Modeling

KU-BIG 학술부

1. Modeling 이란?

데이터의 패턴을 구하기 위한 설정

KU-BIG

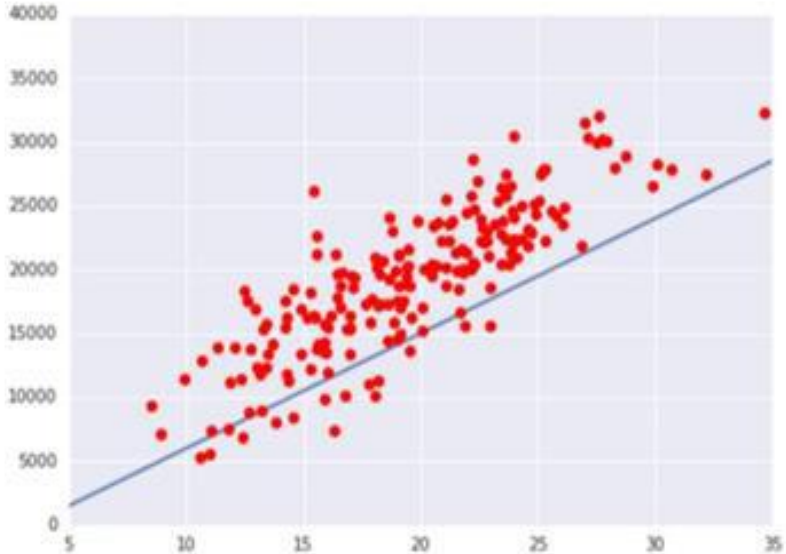
2. Model의 종류

1. Linear Base Model
2. Classification
3. Ensemble
4. Clustering
5. Neural Network

2-1-1. Linear Base Model

$$y_i = \alpha + \beta x_i + \varepsilon_i .$$

데이터의 추세를 1차식으로 나타내는 방법.



가장 간단하지만 한계가 많다.

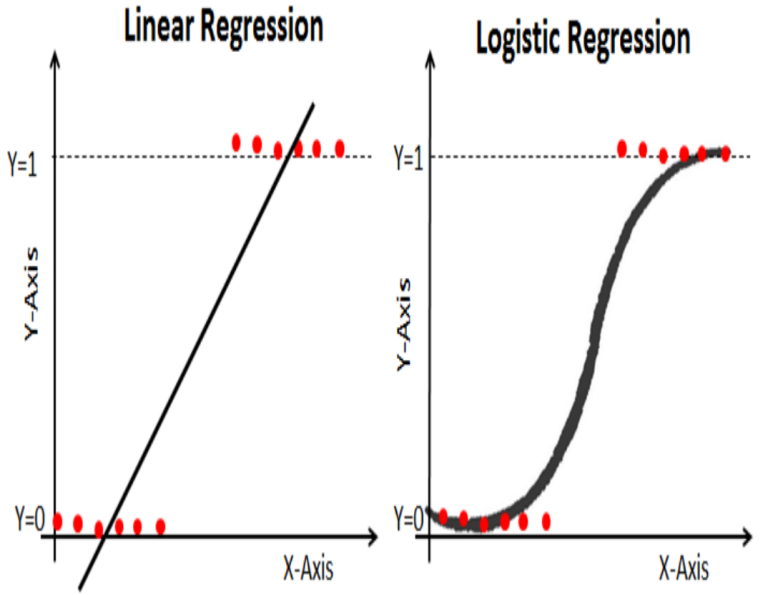
- **오차항이 정규분포를 따른다는 가정**
- **예측 변수 x 끼리 선형적으로 독립**

KU-BIG

2-1-2. Generalized Linear Base Model

종속변수가 정규분포이지 않은 경우의 Linear Model

- Linear Base Model 에서, 종속변수인 y가 연속적이지 않다면?(범주형, 이항반응 변수 등)
- 생존분석 등을 선형으로 나누기는 어렵다.
- Logistic Regression에서는, 일반적으로 0.5 이상을 1, 그 미만을 0으로 나누어 출력



https://nittaku.tistory.com/478

2-1-3. Penalized Linear Model

Penalty가 부여된 Linear Model → **Overfitting을 방지**

Bias는 조금 생기더라도, Variance를 획기적으로 줄이는 방법

- **Ridge : 가중치(계수)들의 값을 감소시킴**
- **LASSO : 변수선택 기능이 있음(중요한 변수만 선택하고, 다른 변수 계수는 0 으로).**
- **Elastic Net : Ridge와 LASSO의 혼합(큰 데이터셋에서 유리)**

