확률추론에 기반한 데이터분석

서울대학교 PLSE

김지훈

확률추론

- Prior : P(M). 우리가 알고 있던 지식
- Posterior : P(M|D). 데이터 D를 통해 배운 지식
- 확률 추론 : Prior, D -> Posterior
- 확률 추론 ~ 베이지안 추론, 통계 추론

프로그램 vs 확률추론

- 프로그램 ~ 확률추론
- 프로그램 스펙 ~ 확률 모형(확률함수 정의)
- 구현 ~ 확률 계산

프로그램 vs 확률추론

- 컴퓨터 프로그램을 필요로 하는 분야가 여전히 많음
- 프로그램이 필요한 사람도
  프로그램의 스펙을 정의하기 힘듬
- 프로그램의 스펙을 정하고 나서도
  프로그램을 구현하는것은 쉽지 않음
- 프로그램 스펙은 계속 변화함

프로그램 vs 확률추론

- 확률추론을 필요로 하는 분야가 여전히 많음
- 확률추론이 필요한 사람도
  계산해야 할 확률함수를 정의하기 힘듬
- 확률모형을 정의하고 나서도
  해당 확률을 계산하는것은 쉽지 않음
- 확률모형은 계속 변화함

확률추론 실제 사례

- 확률추론은 많은 분야에 쓰이고 있고,
  더 많이 쓰일것으로 예상

- 올바른 확률모형을 세우는것이 중요

- 문제있는 확률추론은
  오류있는 프로그램만큼이나 많음

- 확률추론은 분석 정확도를 높이는데도 유용

예제 1 : 임상 실험

	처방A	처방B
증상 1	**93%(81/87)**	87%(234/270)
증상 2	**73%(192/263)**	69%(55/80)

Example from Wikipedia

심슨의 역설

	처방A	처방B
증상 1	93%(81/87)	87%(234/270)
증상 2	73%(192/263)	69%(55/80)
종합	78%(273/350)	**83%(289/350)**

Example from Wikipedia

확률추론 해보기

- 확률함수 정의하기:

  P(처방|환자) = P(처방|증상1) × P(증상1|환자)

  + P(처방|증상2) × P(증상2|환자)

- 올바른 확률 추론의 필수 요소:

  데이터 & 적절한 확률 모형

예제 2 : 통계 조사

- 역대 스타크래프트 게임대회 우승자중
  대부분이 라면과 김치를 즐기는것으로 나타나..

- 라면과 김치를 먹으면 스타크래프트를 잘한다?

확률추론 해보기

- 확률 함수 정의하기 :

  P(실력, 입맛|선수)

  = $\Sigma_{\text{국적}}$ P(실력 | 국적) P(입맛 | 국적) P(국적|선수)

