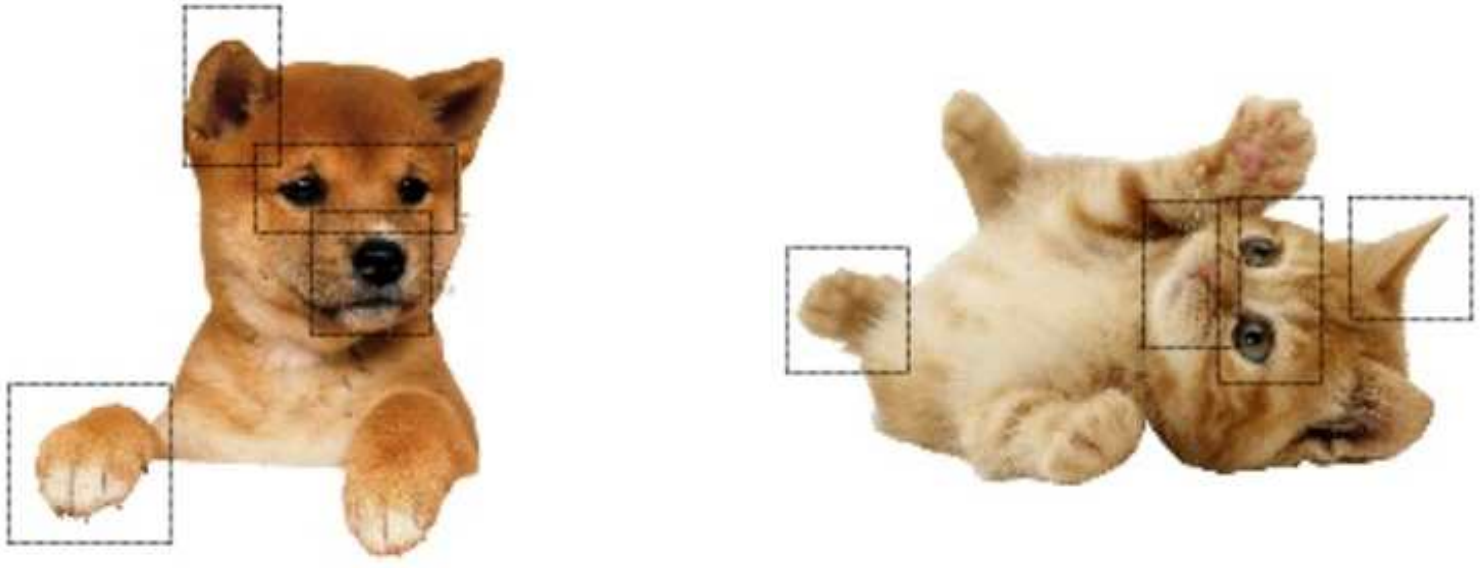
Feature for NLP

강사 : 백병인

pi.paek@modulabs.co.kr

모두의연구소 Research Scientist

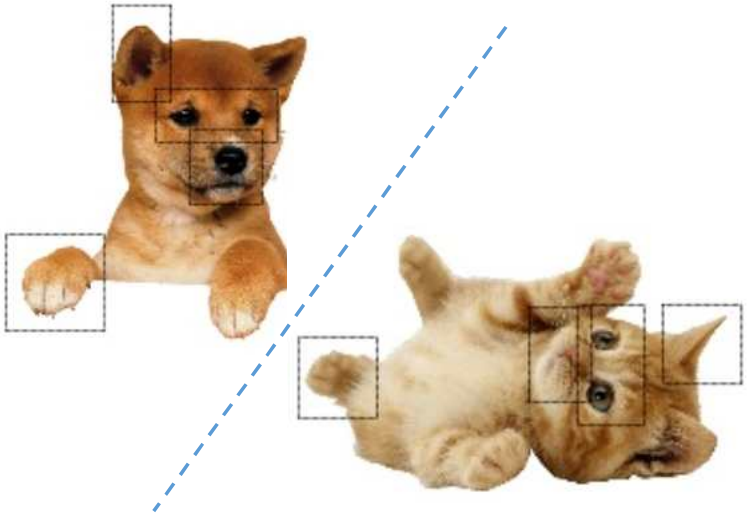
무엇을 보고 다르다는 것을 알 수 있을까?



개와 고양이가 **다르다는 것을 알게 해 주는 중요한 특징**을 **Feature**라고 한다.

https://www.slideshare.net/NaverEngineering/ai-83896428 (그림그리는 AI, 이활석)

Feature는 고정된 것이 아니다.