KU-BIG

2-2. Classification

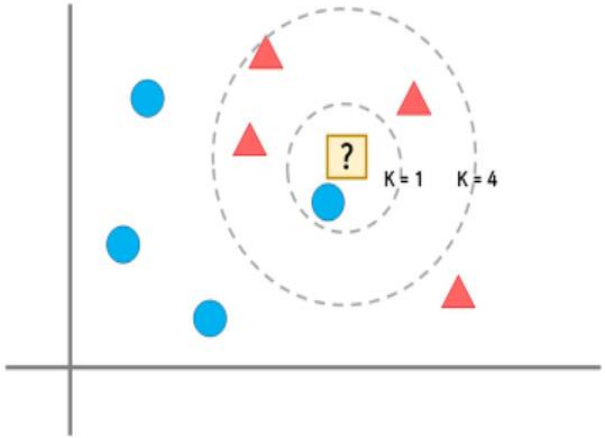
종속변수 Y가 범주형 데이터일 때, 이를 분류하는 방법.

- K-NN
- Decision Tree
- SVM(Support Vector Machine)

2-2-1. K-NN (K- Nearest Neighbor)

가장 고전적이며, 직관적인 Classification

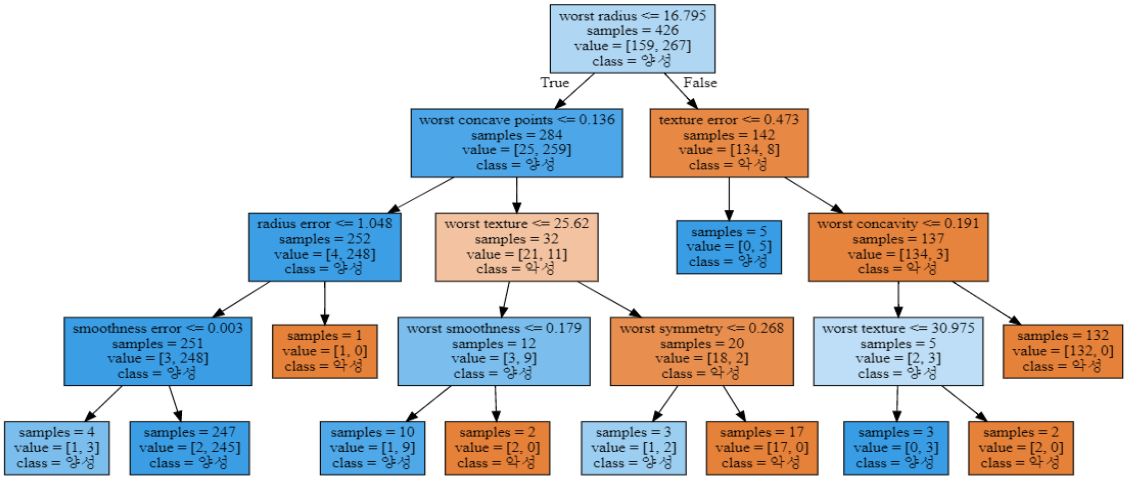
- K-NN은 어떤 데이터가 입력되었을 때,
  주변에 어떤 것이 많은가를 기준으로 분류
- 그러나, 옆의 그림처럼, 애매한 경우가 있다.
  *** 주변에서 몇 개의 샘플을 관찰하고 분류할 것인지가 중요**
- K는 관찰할 주변 데이터의 숫자이다.



https://gomguard.tistory.com/51

2-2-2. Decision Tree

질문에 질문을 이어 데이터를 분류(또는 회귀)해 나가는 기법.



