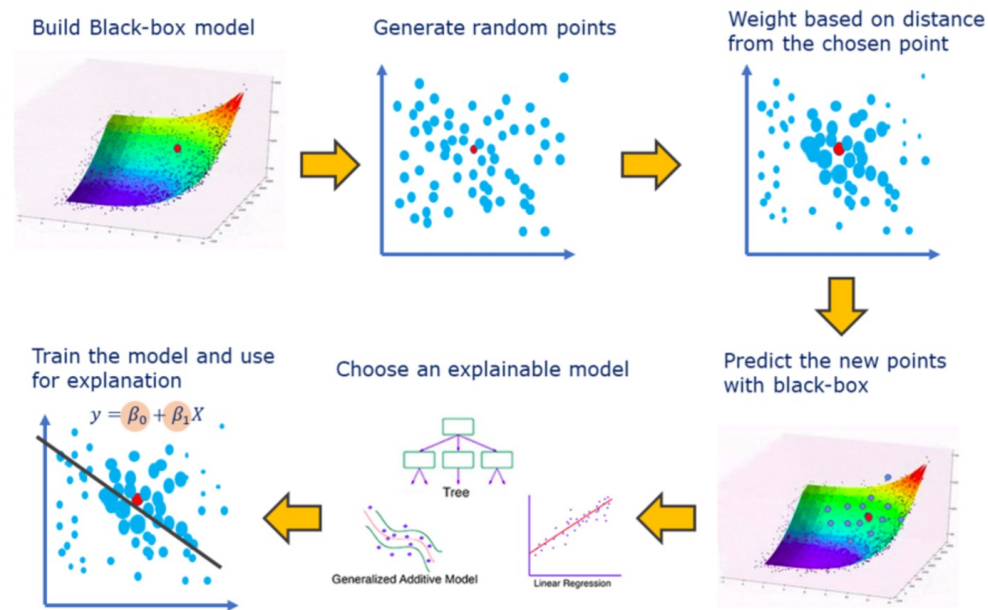


• Core Idea

- 입력값을 조금 바꿨을 때 모델의 예측값이 크게 바뀌면, 그 변수는 중요한 변수이다

• Algorithm 간단히 이해해보기

1. 블랙박스 모델의 예측에 대해 설명이 필요한 관측치 선택
2. 데이터셋을 변형하고 이 새로운 샘플들에 대한 블랙박스 모델 예측 얻기
3. 관심있는 관측치와의 근접도에 따라 새로운 샘플에 가중치 부여
4. 가중치 적용된 해석 가능한 모델을 변형된 데이터셋에 학습시킨다



LIME with Text Data

데이터 변형은 데이터 타입에 따라 다릅니다

- Text

- 원래 텍스트에서 임의로 단어를 제거하여 만들어집니다
- 예) Youtube 댓글에 대해 스팸인지 아닌지 분류해보자

INDEX	CONTENT	CLASS
173	For Christmas Song visit my channel :)	1

- 데이터 변형

- **prob**: 변형된 문장이 스팸인지 아닌지 예측값
- **proximity**: 원본 문장과 근접도

For	Christmas	Song	visit	my	channel	:)	prob	proximity
1	0	1	1	0	0	1	0.17	0.57
0	1	1	1	1	0	1	0.17	0.71
1	0	0	1	1	1	1	0.99	0.71

...

LIME with Image Data

- 하나 이상의 픽셀이 하나의 클래스에 기여
- 이미지를 "슈퍼픽셀"로 분할 및 마스킹 과정을 통해 이미지 변형
 - 유사한 색상의 픽셀을 연결한 것

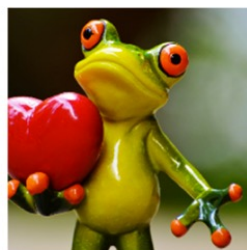


Original Image



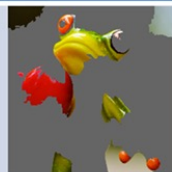
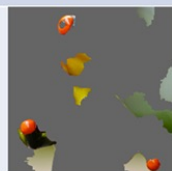
Interpretable Components

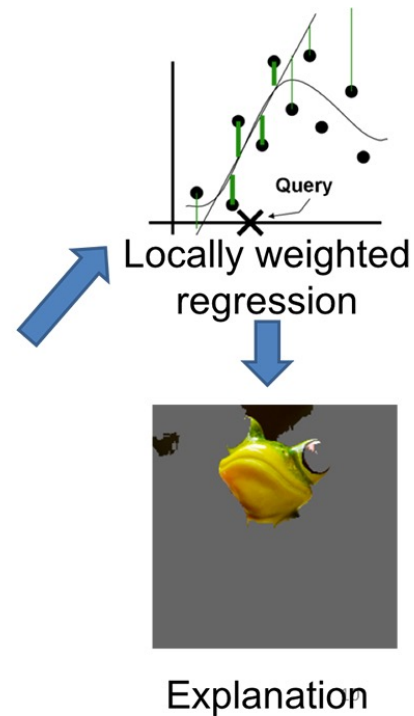
- 슈퍼픽셀 몇 개를 짚어서 회색으로 가린다
- 모델에 넣고 예측값 구하기
 - 예측값이 많이 변하면?
 - 반대로 많이 변하지 않았다면?



