

기계학습과 음성인식

류혁수 역

December 2021

Contents

1 이 책의 목적과 사전지식	7
1.1 이 책의 목적	7
1.2 이 책의 구성	7
1.3 이 책에서 사용하고 있는 수식의 표기	7
1.4 확률론의 기초	7
1.4.1 주변화	7
1.4.2 조건부 확률	7
1.4.3 독립성	7
1.4.4 연속분포와 확률밀도함수	7
2 기계학습에서의 예측	9
2.1 모델에 따른 예측	10
2.2 식별함수의 구성	10
2.3 확률모델의 학습	10
2.4 최적화 알고리즘	10
2.4.1 볼록함수의 최적화	10
2.4.2 지수형 분포에서의 최대 우도 추정	10
2.4.3 은닉변수모델과 EM알고리즘	10
2.4.4 경사에 기반한 국소최적화	10
2.5 사례: 신장과 체중에 따른 나이 추정	10
2.5.1 생성모델 접근법	10
2.5.2 식별모델 접근법	10
2.5.3 식별함수법에 따른 접근법	10
2.6 심층학습	10
2.6.1 식별모델의 구성과 소프트맥스층	10
2.6.2 확률 경사 하강법	10
2.7 모델 선택과 과학습	10
2.7.1 과학습	10
2.7.2 교차 검증	10
2.7.3 정착화	10
2.7.4 조기 종료 (Early Stopping)	10

3 유한상태변환기	11
3.1 유한상태기계	12
3.2 문법과 사전의 표현	12
3.2.1 가중치의 도입	12
3.2.2 변환기의 도입	12
3.3 유한상태변환기의 수학적 정의	12
3.3.1 반환	12
3.3.2 상태집합 Q 와 상태전이집합 E	12
3.3.3 초기상태 I 와 종료상태 F	12
3.3.4 전이 경로와 가중치	12
3.3.5 FST의 등이성	12
3.3.6 대수확률반환과 FST의 확률 해석	12
3.3.7 FST의 결합, 클리니 클로저, 합집합	12
3.4 합성	12
3.4.1 합성연산 알고리즘	12
3.4.2 합성연산의 확률 해석	12
3.4.3 알파벳 문자열의 FST 표현과 합성연산	12
3.5 최단경로문제	12
3.6 FST의 최적화	12
3.6.1 트리밍	12
3.6.2 ϵ 소거	12
3.6.3 가중치와 라벨푸싱	12
3.6.4 결정화	12
3.6.5 최소화	12
3.7 대수확률반환의 가중치를 갖는 비순회FST상의 기대치 계산	12
3.7.1 비순회FST의 위상정렬	12
3.7.2 기대치 계산	12
4 음성인식 시스템	13
4.1 음성인식 시스템의 구성	13
4.2 음성의 단위	13
4.2.1 음소를 통한 음성인식의 생성모델	13
4.2.2 발음모델	13
4.3 음성 분석	13
4.3.1 음성신호 모델	13
4.3.2 분산 푸리에 변환과 주파수 해석	13
4.3.3 필터뱅크 처리	13
4.3.4 캡스트럼 추출과 무상관화	13
4.3.5 대수 에너지	13
4.3.6 세그멘트 분석	13
4.4 음성인식 시스템의 평가	13
4.4.1 인식 성능 평가	13
4.4.2 계산 효율 평가	13

5 음향모델	15
5.1 은닉 마르코프모델	15
5.1.1 강우와 물소리 모델	15
5.1.2 여러 개의 HMM 상태를 갖는 모델	15
5.1.3 비의 추정으로부터 음성인식으로	15
5.2 혼합 정규분포와 연속분포 HMM	15
5.3 음소 문맥 의존 모델	15
5.3.1 결정나무에 따른 음소문맥 클러스터링	16
5.3.2 결정나무를 사용한 음향 모델의 FST 표현	16
5.3.3 응집형 클러스터링에 따른 질문의 자동 생성	16
5.4 신경망 음향모델	16
5.4.1 재귀결합 신경망	16
5.4.2 케이트유닛의 장단기 기억	16
5.5 계열식별학습	16
5.5.1 계열식별학습 기준	16
5.5.2 인식가설을 사용한 최적화 알고리즘	16
5.6 음향 모델 적용 기술	16
5.6.1 성도 길이 정규화에 따른 적용	16
5.6.2 화자 코드의 입력에 따른 적용	16
5.6.3 재학습에 따른 적용	16
6 언어모델	17
6.1 언어모델이란	17
6.2 1그램 언어모델과 Bag-of-words	17
6.3 N그램 언어모델	17
6.4 N그램 언어모델의 학습과 평활화	17
6.4.1 N그램 언어모델의 최대우도추정	17
6.4.2 가산 평활화	17
6.4.3 선형보간 평활화	17
6.4.4 Witten-Bell 평활화	17
6.4.5 Good-Turing 추정법	17
6.4.6 Katz 평활화	17
6.4.7 절대할인법	17
6.4.8 Kneser-Ney 평활화	17
6.5 N그램 언어모델의 FST에 따른 표현	17
6.6 최대 엔트로피 모델과 식별 언어모델	18
6.6.1 최대 엔트로피 원리에 기반한 언어모델	18
6.6.2 문장 레벨의 최대 엔트로피 모델	18
6.6.3 음성인식을 위한 식별 언어모델	18
6.7 신경망 언어모델	18
6.7.1 신경망에 따른 후속 단어 예측	18
6.7.2 단어의 분산표현	18
6.7.3 신경망 언어모델에 따른 rescoring	18