2-2-3. SVM (Support Vector Machine)

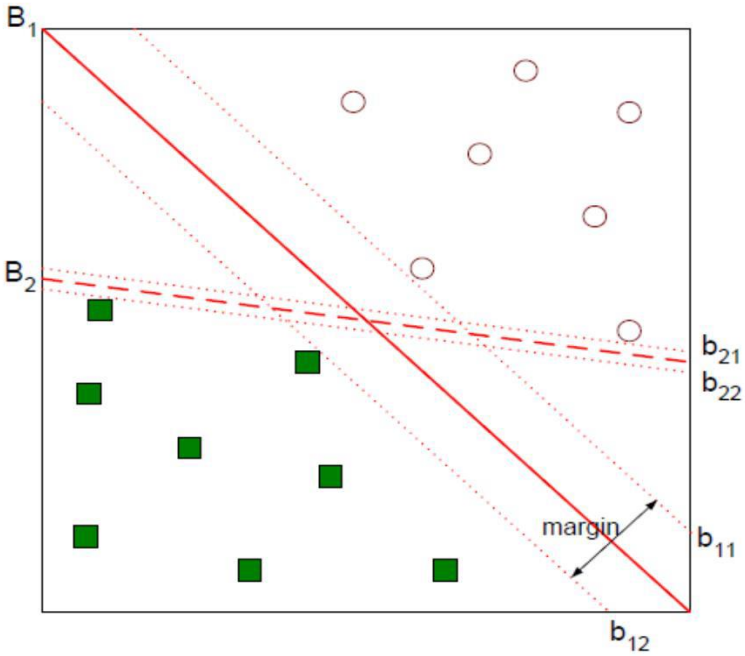
- Linear SVM

오른쪽 그림에서, 더 나은 분류경계면은 B1

그림과 같이, **Margin을 최대화**하는
선형 경계면을 찾는 기법이 Linear SVM

* Margin : 분류경계면에서 가장 가까운 점과의 거리

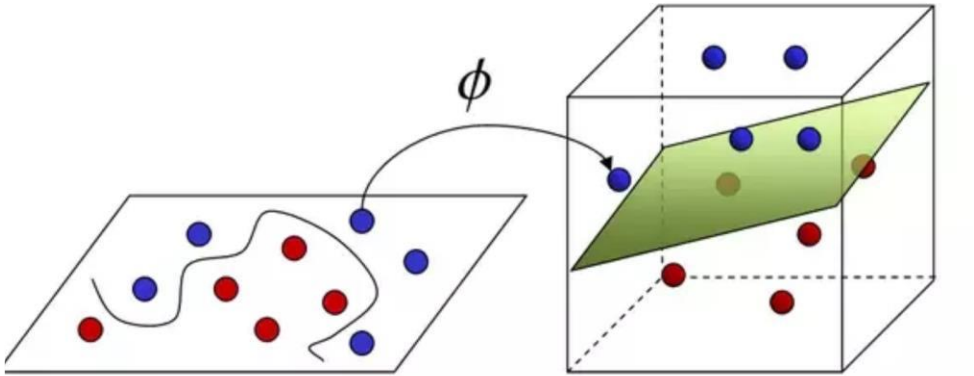
- 딥러닝 이전 우수한 분류기법으로 주목됨



2-2-3. SVM (Support Vector Machine)

- Kernalized SVM

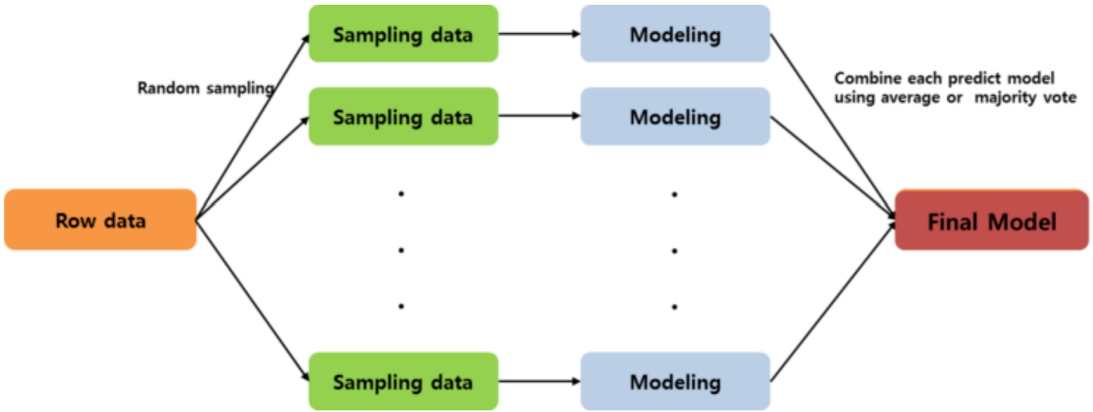
  선형 분류경계선만으로 분류가 불가능할 때, 이를 공간으로 옮겨 분류를 수행.
  Margin을 최대화하는 분류경계면을 연산



