - 특징 공간을 미리 나누는 보르노이 도형 방법 활용 가능


보르노이 도형의 예

혼합 모델 (1/4)

- 목표
  - 두 개 이상의 서로 다른 확률 분포의 혼합으로 X를 모델링하는 것이 목표
- 가우시언 혼합
  - 여러 개의 모드를 가진 경우, 여러 개의 가우시언을 혼합
    - 혼합할 요소 분포는 어떤 분포도 가능하나, 가우시언으로 국한
  - 다음의 매개 변수 추정이 필요
    - 가우시언의 개수 K
      - 고정되어 있다고 가정
    - k번째 가우시언의 매개 변수
      - K개 각각의 평균 벡터와 공분산 행렬
    - k번째 가우시언의 가중치 π_k (혼합 계수)
      - 가우시언 각각의 영향력
      - 확률 값이므로, $0 \le \pi_k \le 1$ 과 $\sum_{k=1}^{K} \pi_k = 1$ 만족
  - 가중치가 포함된 식
    - $p(x) = \sum_{k=1}^{k} \pi_k N(x|\mu_k, \Sigma_k)$
    - 주어진 값 $X = \{x_1, x_2, \cdots, x_N\}$
    - 추정할 값 $\Theta = \{\pi = (\pi_1, \cdots, \pi_k), (\mu_k, \Sigma_k), \cdots, (\mu_k, \Sigma_k)\}$

혼합 모델 (2/4)

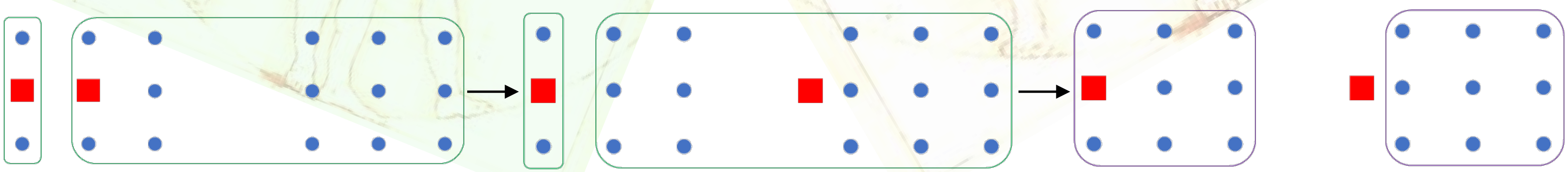
- 가우시언 혼합 (Cont.)
  - 최대 우도 추정 (로그 우도)
    - $ln\, p(X|\Theta) = \sum_{i=1}^{N} ln(\sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k))$
      - $\hat{\Theta} = arg\, max\, ln\, p(X|\Theta)$
      - X에 대해 최대 우도를 갖는 매개 변수 집합 Θ
- EM 알고리즘
  - K개의 가우시언과 혼합 계수 π를 추정하는 최적화 문제 해결 방법
    - 최대 우도 추정과 같이 미분을 이용한 방법 적용으로는 해결 불가
  - E (Expectation) 단계
    - 샘플이 어느 가우시언에 속하는지 추정하는 단계
    - 은닉 변수 z 사용
      - 샘플 x_i가 j 번째 가우시안에서 발생했다고 가정, $p(x_i|z_j = 1) = N(x_i|\mu_j, \Sigma_j)$
    - 소속 정도를 확률로 표현하는 연성 소속 방법 이용
      - 조건부 확률 (사후확률) $P(z_j = 1|x_i) = \frac{P(z_j=1)p(x_i|z_j = 1)}{p(x_i)} = \frac{\pi_j N(x_i|\mu_j, \Sigma_j)}{\sum_{k=1}^{K} \pi_j N(x_i|\mu_j, \Sigma_j)}$
  - M (Maximization) 단계
    - 매개 변수 집합 Θ (μ, Σ, π)의 추정 단계
    - E와 M 단계를 번갈아 반복하다 수렴 조건이 반복되면 종료
      - 수렴 조건은 로그 우도의 값이 이전 것에 비해 좋아지지 않으면 수렴했다고 결정

혼합 모델 (3/4)

- EM 알고리즘 (Cont.)
  - j번째 가우시언의 평균 벡터 μ_j 추정
    - 미분의 성질에 따라, $ln\, p(X|\Theta)$는 최대 점에서 μ_j 로 미분한 값이 0
    - 미분 결과를 정리하면, $\mu_j = \frac{1}{N_j}\sum_{i=1}^{N} P(z_j = 1|x_i),\ N_j = \sum_{i=1}^{N} P(z_j = 1|x_i)$
      - N_j : j번째 가우시언에 소속된 샘플의 개수
      - μ_j : j번째 가우시언에 속할 확률을 가중치로 두는 평균 벡터
  - j번째 가우시언의 공분산 행렬 $\sum_j$ 추정
    - 미분의 성질에 따라, $ln\, p(X|\Theta)$는 최대 점에서 $\sum_j$ 로 미분한 값이 0
    - 미분 결과를 정리하면, $\sum_j = \frac{1}{N_j}\sum_{i=1}^{N} P(z_j = 1|x_i)\,(x_i - \mu_j)(x_i - \mu_j)^{T}$
      - $\sum_j$: j번째 가우시언에 속할 확률을 가중치로 두는 공분산 행렬
  - j번째 가우시언의 혼합 계수 벡터 π_j 추정
    - 혼합 계수는 $0 \leq \pi_k \leq 1$ 과 $\sum_{k=1}^{K} \pi_k = 1$ 를 만족해야 하므로, 제약이 발생
    - 조건부 최적화 문제
      - 제약이 있는 최적화 문제
      - 라그랑제 승수법을 이용, 최적화하려는 값에 라그랑제 승수 (α) 항을 더하여 해결
      - $ln\, p(X|\Theta) + \alpha(\sum_{k=1}^{K} \pi_k - 1)$

혼합 모델 (4/4)

- **EM 알고리즘 (Cont.)**
  - 라그랑제 승수법을 이용한 식을 최대 점에서 π_j 로 미분한 값이 **0**
  - 미분 결과를 정리하면, $\pi_j = -\frac{N_j}{\alpha}$
  - $\sum_{k=1}^{K} \pi_k = 1$를 적용하면, $\alpha = -N$, $\pi_j = \frac{N_j}{N}$
    - π_j : 소속된 샘플의 개수를 전체 샘플의 개수로 나눈 것
- **EM 알고리즘 부연 설명**
  - 낮은 수렴 속도로 인해 k-평균 알고리즘의 결과를 초기값으로 사용 가능
    - k-평균 알고리즘 : 샘플 각각을 가장 가까운 프로토타입에 할당한 뒤, 각 프로토타입을 자신에게 배정된 샘플의 평균으로 대치하는 알고리즘
  - 그리디 알고리즘이므로 전역 최적 해가 아닌 지역 최적 해로 수렴 가능
    - 그리디 알고리즘 : 탐색 공간 전체가 아닌, 현재 해의 이웃을 조사하고 이동하는 알고리즘
  - 불완전한 데이터가 주어진 경우를 위한 최대 우도 추정법의 일종
    - 불완전한 데이터는 샘플의 가우시안 소속 정보를 모르는 것에서 발생



k-평균 알고리즘의 예