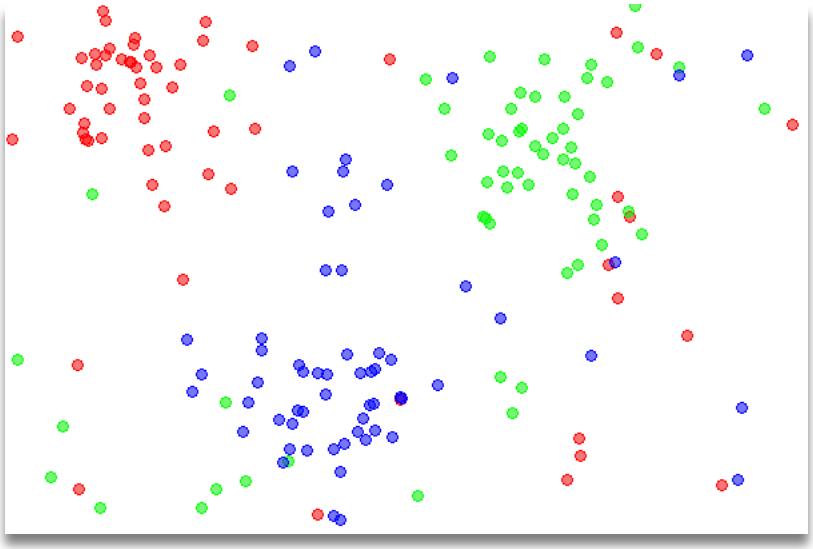
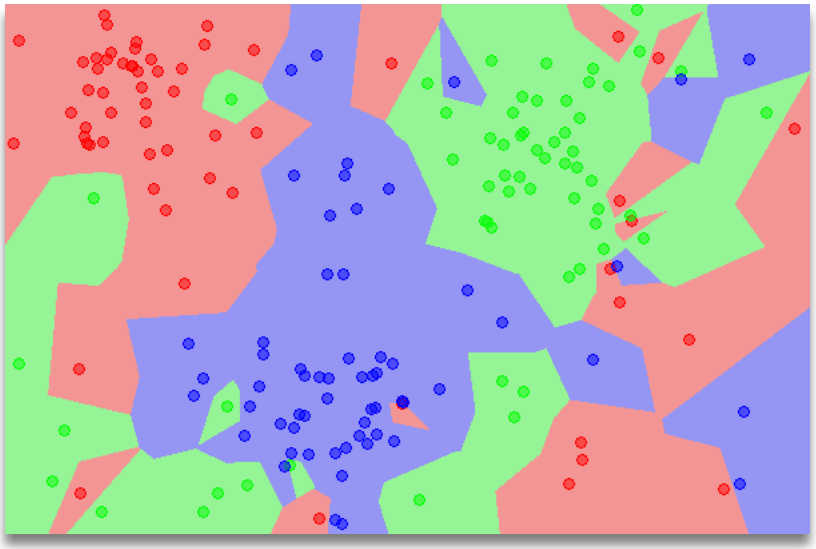
분류기

허석진

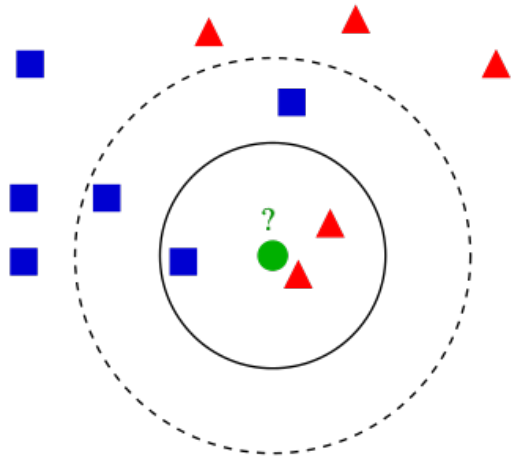
k-최근접 이웃 알고리즘 (k-NN)