2-3. Ensemble

주어진 데이터로부터 여러 Model을 만들고, 이를 절충해서 최적의 Model을 산정하는 방식

2-3-1. Bagging



Raw data로부터 세부 표본을 랜덤 추출하여 Model을 여러 개 생성한 후에, 각 결과의 다수결 또는 평균 산정을 통해 최종 Model을 선정하는 방식.

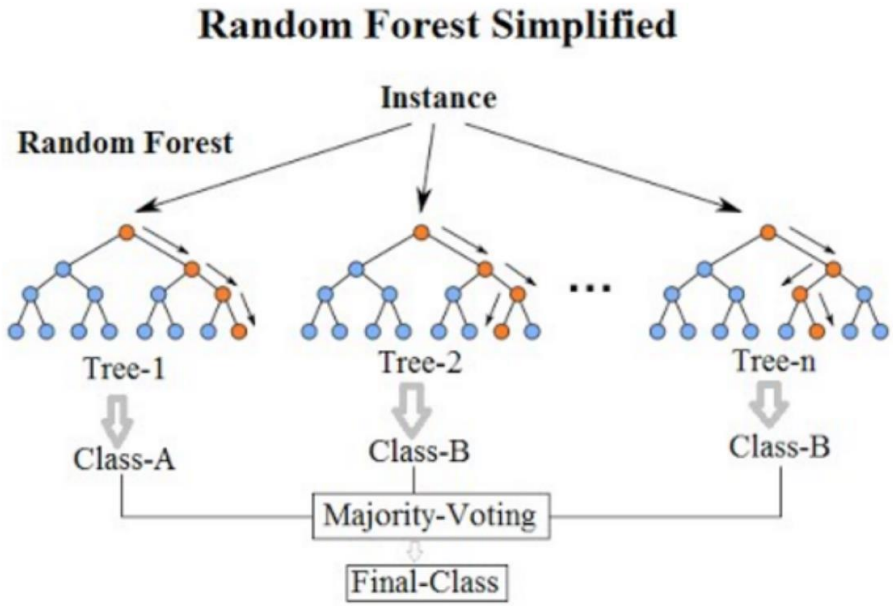
- Model의 성능은 개선되나, 설명력이 떨어지게 되는 단점이 있다.

2-3-2. Random Forest

- Decision Tree의 Ensemble 버전
- 단순 Bagging과는 차이가 있다.

Randomness를 표본 추출 뿐 아닌, Feature 선택에도 부여함.

Tree간의 이질성 강화, 상관성 감소



2-3-3. Boosting

- Bagging이 병렬적인 연산을 수행한다면, Boosting은 직렬적인 연산을 수행한다.
- 이전 시행의 결과가 다음 시행에 영향을 미침(Bagging은 독립적)
- Boosting에도 여러 종류가 존재.

1. **Ada Boost**
2. **Gradient Boost**
3. **XG Boost**

2-3-3. Boosting

- **Ada Boost**

  데이터셋에서 샘플을 추출하여 여러 분류기에 적용해서 학습시킨다.
  시행 결과 잘못 분류된 데이터를 집중적으로 학습하여 다음 fitting에 활용.
  Noise나 Outlier가 심한 데이터의 경우, 문제가 발생할 수 있다.

- **Gradient Boost**

  Gradient Descent를 Ada Boost에 적용한 기법.
  Outlier, Noise 문제를 해결할 수 있으나, 연산량이 많아짐

- **XG Boost**

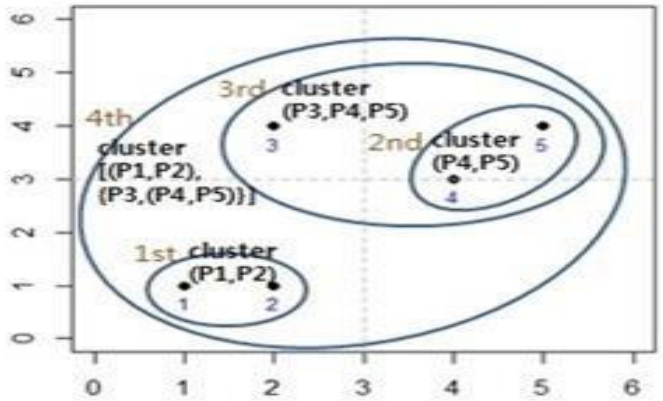
  Gradient Boost의 많은 연산량을 CPU 분산처리기법 등을 통해 소화하는 방식.