- 국적 정보 X: 실력과 입맛 간에 상관관계 O
- 국적 정보 O: 실력과 입맛은 서로 독립

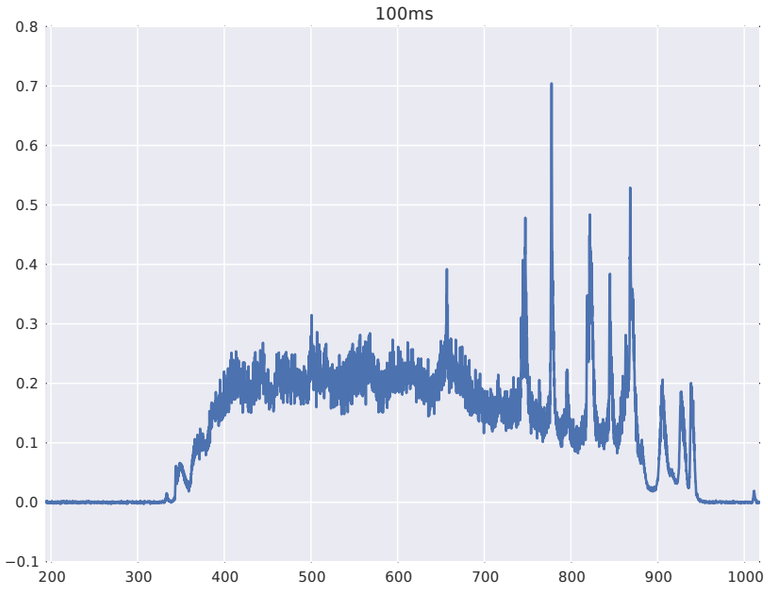
같은 데이터 & 다른 해석

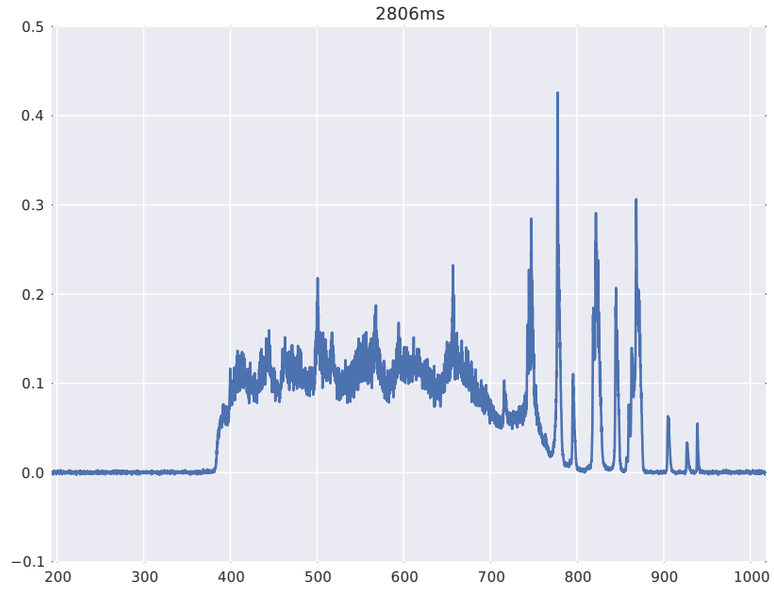
- 조금더 민감한(?) 예제들
  - 성공한 사람들의 비결…
  - 게임은 사람을 폭력적으로 만들어…
  - A를 섭취하면 암 걸릴 위험이 적은것으로…
- 주장의 근거를 잘 살펴봐야 함

