

충북대학교
산업인공지능학과

분류 및 회귀 모델

2023254010

사수진

1. 현장에서 분류를 적용할 학습 데이터 수집

분류



제목	기록물철
[외부강의]제목*****	감사관리
[일상감사요청]제목*****	감사관리
[일상감사의견]제목*****	감사관리
[지급신청]제목*****	감사교육일반관리
[지출품의]제목*****	감사교육일반관리
[당직근무]제목*****	감사일반관리
[외부강의]제목*****	감사일반관리
[지급신청]제목*****	감사일반관리
[지급신청]제목*****	개인정보보호관리
[계약의뢰]제목*****	건축토목분야설계및공사감독관리
[계약의뢰검수]제목*****	건축토목분야설계및공사감독관리
[계약체결]제목*****	건축토목분야설계및공사감독관리
[계약체결변경]제목*****	건축토목분야설계및공사감독관리
[일상감사요청]제목*****	건축토목분야설계및공사감독관리
[일상감사의견]제목*****	건축토목분야설계및공사감독관리
[지급신청(일반계약)]제목*****	건축토목분야설계및공사감독관리
[지급신청(직접구매)]제목*****	건축토목분야설계및공사감독관리
[지급신청]제목*****	건축토목분야설계및공사감독관리
[직접구매검수]제목*****	건축토목분야설계및공사감독관리
[직접구매신청]제목*****	건축토목분야설계및공사감독관리
[계약의뢰]제목*****	경영평가관리
[계약의뢰검수]제목*****	경영평가관리
[계약체결]제목*****	경영평가관리
[일상감사요청]제목*****	경영평가관리
[지급신청(일반계약)]제목*****	경영평가관리
[지급신청(직접구매)]제목*****	경영평가관리
[지급신청]제목*****	경영평가관리
[지출품의]제목*****	경영평가관리
[직접구매검수]제목*****	경영평가관리
[직접구매신청]제목*****	경영평가관리

1.데이터 수집: 전자문서 제목에 대괄호 내에 있는 정보와 기록물철 정보가 포함된 데이터를 수집해야 합니다.

2.데이터 전처리: 수집한 데이터를 전처리하여 학습에 사용할 수 있는 형태로 가공해야 합니다. 대괄호 내에 있는 정보를 추출하고, 기록물철 데이터와 연결하여 하나의 훈련 데이터로 구성합니다. 이때, 대괄호 내의 정보와 기록물철 데이터를 적절한 방식으로 표현하고 수치화하는 것이 중요합니다.

3.특징 추출: 대괄호 내에 있는 정보와 기록물철 데이터를 기반으로 특징을 추출합니다. 이는 해당 정보의 속성을 정의하고, 각 문서의 특징을 수치적으로 표현하는 과정입니다.

4.분류 모델 학습: 전처리된 데이터와 특징을 사용하여 분류 모델을 학습시킵니다. 대표적인 분류 모델로는 로지스틱 회귀, 의사결정 트리, 랜덤 포레스트, 서포트 벡터 머신 등이 있습니다. 선택한 모델에 데이터를 입력하고, 모델을 학습시킵니다.

5.모델 평가: 학습된 분류 모델을 평가하여 예측 성능을 측정합니다. 이를 위해 테스트 데이터셋을 사용하여 모델의 예측 정확도, 정밀도, 재현율, F1 스코어 등을 평가합니다. 필요에 따라 모델의 하이퍼파라미터를 조정하고, 교차 검증 등의 기법을 활용하여 모델을 향상시킬 수 있습니다.

2. 현장에서 회귀를 적용할 학습 데이터 수집



회귀

년도	제목	건수
2022	[계약의뢰]제목*****	629
2023	[계약의뢰]제목*****	276
2022	[계약의뢰검수]제목*****	613
2023	[계약의뢰검수]제목*****	162
2022	[계약의뢰변경]제목*****	37
2023	[계약의뢰변경]제목*****	15
2022	[계약체결]제목*****	608
2023	[계약체결]제목*****	235
2022	[계약체결변경]제목*****	50
2023	[계약체결변경]제목*****	16
2022	[교육수입결의]제목*****	18
2023	[교육수입결의]제목*****	6
2022	[급여결의]제목*****	13
2023	[급여결의]제목*****	4
2022	[당직근무]제목*****	227
2023	[당직근무]제목*****	83
2022	[대체결의]제목*****	784
2023	[대체결의]제목*****	234
2022	[방문자숙소수입결의]제목*****	20
2023	[방문자숙소수입결의]제목*****	6
2022	[부패영향평가결과]제목*****	8
2023	[부패영향평가결과]제목*****	10
2022	[부패영향평가요청]제목*****	9
2023	[부패영향평가요청]제목*****	10
2022	[수입결의]제목*****	1085
2023	[수입결의]제목*****	413
2022	[예산경정]제목*****	13
2023	[예산경정]제목*****	3
2022	[예산변경]제목*****	43
2023	[예산변경]제목*****	16

