**LIME: “Why Should I Trust You?” Explaining the Predictions of Any Classifier**

00. Summary of LIME

|  |
| --- |
| **Interpretable explanation:** 사람이 이해할 수 있는 형태의 표현을 사용한 설명   * Text classification의 경우 **각 word의 존재 여부를 표시한 binary vector**가 될 수 있다. * Image classification의 경우 **contiguous patch 또는 super-pixel이 존재하는지의 여부를 표시한 binary vector**가 될 수 있다.   **:** 설명되는 instance에 대한 **original representation**  **:** 그 original representation의 **interpretable** representation에 대한 **binary vector**  **:**   * **g:** **explanation을 모델로 정의**한 것 (domain: , 따라서 g는 **각 interpretable component의 존재 여부**를 나타낸다고 할 수 있음) * **G:** 잠재적으로 interpretable한 **모델의 class** (linear models, decision trees, …)   **:** explanation g에 대한 **measure of complexity**  **:** 설명 가능한 모델을 이렇게 나타낼 수 있다.   * **:** x가 **특정 class에 속할 확률** (또는 binary indicator) * **:** 어떤 instance z와 x 사이의 **proximity measure** (x에 대하여 local하게 정의)   **:** 모델 g가 f를 추정할 때 **얼마나 unfaithful한지, 즉 loss가 큰지**를 나타낸다.   * Interpretability와 local fidelity(x에 대하여 local하게 잘 추정함)를 위하여 **를 충분히 낮게 하면서 를 최소화**해야 한다. * **LIME에 의하여 만들어지는 explanation**은 다음 수식과 같다. |

01. Local Explanation을 위한 샘플링

|  |
| --- |
| Locality-aware loss **를 설명 가능한 모델 f에 대한 어떤 가정 없이 최소화**해야 하므로 **explainer는 model-agnostic**해야 한다.   * 즉, f의 local behavior를 학습하기 위하여 **샘플을 이용하여 를 추정**해야 한다. * 0이 아닌 의 element를 이용하여 의 주변에 있는 instance를 샘플링한다.   Perturbed sample 에서 **original representation에 있는 샘플 를 복원하고 를 찾는다.**   * Associated label이 있는 Perturbed sample의 dataset Z에 대하여 **(1)을 최적화하여 설명 를 얻을 수** 있다. |

02. Sparse Linear Explanations

|  |
| --- |
| 여기서 **G는 인 linear model의 class**라고 하자. 여기서 locally weighted square loss 를 다음과 같다고 하자.  Text classification에서 interpretable representation을 bag of word라고 하고 **number of word의 값을 최대 K**로 제한하여 **explanation이 interpretable하다는 것을 보장**한다.   * 여기서는 K의 값을 constant value로 이용한다. * Image classification에서는 word 대신 super-pixel을 이용한다.     여기서는 다음과 같이 한다.  1. **K개의 feature를 처음에 선택**하는데, 이때 **Lasso 알고리즘**을 이용한다.  2. **Least-square를 이용하여 weight을 학습**한다. (K-LASSO in Algorithm 1)   * 이 알고리즘이 individual prediction에 대한 설명을 생성하고, 이것의 complexity는 dataset의 크기가 아닌 **f(x)를 계산하는 데 걸리는 시간과 샘플의 개수 N**에 의해 결정된다. |

03. Examples

|  |
| --- |
| **1. Text Classification with SVMs**  오른쪽 사진은 **SVM with RBF kernel을 Christianity와 Atheism을 구분하도록 학습**시킨 모델을 이용한 예측을 설명한 것이다.   * 94% held-out accuracy * 특정 instance에 대한 설명은 **prediction이 해당 분야와 무관한 임의의 이유(Posting, Host, Re 등)에 의해 만들어졌다**는 것을 알 수 있다.   **2. Deep networks for images**  Sparse linear explanation을 image classifier에 이용할 때 **특정 class에 대한 positive weight이 있는 super-pixel을 강조**하고 싶을 수 있다.   * 여기서는 **Google의 pre-trained Inception neural network**의 예측을 설명한다. * K=10 |

04. Submodular pick for explaining models

|  |  |  |
| --- | --- | --- |
| 여기서는 **개별 instance의 집합을 이용하여 모델의 global understanding을 도출**하려고 한다.   * Instance의 집합 X에서 B개의 instance를 선택하여 사용자에게 설명하는 문제를 **pick step**이라고 한다.   1. instance의 집합 X가 있다고 할 때 **각 instance의 interpretable component의 local importance**를 나타내는, 차원이 인 **explanation matrix W**를 만들 수 있다.   * Instance 와 explanation 에 대하여 이다.   2. 각 component에 대한 **global importance**를 다음과 같이 정의한다.   * Image에 대해서 I는 **서로 다른 이미지에서 비교할 수 있는 superpixel**에 대한 feature를 측정해야 한다.   **<Algorithm>**   |  | | --- | | 1. Algorithm 1을 이용하여 **각 instance의** **weight w**를 구하고 그것을 **인** 라고 하자.  2. 각 element (feature) j에 대하여 다음을 구한다.  3. V를 정의하고, **B개의 instance를 다음과 같이 선택**하여 다음 수식의 값을 greedy하게 최적화한 후 V를 반환한다.  이때 **marginal converge gain** of adding instance i는 다음과 같다. |   **<Example>**   |  | | --- | | 여기서는 W에 대하여 **인 binary**이다.  여기서 importance function I는 f2를 f1보다 높게 점수를 매긴다. 즉 인데 이는 **f2가 더 많은 instance를 설명하기 때문**이다.  이때 **row 2**가 선택되었다면 다음에 **인 row 5**가 선택된다. 즉 **explanation은 중복되면 안 된다.** | |