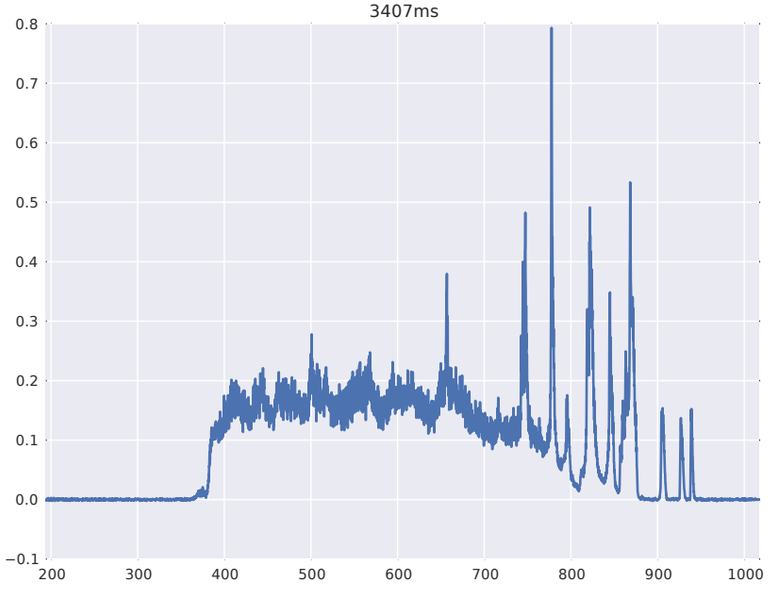
예제3 : 샘플 분류

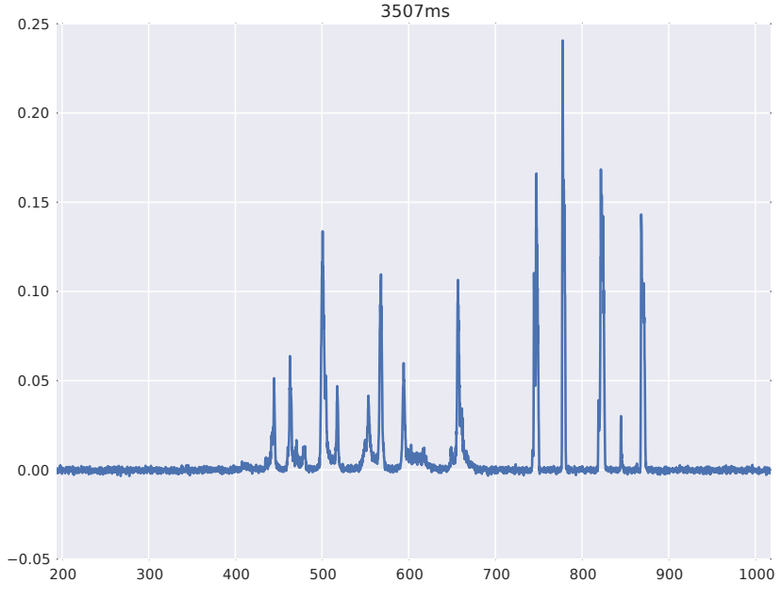
- 목표 : 스펙트럼 데이터를 보고 샘플 종류 추정하기
- 확률 함수 : P(샘플종류 | 측정데이터=d)

샘플 1

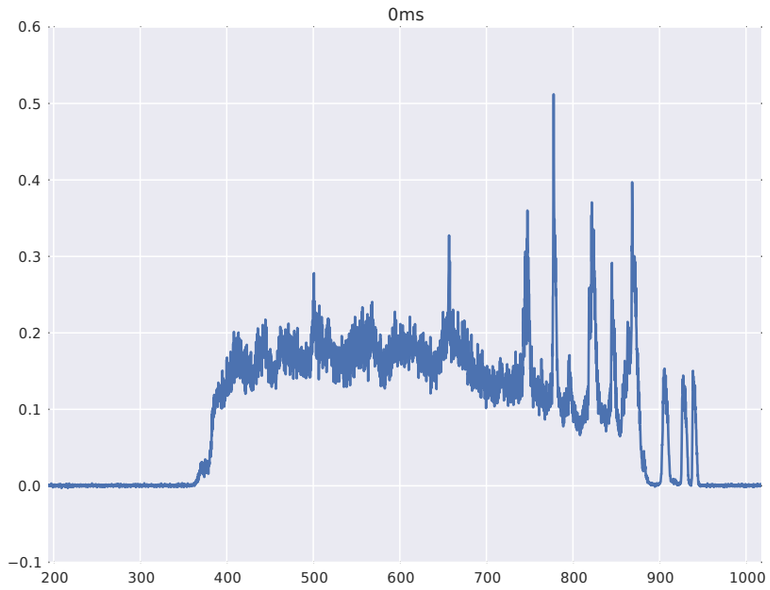

100ms
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
200
300
400
500
600
700
800
900
1000

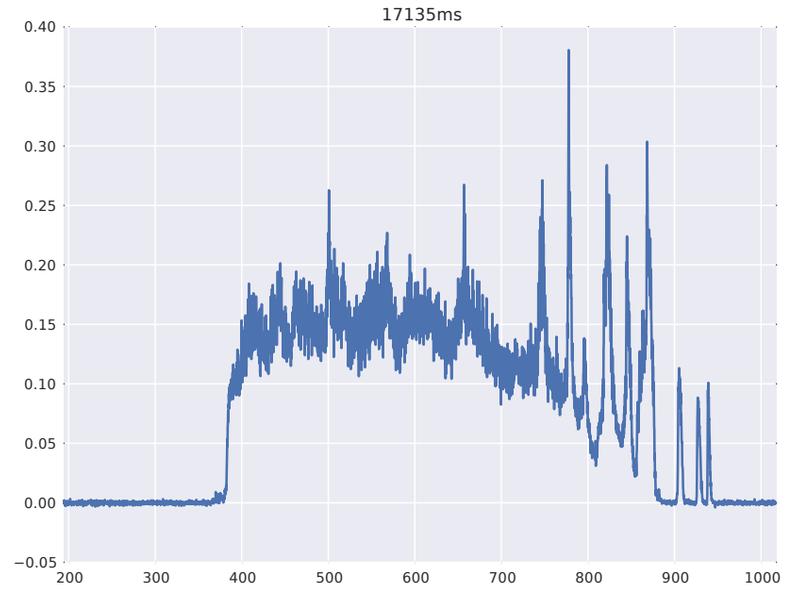

2806ms
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
200
300
400
500
600
700
800
900
1000