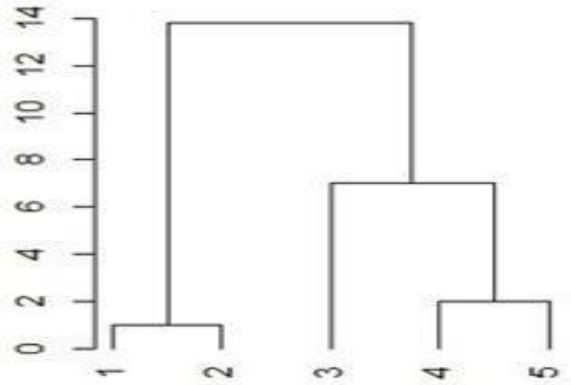
2-4. Clustering

비슷한 Data의 포인트들을 하나의 Cluster로 묶어, 데이터셋을 여러 Cluster로 나타내는 방식

2-4-1. Hierarchical Clustering

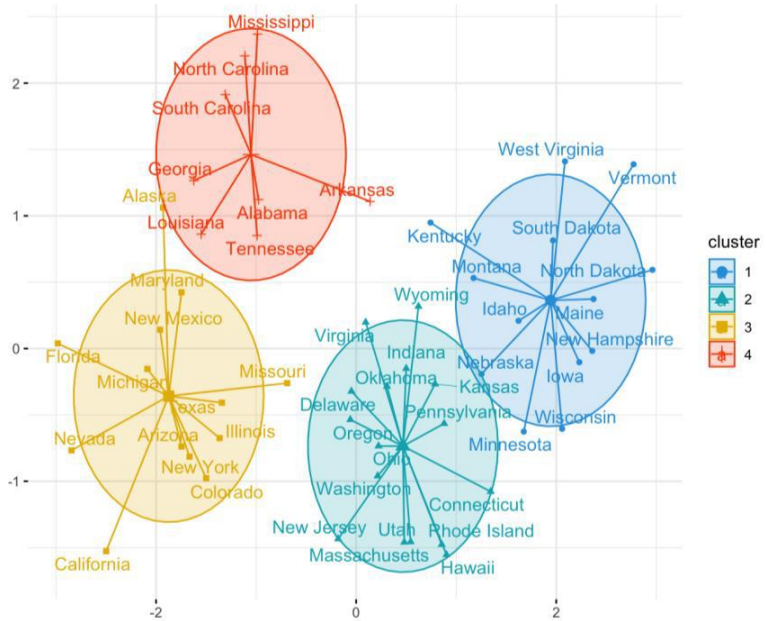
- 하나의 클러스터를 더 큰 클러스터가 포함하는 것을 반복하여 나타내는 방식.
- 동물 - 개 – 웰시코기와 같은 분류





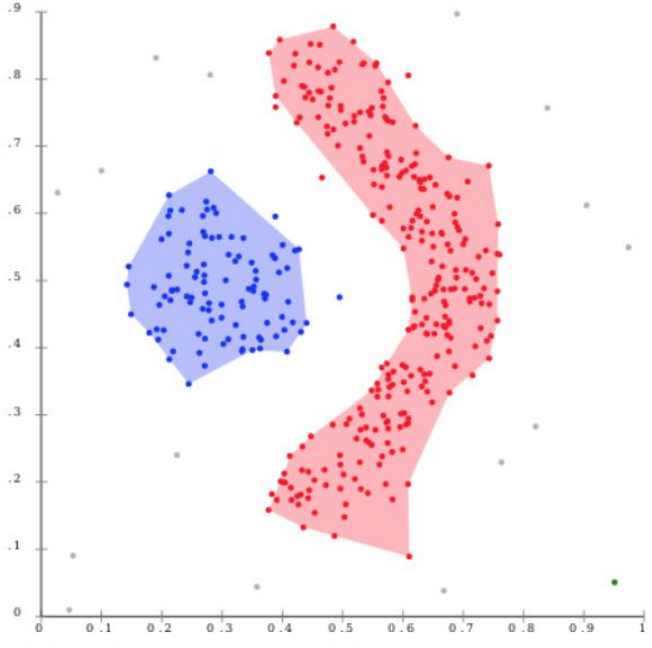
2-4-2. K-Means Clustering

- 군집 간의 거리는 최대화,
  군집 내의 데이터간 거리는
  최소화
- 비계층적 Clustering
- 사전에 클러스터의 개수를 잘
  설정해야 함(K).
- 거리를 기반으로 한 Clustering