7 대어휘 연속 음성인식	19
7.1 FST의 합성과 확률모델	19
7.1.1 디코딩 네트워크의 구성과 탐색오류	20
7.1.2 disambiguation 심볼	20
7.2 대어휘 연속 음성인식의 탐색문제	20
7.3 대규모 FST 합성 기술	20
7.3.1 온 더 플라이 합성	20
7.3.2 디스크 기반 인식 시스템	20
7.4 N-Best 리스트 및 lattice 생성	20
7.4.1 lattice 생성	20
7.4.2 lattice로부터 N-Best 리스트 생성	20
8 심층학습의 발전	21
8.1 여러가지 신경망 요소	21
8.1.1 포화되지 않는 활성화 함수	21
8.1.2 Dropout	21
8.1.3 배치 정규화	21
8.1.4 합성곱/풀링층	21
8.2 신경망 고속화	21
8.2.1 가중치의 양자화	21
8.2.2 특이값 분해에 따른 가중치 행렬 압축	21
8.2.3 knowledge distillation에 따른 모델 변환	21
8.3 End-to-end 음성인식	21
8.3.1 CTC	21
8.3.2 Encoder-Decoder End-to-end 음성인식	21

Chapter 1

이 책의 목적과 사전지식

이 챕터에서는 이 책의 목적과 이 책을 읽어 나감에 있어서 필요한 사전 지식을 설명한다. 이 챕터에서는 우선 1.1에서 이 책의 목적을 기술한다. 그 다음으로 1.2에서는 이 책의 구성을 장마다 설명한다. 1.3에서는 이 책에서 사용하는 수식의 표기법에 대해서 설명하고, 1.4에서는 이 책에서 사용하는 범위에서 확률론의 기초를 설명한다.

1.1 이 책의 목적

이 책에서는 음성인식 시스템을 구성하는 기술에 대해 **기계학습** (Machine Learning)의 관점에서 설명한다. 사람의 음성을 텍스트로 변환하는 음성인식 시스템은 굉장히 오랜 기간 연구되어 왔으며, 그 배경에는 여러가지 요소 기술의 정수가 들어 있다.

1.2 이 책의 구성

1.3 이 책에서 사용하고 있는 수식의 표기

1.4 확률론의 기초

1.4.1 주변화

1.4.2 조건부 확률

1.4.3 독립성

1.4.4 연속분포와 확률밀도함수

Chapter 2

기계학습에서의 예측

2.1 모델에 따른 예측

2.2 식별함수의 구성

2.3 확률모델의 학습

2.4 최적화 알고리즘

2.4.1 볼록함수의 최적화

2.4.2 지수형 분포에서의 최대 우도 추정

2.4.3 은닉변수모델과 EM알고리즘

2.4.4 경사에 기반한 국소최적화

2.5 사례: 신장과 체중에 따른 나이 추정

2.5.1 생성모델 접근법

2.5.2 식별모델 접근법

2.5.3 식별함수법에 따른 접근법

2.6 심층학습

2.6.1 식별모델의 구성과 소프트맥스층

2.6.2 확률 경사 하강법

2.7 모델 선택과 과학습

2.7.1 과학습

2.7.2 교차 검증

2.7.3 정칙화

2.7.4 조기 종료 (Early Stopping)