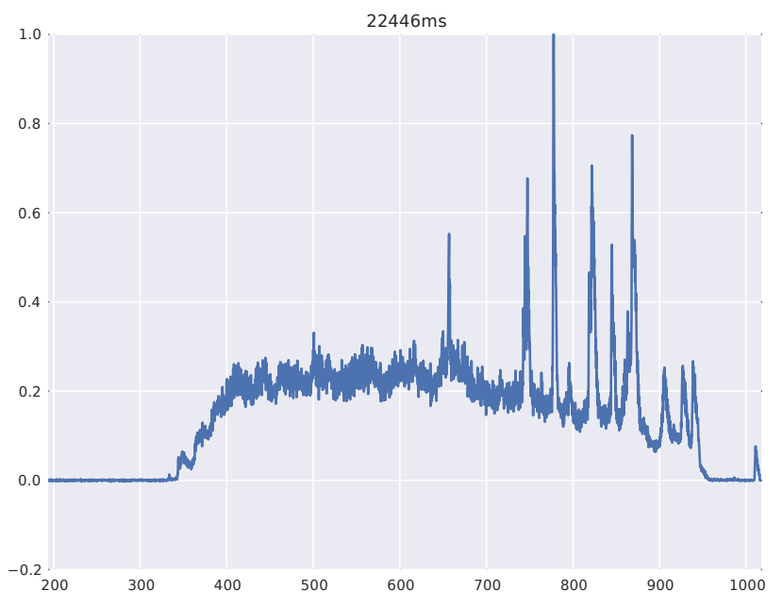

3407ms
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
200
300
400
500
600
700
800
900
1000

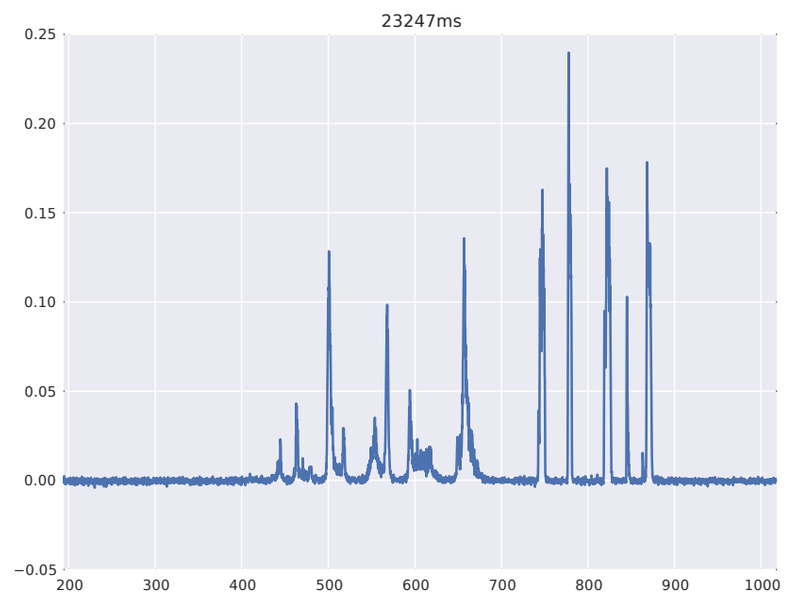

3507ms
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
200
300
400
500
600
700
800
900
1000

샘플 2


0ms
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
200
300
400
500
600
700
800
900
1000


17135ms
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
200
300
400
500
600
700
800
900
1000


22446ms
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
200
300
400
500
600
700
800
900
1000


23247ms
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
200
300
400
500
600
700
800
900
1000

샘플 분류하기

- 스펙트럼의 피크 정보를 확용하는것이 유용:
  - 샘플 데이터를 지도학습에 이용 : ~70% 정확도
  - Peak 데이터를 지도학습에 이용 : 〉90% 정확도
- 관건 : 노이즈가 섞인 스펙트럼에서 피크 뽑아내기

피크 찾기

- 여러 분야에 쓰임 :
  - 분광분석 (빛의 파장)
  - 질량분석(질량/전하 비율)
  - 시계열 데이터
- 눈으로는 찾기 쉬운데 정량적으로 정의하기는 힘듬