1.데이터 수집: 전자 문서 제목을 수집하고, 대괄호 내에 있는 정보와 해당 년도를 추출합니다. 예를 들어, "[계약의뢰]제목****" 라는 제목에서 대괄호 내의 정보 " 계약의뢰 " 와 해당 년도를 추출합니다.

2.데이터 전처리: 추출한 대괄호 내의 정보와 해당 년도를 사용하여 데이터 포인트를 생성합니다. 예를 들어, " 계약의뢰 " 를 대괄호 정보로 사용하고, 해당 년도를 회귀 분석의 종속 변수로 사용합니다. 독립 변수로는 대괄호 내의 정보를 사용합니다.

3.데이터 집계: 수집한 데이터를 기반으로 대괄호 내 정보에 따른 각 년도의 건수를 집계합니다. 이를 통해 독립 변수와 종속 변수 간의 관계를 분석할 수 있습니다.

4.데이터 분할: 수집한 데이터를 학습 세트와 테스트 세트로 분할합니다. 학습 세트를 사용하여 회귀 모델을 학습하고, 테스트 세트를 사용하여 모델의 성능을 평가합니다.

5.회귀 분석: 분할된 데이터 세트를 사용하여 회귀 모델을 학습합니다. 대괄호 내의 정보를 독립 변수로, 해당 년도의 건수를 종속 변수로 사용하여 회귀 분석을 수행합니다.

3. 어떠한 학습을 수행할지?

학습 방법



전자문서 제목에 대괄호 내에 있는 정보를 기반으로 년도별 건수를 예측하기 위해 사용할 수 있는 몇 가지 학습 방법은 다음과 같습니다.

1. 선형 회귀: 대괄호 내의 정보를 독립 변수로 사용하여 선형 회귀 모델을 학습할 수 있습니다. 선형 회귀는 종속 변수와 독립 변수 사이의 선형적인 관계를 모델링하는 방법입니다. 대괄호 내의 정보를 수치형으로 변환하고, 해당 건수를 종속 변수로 사용하여 선형 회귀 모델을 학습시킬 수 있습니다.

2. 의사결정 트리: 대괄호 내의 정보를 기반으로 한 의사결정 트리 모델을 학습할 수 있습니다. 의사결정 트리는 특징들의 조건에 따라 분기하는 계층적인 결정 규칙을 생성하는 분류 및 회귀 분석 알고리즘입니다. 대괄호 내의 정보를 기반으로 건수를 예측하는 의사결정 트리 모델을 학습시킬 수 있습니다.

3. 랜덤 포레스트: 대괄호 내의 정보를 기반으로 한 랜덤 포레스트 모델을 학습할 수도 있습니다. 랜덤 포레스트는 의사결정 트리의 앙상블 기법으로, 여러 개의 의사결정 트리를 사용하여 예측을 수행합니다. 대괄호 내의 정보를 특징으로 사용하여 랜덤 포레스트 모델을 학습시킬 수 있습니다.

4. 신경망: 대괄호 내의 정보를 입력으로 사용하여 신경망 모델을 학습할 수 있습니다. 신경망은 복잡한 비선형 관계를 모델링할 수 있는 강력한 기법입니다. 대괄호 내의 정보를 수치형으로 변환하고, 해당 건수를 예측하는 신경망 모델을 학습시킬 수 있습니다.

위의 방법들은 대괄호 내의 정보를 활용하여 건수를 예측하기 위해 일반적으로 사용되는 학습 방법입니다. 선택한 방법은 데이터의 특성과 예측 목표에 따라 달라질 수 있으므로, 데이터를 탐색하고 실험을 통해 가장 적합한 학습 방법을 선택하는 것이 중요합니다.