인용 및 참고문헌

Chapter 3

유한상태변환기

3.1 유한상태기계

3.2 문법과 사전의 표현

3.2.1 가중치의 도입

3.2.2 변환기의 도입

3.3 유한상태변환기의 수학적 정의

3.3.1 반환

3.3.2 상태집합 Q 와 상태전이집합 E

3.3.3 초기상태 I 와 종료상태 F

3.3.4 전이 경로와 가중치

3.3.5 FST의 등이성

3.3.6 대수학률반환과 FST의 확률 해석

3.3.7 FST의 결합, 클리니 클로저, 합집합

3.4 합성

3.4.1 합성연산 알고리즘

3.4.2 합성연산의 확률 해석

3.4.3 알파벳 문자열의 FST 표현과 합성연산

3.5 최단경로문제

3.6 FST의 최적화

3.6.1 트리밍

3.6.2 ϵ 소거

3.6.3 가중치와 라벨푸싱

3.6.4 결정화

Chapter 4

음성인식 시스템

4.1 음성인식 시스템의 구성

4.2 음성의 단위

4.2.1 음소를 통한 음성인식의 생성모델

4.2.2 발음모델

4.3 음성 분석

4.3.1 음성신호 모델

4.3.2 분산 푸리에 변환과 주파수 해석

4.3.3 필터뱅크 처리

4.3.4 캡스트림 추출과 무상관화

4.3.5 대수 에너지

4.3.6 세그멘트 분석

4.4 음성인식 시스템의 평가

4.4.1 인식 성능 평가

4.4.2 계산 효율 평가

인용 및 참고문헌

Chapter 5

음향모델

5.1 은닉 마르코프모델

5.1.1 강우와 물소리 모델

5.1.2 여러 개의 HMM 상태를 갖는 모델

5.1.3 비의 추정으로부터 음성인식으로

5.2 혼합 정규분포와 연속분포 HMM

5.3 음소 문맥 의존 모델

5.3.1 결정나무에 따른 음소문맥 클러스터링

5.3.2 결정나무를 사용한 음향 모델의 FST 표현

5.3.3 응집형 클러스터링에 따른 질문의 자동 생성

5.4 신경망 음향모델

5.4.1 재귀결합 신경망

5.4.2 게이트유닛의 장단기 기억

5.5 계열식별학습

5.5.1 계열식별학습 기준

5.5.2 인식가설을 사용한 최적화 알고리즘

5.6 음향 모델 적응 기술

5.6.1 성도 길이 정규화에 따른 적응

5.6.2 화자 코드의 입력에 따른 적응

5.6.3 재학습에 따른 적응

인용 참고 문헌

Chapter 6

언어모델

6.1 언어모델이란

6.2 1그램 언어모델과 Bag-of-words

6.3 N그램 언어모델

6.4 N그램 언어모델의 학습과 평활화

6.4.1 N그램 언어모델의 최대우도추정

6.4.2 가산 평활화

6.4.3 선형보간 평활화

6.4.4 Witten-Bell 평활화

6.4.5 Good-Turing 추정법

6.4.6 Katz 평활화

6.4.7 절대할인법

6.4.8 Kneser-Ney 평활화

6.5 N그램 언어모델의 FST에 따른 표현

6.6 최대 엔트로피 모델과 식별 언어모델

- 6.6.1 최대 엔트로피 원리에 기반한 언어모델
- 6.6.2 문장 레벨의 최대 엔트로피 모델
- 6.6.3 음성인식을 위한 식별 언어모델

6.7 신경망 언어모델

- 6.7.1 신경망에 따른 후속 단어 예측
- 6.7.2 단어의 분산표현
- 6.7.3 신경망 언어모델에 따른 rescoring

인용 및 참고문헌

Chapter 7

대어晦 연속 음성인식

본 챕터에서는 지금까지의 챕터에서 설명한 각종 기법을 취합하여 대어晦 연속 음성 인식 (large vocabulary continuous speech recognition; LVCSR) 엔진을 구성하는 방법에 대해 설명한다.

7.1 FST의 합성과 확률모델

지금까지 도입했던 음성인식의 요소를 나타내는 FST를 합성하고, 합성한 FST의 최단 경로 문제의 풀이로서 음성인식결과를 얻는 방법을 고찰한다. 음성인식의 통계모델을 표현하는 FST는 일반적으로는 아래의 4 종류가 있다.

- G : 단어열 Acceptor (6.5)
- L:문맥에 의존하지 않는 음소열로부터 단어열 변환 (4.2.2)
- C:문맥에 의존하는 음소열로부터 문맥에 의존하지 않는 음소열로 변환 (5.3)
- H:HMM state 시퀀스로부터 문맥에 의존하는 음소열로 변환 (5.3)

7.1.1 디코딩 네트워크의 구성과 탐색오류

7.1.2 disambiguation 심볼

7.2 대어휘 연속 음성인식의 탐색문제

7.3 대규모 FST 합성 기술

7.3.1 온 더 플라이 합성

7.3.2 디스크 기반 인식 시스템

7.4 N-Best 리스트 및 lattice 생성

7.4.1 lattice 생성

7.4.2 lattice로부터 N-Best 리스트 생성

인용 및 참고문헌

Chapter 8

심층학습의 발전

8.1 여러가지 신경망 요소

8.1.1 포화되지 않는 활성화 함수

8.1.2 Dropout

8.1.3 배치 정규화

8.1.4 합성곱/풀링층

8.2 신경망 고속화

8.2.1 가중치의 양자화

8.2.2 특이값 분해에 따른 가중치 행렬 압축

8.2.3 knowledge distillation에 따른 모델 변환

8.3 End-to-end 음성인식

8.3.1 CTC

8.3.2 Encoder-Decoder End-to-end 음성인식

인용 및 참고문헌