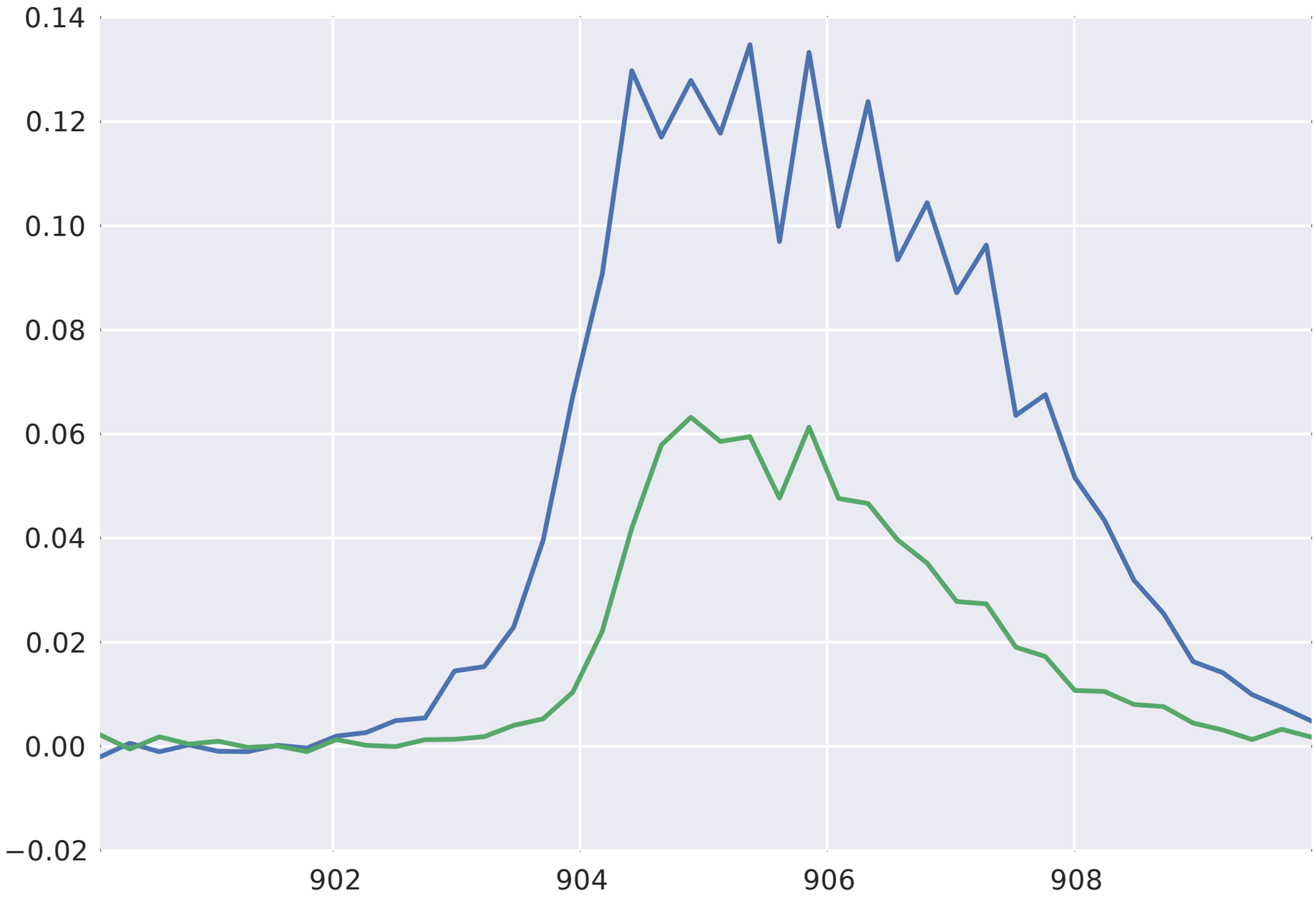
기존의 피크 찾기 방법

- 일반적으로 세 단계로 구성됨:

  1. Smoothing (화이트 노이즈 제거)
  2. Baseline correction (완만한 굴곡 제거)
  3. Peak finding criterion (피크 추출)

- 문제점 : 데이터의 각 요소가 서로 독립이라 가정

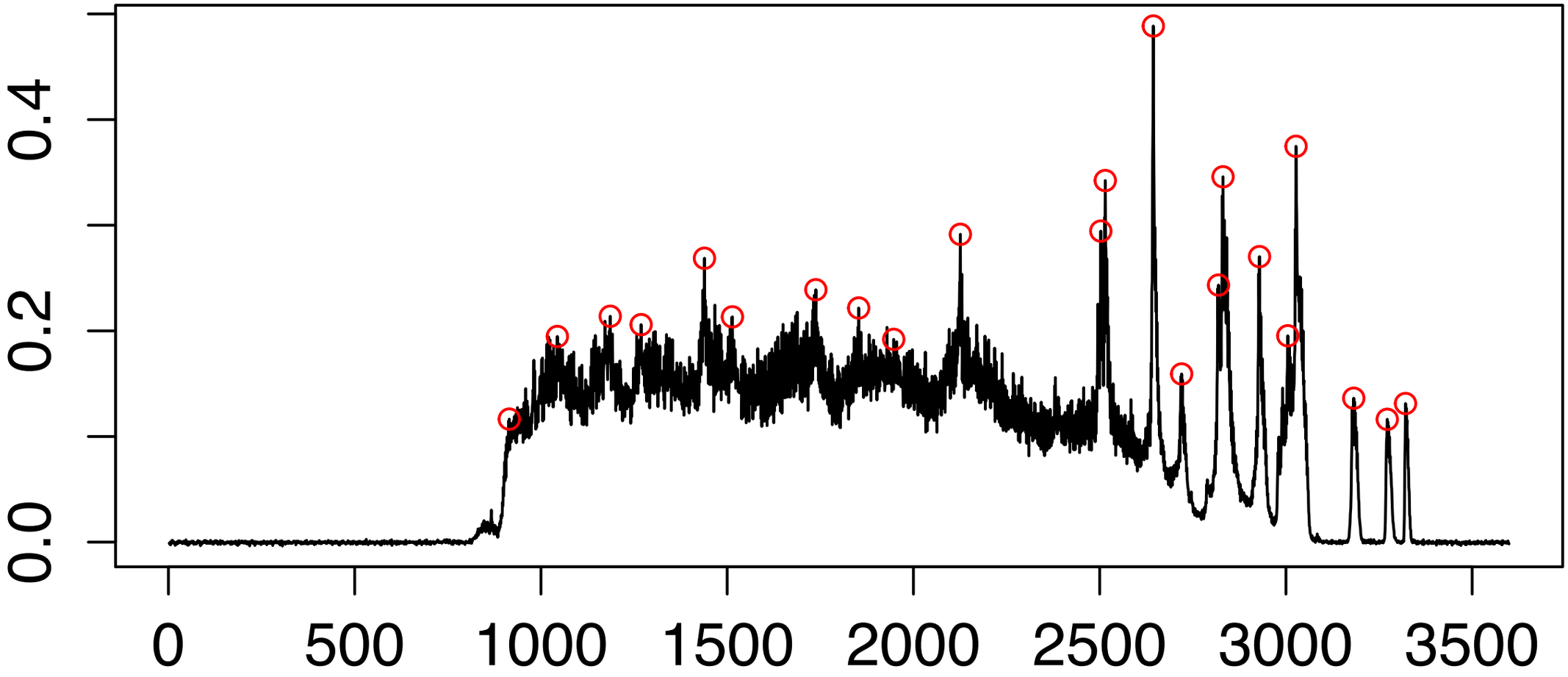
기존 방법의 한계



확률추론에 기반한 피크 찾기

- 아이디어:
  - 샘플간의 상관관계를 확률 모형으로 : P(D|H)
  - 각 샘플의 노이즈제거, 피크 찾기 : P(H|D)
- 개별 데이터만 볼때보다 노이즈 속 피크를 찾기 용이

피크찾기 사례



피크찾기 : 남은 문제

- 데이터에 맞는 확률모형 만들기 : 현재는 손으로..
- N 차원 데이터, N^m 상관관계 :
  - N차원 데이터를 n개의 그룹으로 나눠서 생각하기
  - N차원 몬테카를로 방법에서도 중요한 문제

정리

- 확률 추론 : Prior, D -> Posterior
- 데이터를 기술하는 확률 모형이 중요
- 확률 모형 세우기 : 기계학습도 유용하게 쓰임
- 피크 찾기 : 확률추론으로 더 높은 정확도