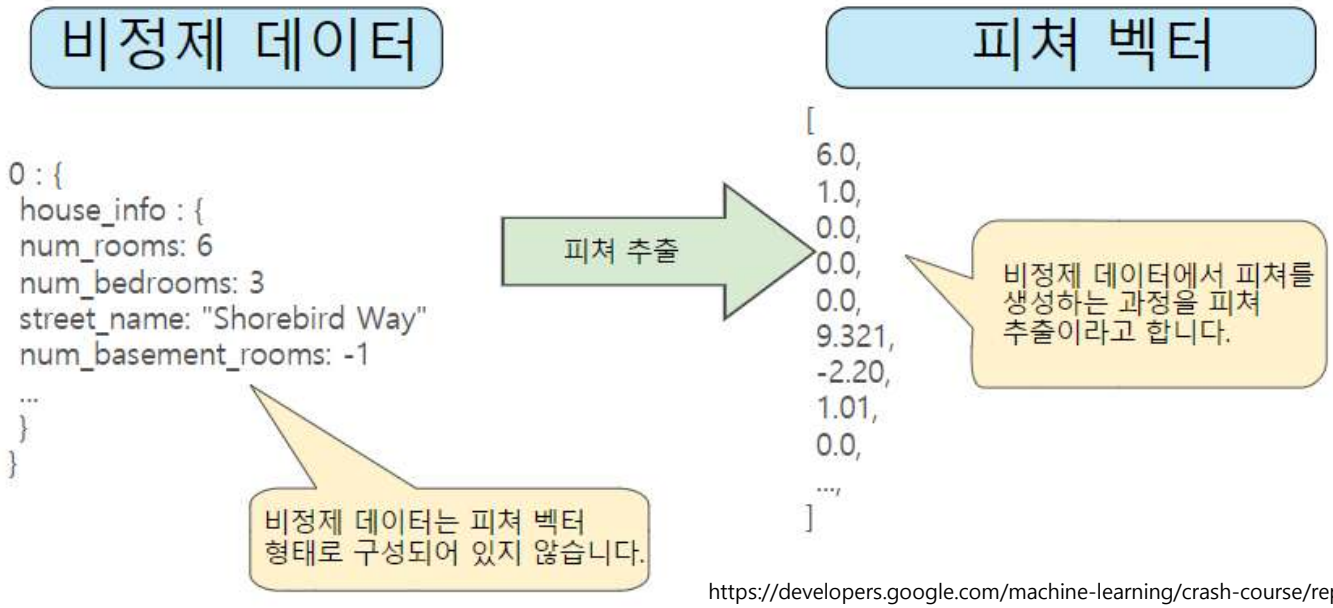
개와 고양이를 구분할 때 중요한 Feature와, 고양이들끼리의 종류를 구분할 때 중요한 Feature는 다르다.
 * 고양이 종류를 구분할 때, 고양이 눈깔 모양은 그리 크게 도움이 안된다.
 * 고양이 종류를 구분하기 위해서는 훨씬 더 섬세하고 복합적인 Feature가 필요하다.

Feature는 구분하려는 레이블들 사이의 경계선이 어디에 그어지느냐에 따라 규정된다.

Feature란?

- Individual measurable property or characteristic of a phenomenon being observed. (from Wikipedia)
  - **Individual** : 독립적이어야 한다.
  - **Measurable** : 숫자로 표현 가능해야 한다.
    (엄밀하게는 크기를 비교할 수 있는 측정 기준이 있다는 뜻이다.)
  - **Observable** : 관측 가능해야 한다.
  - **Characteristic** : 어떤 것 X가 다른 것 Y와 구별되게 하는 특징을 규정해야 한다.
  - **Property** : X의 하위 요소이다.

Feature Representation (Vector)