Original Image
 $P(\text{tree frog}) = 0.54$

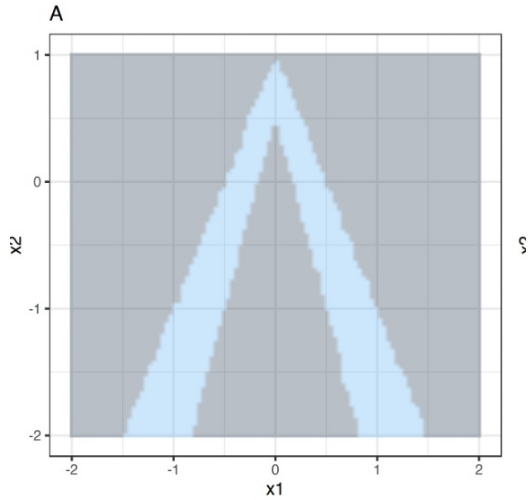


Perturbed Instances	$P(\text{tree frog})$
	<div><div></div></div> 0.85
	<div><div></div></div> 0.00001
	<div><div></div></div> 0.52



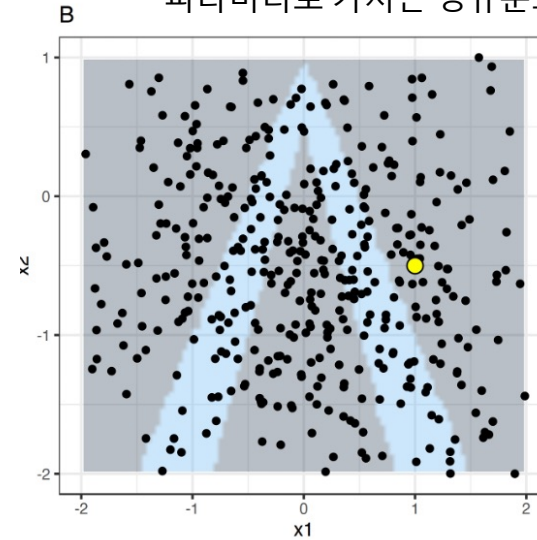
LIME with Tabular Data

A: 블랙 박스 모델은 변수 x_1 와 x_2 가 주어졌을 때 두 클래스 중 하나를 예측

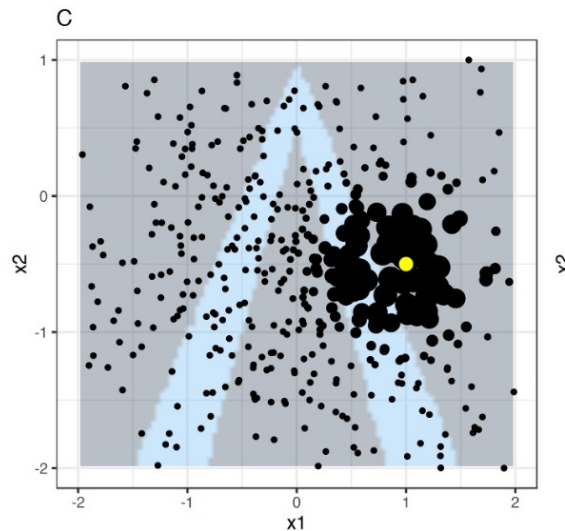


B: 노란점: 설명하고자 하는 관심있는 관측치

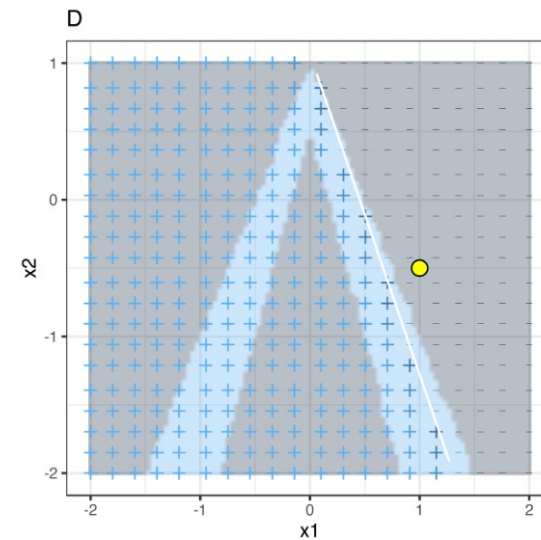
검은점: 학습 데이터에서 해당 변수의 평균과 분산을
파라미터로 가지는 정규분포에서 추출된 데이터



C: 관심 있는 인스턴스 근처의 점들에 높은 가중치 부여



D: 가중된 샘플 형태를 학습한 모델의 분류 결과 표시



- 빠르게 예제 하나로 **쉬운 시각화** 가능
- 보여주고 싶은 대표 케이스를 설명하기엔 좋음
- 어떤 이벤트가 일어난 특정 시점, 특정 데이터 포인트를 안다면 활용하기 좋을 것 같음
- Jupyter notebook 호환성
- 한 지점밖에 못 보기 때문에, 구간을 볼 때 취약
- Lime model 의 Feature 계수들을 활용성?
 - Util 함수 지원 X
 - dataframe 화 제공 X -- Single instance 에 초점을 맞추어 전반적으로 계수를 보게 배려하지 않고 있음
 - Tabular data 의 경우 feature 계수가 probability 로 나오기 때문에, feature 계수를 추출하려면,
 - List 로 single instance 에 대한 feature 값들을 받아서 전처리 필요
 - 처리된 features 들을 data frame 화 해야하는데, 만약 feature 갯수가 많다면, 전처리 하는 과정에서 놓치거나 오류가 나올 수 있는 가능성이 많아 굉장히 꼼꼼히 보거나 features 에 대해 빠삭히 아는 상태에서 처리해야할 것 같음