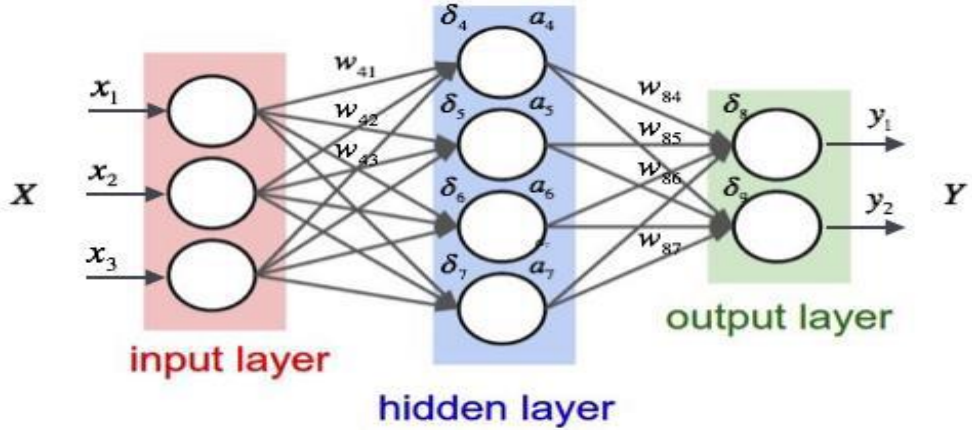
2-4-3. DBSCAN

- 밀도 기반의 클러스터링. 어느 점을 기준으로 반경 x내에 n개 이상의 데이터가 있으면, 클러스터링 하는 방식
- 비계층적 Clustering
- 사전에 Cluster의 개수를 설정할 필요가 없다.
- Noise나 Outlier에 취약하지 않음



KU-BIG

2-5. Neural Network



Input layer, Hidden layer, output layer로 구성되어, Hidden layer가 2개 이상인 경우에는 Deep Neural Network(DNN)이라 하고, 이것이 곧 딥러닝이다.

3. Loss Function (Cost Function)

- Model이 예측한 값과, 실제 값과의 차이를 나타내는 함수.
  즉, 모델이 얼마나 설명을 **못하는가**를 수치로 나타낸 것이다.

$$loss(f) = (y - \hat{y})^2$$

$$Cost\ Function = \frac{1}{m}\sum_{i=1}^{m} Loss\ Function^{(i)}$$

- 중요하다고 판단되는 Data의 패턴에 따라 Loss Function의 종류도 달라질 수 있다.
- MSE, Cross Entropy등이 있다.

KU-BIG

4. Optimization

Cost Function의 값을 최소화하기 위한 Parameter의 Set을 찾는 과정

4-1. Gradient Descent

- Cost Function의 값을 나타내면, 2차 포물선이 나오는데
  이 그래프의 기울기가 0이 되는 지점을 찾는 Algorithm(Cost Function이 최소가 되는 지점).
- 초기값을 최적 모수와 얼마나 가깝게 잡느냐에 따라 학습 속도가 달라진다.

- Single-variable case

f(x): cost function

Step 0	Choose a random point $x_0 \in \mathbb{R}$ and let $i = 0$	임의의 지점 x에서 시작.
Step 1	Compute $x_{i+1} = x_i - \alpha_i f'(x_i)$	기울기의 반대 방향으로 위치를 차츰차츰 내려가면서 갱신.
Step 2	If the termination condition (e.g., $\|f'(x_i)\| < \epsilon$) is satisfied, ***stop***	기울기 값이 충분히 작을 때까지 반복.
	Otherwise, let $i := i + 1$ and go to ***Step 1***	

KU-BIG

수고하셨습니다!