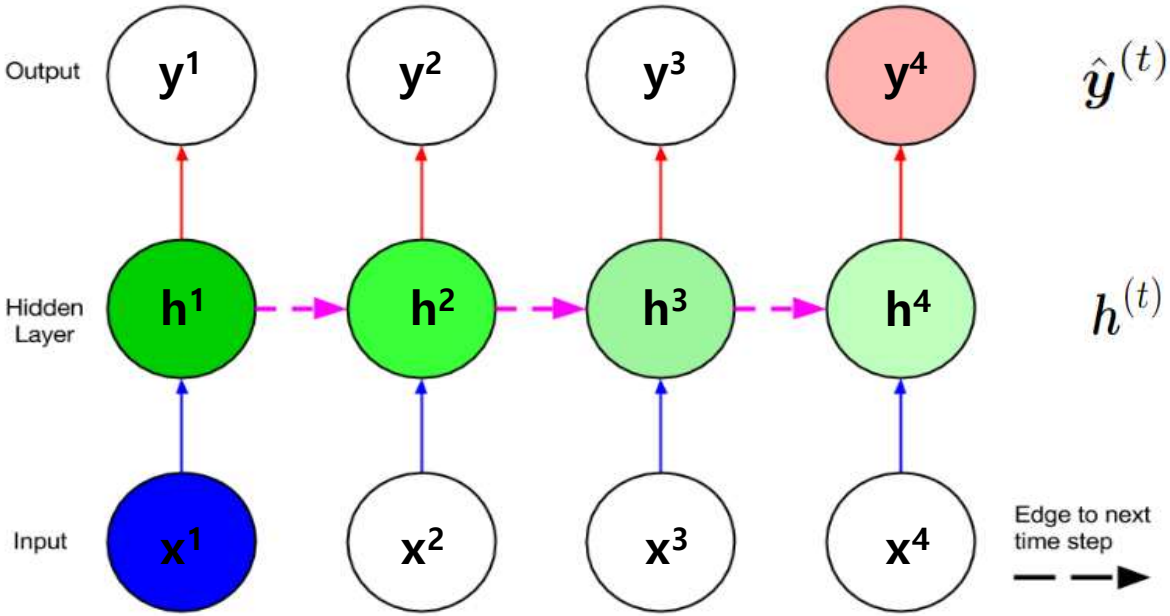
https://developers.google.com/machine-learning/crash-course/representation/feature-engineering?hl=ko

머신러닝의 방법으로 데이터를 레이블로 분류하기 위해서 Feature를 Vector의 형태로 표현한다.
데이터의 하위 요소가 N개가 명확히 정의되어 있다면, 데이터를 N차원의 Vector로 표현 가능하다.

그런데, 데이터의 하위요소가 명확하게 규정되어 있지 않다면? (그냥 데이터와 레이블만 주어진다면?)

State Revisited

- RNN은 state를 feature로 사용하는 구조의 Neural Network



Feature for RNN Inference

$$\hat{\boldsymbol{y}}^{(t)} = \mathrm{softmax}(W^{\mathrm{yh}}\boldsymbol{h}^{(t)} + \boldsymbol{b}_y)$$

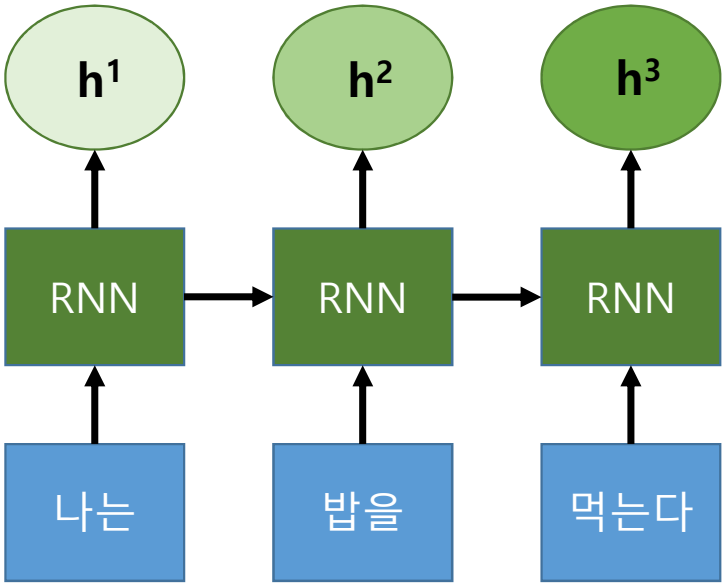
RNN의 출력(점원의 대답) y^t은 새로운 상태벡터 h^t에 가장 적절한 것으로 선택된다.

$$\boldsymbol{h}^{(t)} = \sigma(W^{\mathrm{hx}}\boldsymbol{x}^{(t)} + W^{\mathrm{hh}}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_h)$$

새로운 상태벡터 h^t는 이전 상태벡터 h^{t-1}과 새로운 입력벡터 x^t의 벡터 선형결합(linear combination)으로 결정된다.

문장 안에서 state의 변화가 있을까?

- 과연 문장을 sequence로 modeling하는 게 타당할까?



state의 변화로 모델링?

'나'와 '밥'과 '먹는다'는 시차를 두고 순서대로 발생하는 별개의 사건이 아니라 하나의 사건 속에 **동시에 존재하는 구성 요소** 아닌가요?

RNN의 약점

- 다른 Neural Network보다 학습속도가 느리다. 시간순차적 학습 때문에 병렬처리에 약점을 지니기 때문

- RNN의 state에는 sequence의 가장 마지막 input이 가장 많이 반영되어 있을 가능성이 높다. 그래서 sentence classification에는 정확도가 높지 않다.

- 그렇다면 어떤 대안을 생각해 볼 수 있을까?
  - 혹시 CNN은 어떨까?