<용어체크>

베이즈 정리

두 확률 변수의 사전 확률과 사후 확률 사이의 관계를 나타내는 정리를 말하며. 베이즈 확률론 해석에 따르면 베이즈 정리는 사전확률로부터 사후확률을 구하는 것이다. 베이즈 정리는 불확실성 하에서 의사결정문제를 수학적으로 다룰 때 중요하게 이용된다. 특히, 정보와 같이 눈에 보이지 않는 무형자산이 지닌 가치를 계산할 때 유용하게 사용된다. 전통적인 확률이 연역적 추론에 기반을 두고 있다면 베이즈 정리는 확률임에도 귀납적, 경험적인 추론을 사용한다.

베이즈 네트워크

변수를 표현하는 노드와 변수들 간의 의존관계를 표현하는 호의 방향성 비순환 그래프이다. 노드 A 에서 노드 B 까지의 호가 있다면 A 는 B 의 부모이며 노드가 값을 포함하고 있다면 증거 노드라고 한다.

나이브 베이즈 분류기

특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류기의 일종으로 통계 및 컴퓨터 과학 문헌에서 , 나이브 베이즈는 단순 베이즈, 독립 베이즈를 포함한 다양한 이름으로 알려져 있다. 나이브 베이즈 분류는 텍스트 분류에 사용됨으로써 문서를 여러 범주 중 하나로 판단하는 문제에 대한 대중적인 방법으로 남아있다.

<학습내용>

베이지안 네트워크

베이지안 분류기

확률 모델을 이용한 베이지안 분류기 생성

<학습목표>

베이지안 네트워크의 개념을 설명할 수 있다.

베이즈 정리를 이용한 베이지안 분류기 개념을 설명할 수 있다.

확률 모델로 구성된 베이지안 분류기 모델을 설명할 수 있다.

Q. 요즘 같이 스팸 메일이 많이 오는 경우 어떻게 스팸 메일과 정상 메일을 쉽게 분류할 수 있을까요?

: 하루에도 수십에서 수백통씩 메일을 주고 받는 경우 스팸 메일과 정상 메일을 분류하기란 쉽지 않습니다. 그래서 포털사이트나 대기업 등에서는 스팸 메일 필터 알고리즘을 이용하는 경우가 있죠. 예전에는 제목 또는 메일을 보낸 송신자로 필터를 적용하였으나 요즘 워낙 교묘하게 운용되다 보니 새로운 방법이 주목받고 있습니다. 바로 나이브 베이즈 분류기인데요, 문서 내에 포함되어 있는 단어들을 분류하여 스팸 메일에 사용되는 단어인지 정상 메일에 사용되는 단어인지를 확률적으로 계산하여 분류하는 알고리즘 입니다. 메일 내에 포함된 단어를 사람들이 일일이 이 단어는 스팸, 저 단어는 정상하고 구분하기에는 비효율적이라서 알고리즘을 이용하여 정상으로 분류한 메일과 스팸으로 분류한 메일을 따로 구분하여 단어들을 뽑아내고 확률을 계산합니다. 그러면 나중에 새로운 메일이 도착했을 때 메일 내에 포함된 단어가 어느 쪽에 속할 확률이 더 높은가를 자동으로 산출하여 스팸 메일을 분류할 수 있습니다. 편리하면서도 신기하죠?

베이즈 네트워크

- 확률 개념을 도입해 추론 규칙을 개선한 전문가 시스템으로 제안한 것이다.
- 불확실성을 포함한 사건의 예측과 관측 결과를 활용한다.
- 장애 진단에 사용하는 그래픽 확률 모델이다.
- 확률 변수 사이의 확률 의존 관계 정보를 유향 그래프로 나타내는 네트워크로 시스템으로 구성한다.

나이브 베이즈 분류기

- 기계학습에서 데이터의 속성들 사이의 독립 사건을 가정하는 베이즈 정리를 적용한 확률 분류기의 일종이다.
- 지도 학습에 효율적이다.
- 파라미터 추정에 사용되는 학습 데이터 양이 많이 필요하지 않다.
- 간단한 디자인과 단순한 가정 이 실제 응용에 효율적으로 동작한다.

베이즈 분류기 모델

- 스팸 메일이나 문서 분류기 등의 학습 데이터를 활용한 모델에 적합하다.
- 학습 데이터의 사전 확률이 나오지 않는 경우 사후 확률이 부정확할 수 있으므로 바이어스 등을 활용한다.
- 입력 데이터가 실수인 경우 데이터 범위를 이용한 카테고리화로 사전 확률을 적용한다.