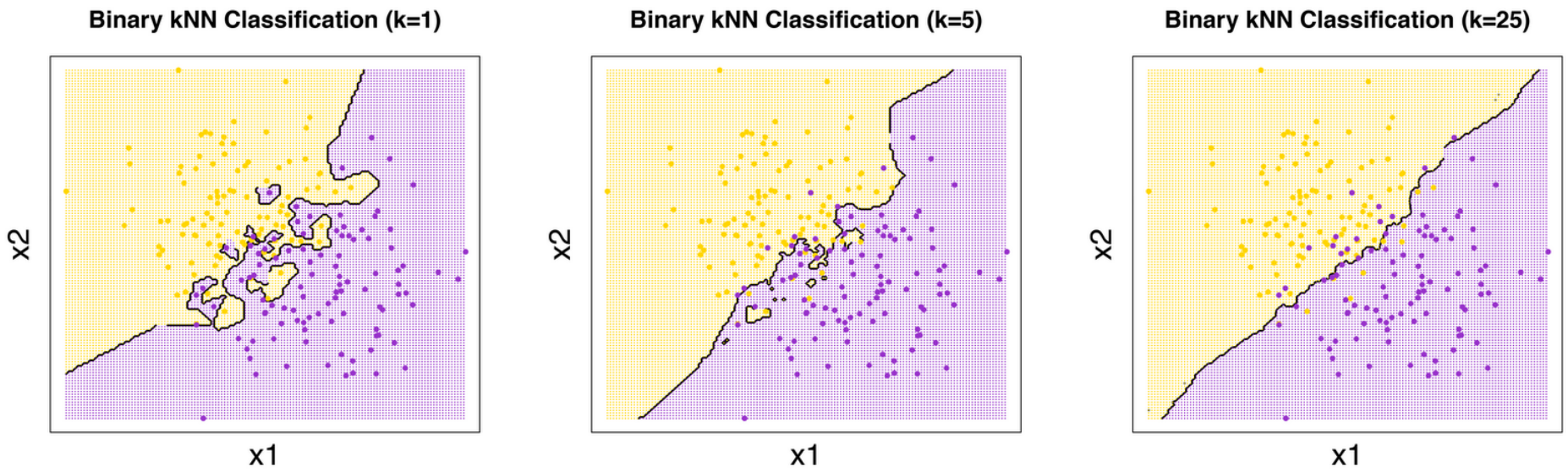
- 분류되어있는 훈련 데이터(좌) 사용
- 훈련 데이터 중 가장 가까운 k 개에 따른 분류 방법

k-최근접 이웃 알고리즘 (k-NN)



- 훈련 데이터: 파란색, 빨간색 점
- 분류하려는 초록색 점에서 최근접한 k 개의 훈련 데이터로 결정

k-최근접 이웃 알고리즘 (k-NN)



- 몇 개의 k가 좋은지?
- k가 증가함에 따라 과적합 감소, 잘못된 분류 증가
- 거리에 반비례하는 가중치를 주는 방법도 있음

나이브 베이즈 분류 (Naïve Bayes)

- 베이즈 정리
  - $P(H|X) = \frac{P(X|H)P(H)}{P(X)} = \frac{P(X|H)P(H)}{\sum P(X|H_i)P(H_i)}$
  - H: 확률을 판정하려는 사건, H_i: 모든 가능한 사건
  - X: 관측된 사건
- 암 진단에 관한 예시 (빨간색 값이 주어진 값)
  - H: 암 발병, $P(H) = 0.01$
  - H': 암이 아님, $P(H') = 0.99$
  - $P(X|H) = 0.9$: 암이 걸렸을 때 진단률
  - $P(X|H') = 0.1$: 암이 걸리지 않았을 때 진단률
  - X: 암 진단, $P(X) = 0.9*0.01+0.1*0.99 = 0.108$
  - $P(H|X) = 0.083$ (암 진단을 받았을 때 암 발병율)

나이브 베이즈 분류 (Naïve Bayes)

- 나이브 베이즈 분류 알고리듬
  - $P(X|H_i)$는 조건부 확률 $P(x_k|H_i)$들의 단순 곱 (x_k들의 독립을 가정)
  - $P(H_i|X)\left(=\frac{P(X|H_i)P(H_i)}{P(X)}\right)$ 이 최대가 되는 H_i 탐색
  - X를 H_i로 분류
  - 여기서 X가 이산형이면 x_k가 나타나는 개수로부터 확률 계산, X가 연속형이면 정규분포를 가정하고 평균과 표준편차에 따른 확률밀도함수 사용
- 왜 Naïve 인가? - x_k들이 독립이라고 가정하기 때문에.

스팸 이메일 분류

- 사전 확률
  - $P(S)$: 전체 메일 중 스팸의 비율
  - $P(H) = 1 - P(S)$: 전체 메일 중 스팸이 아닌 비율 (햄이라고 부름)
- 단어 w가 메일 내용에 포함될 때 그 메일이 스팸일 확률
  - $P(S|w) = \frac{P(w|S) \cdot P(S)}{P(w)}$
  - $P(w|S)$: 스팸인 메일에서 단어 w가 포함되는 확률
  - $P(w)$: 단어 w가 전체 메일에서 포함되는 확률

스팸 이메일 분류

- 나이브 베이즈 분류
  - 스팸에서 단어 w_k가 포함되는 확률 $P(w_k|S)$ 파악
  - 햄에서 단어 w_k가 포함되는 확률 $P(w_k|S)$ 파악
  - $P(S|w_k) = \frac{P(w_k|S)\cdot P(S)}{P(w_k)} = \frac{P(w_k|S)\cdot P(S)}{P(w_k \cap S)+P(w_k \cap H)} = \frac{P(w_k|S)\cdot P(S)}{P(w_k|S)\cdot P(S)+P(w_k|H)\cdot P(H)}$
  - $P(S|w_1), P(S|w_2), \ldots, P(S|w_n)$을 곱하여(Naïve!) 단어들의 집합 W에 대해 $P(S|W)$ 계산
  - $P(H|W)$도 유사한방법으로 계산하여 $P(S|W)$ 이 더 크면 스팸 (또는 $P(S|W)$이 특정 값보다 크면 스팸)