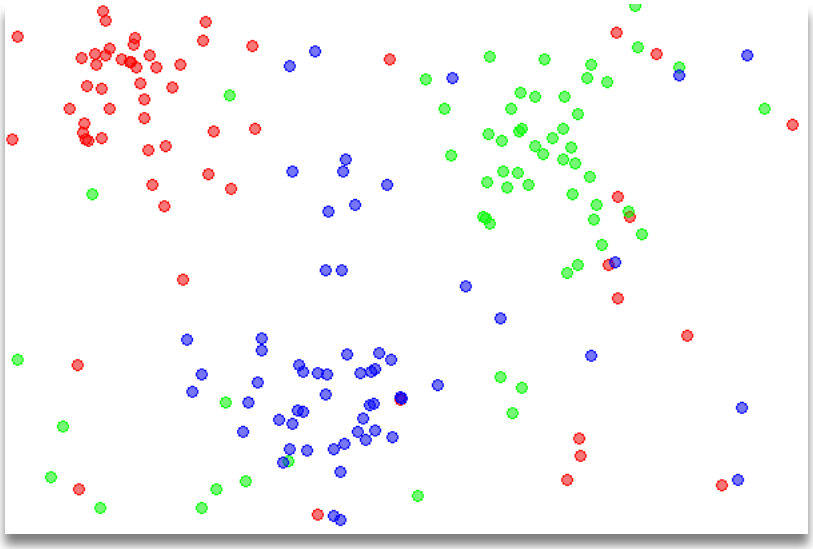
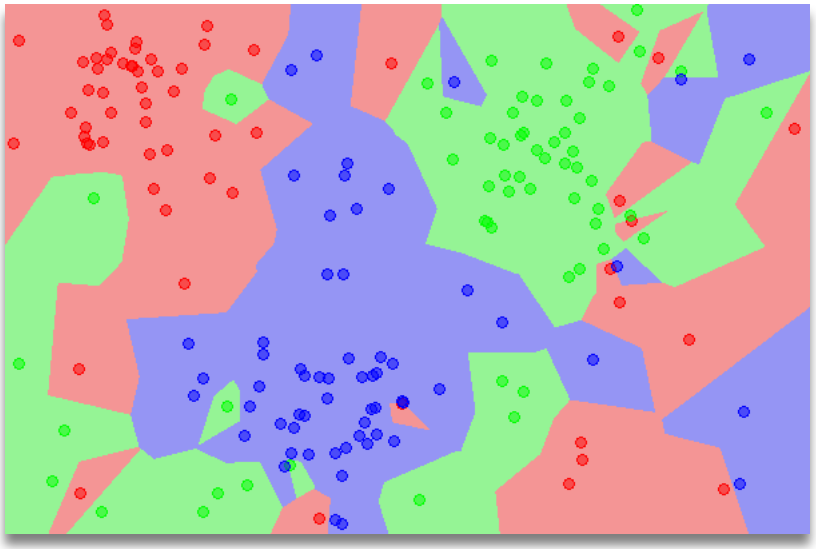
분류기

허석진

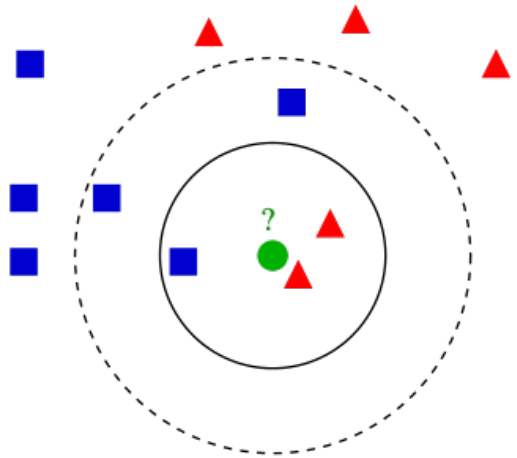
k-최근접 이웃 알고리즘 (k-NN)





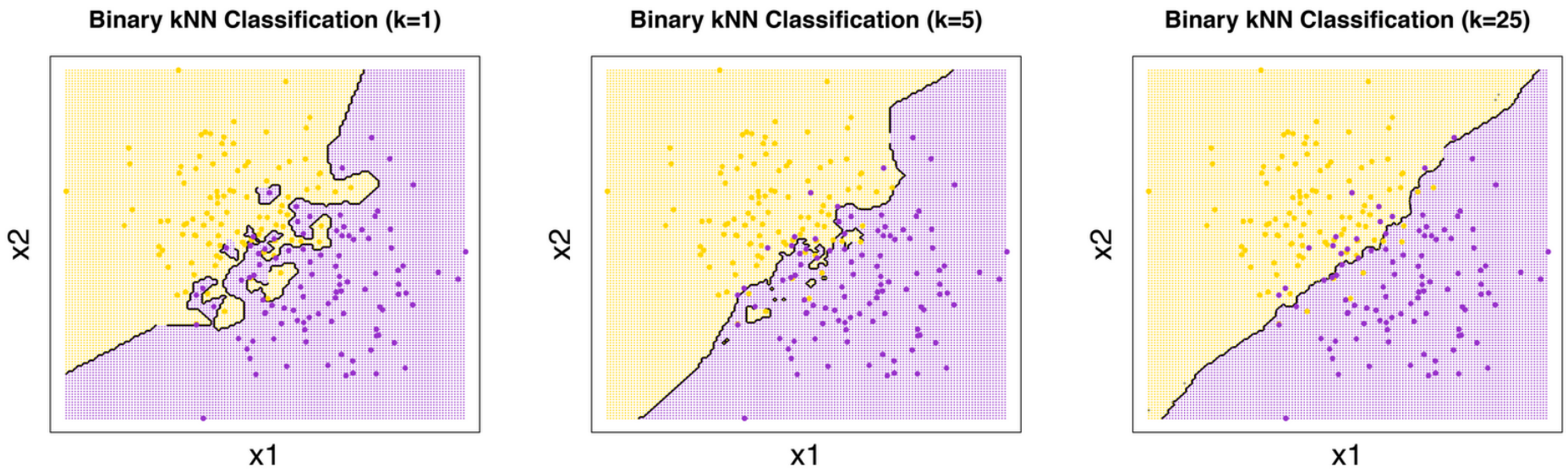
- 분류되어있는 훈련 데이터(좌) 사용
- 훈련 데이터 중 가장 가까운 k 개에 따른 분류 방법

k-최근접 이웃 알고리즘 (k-NN)



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

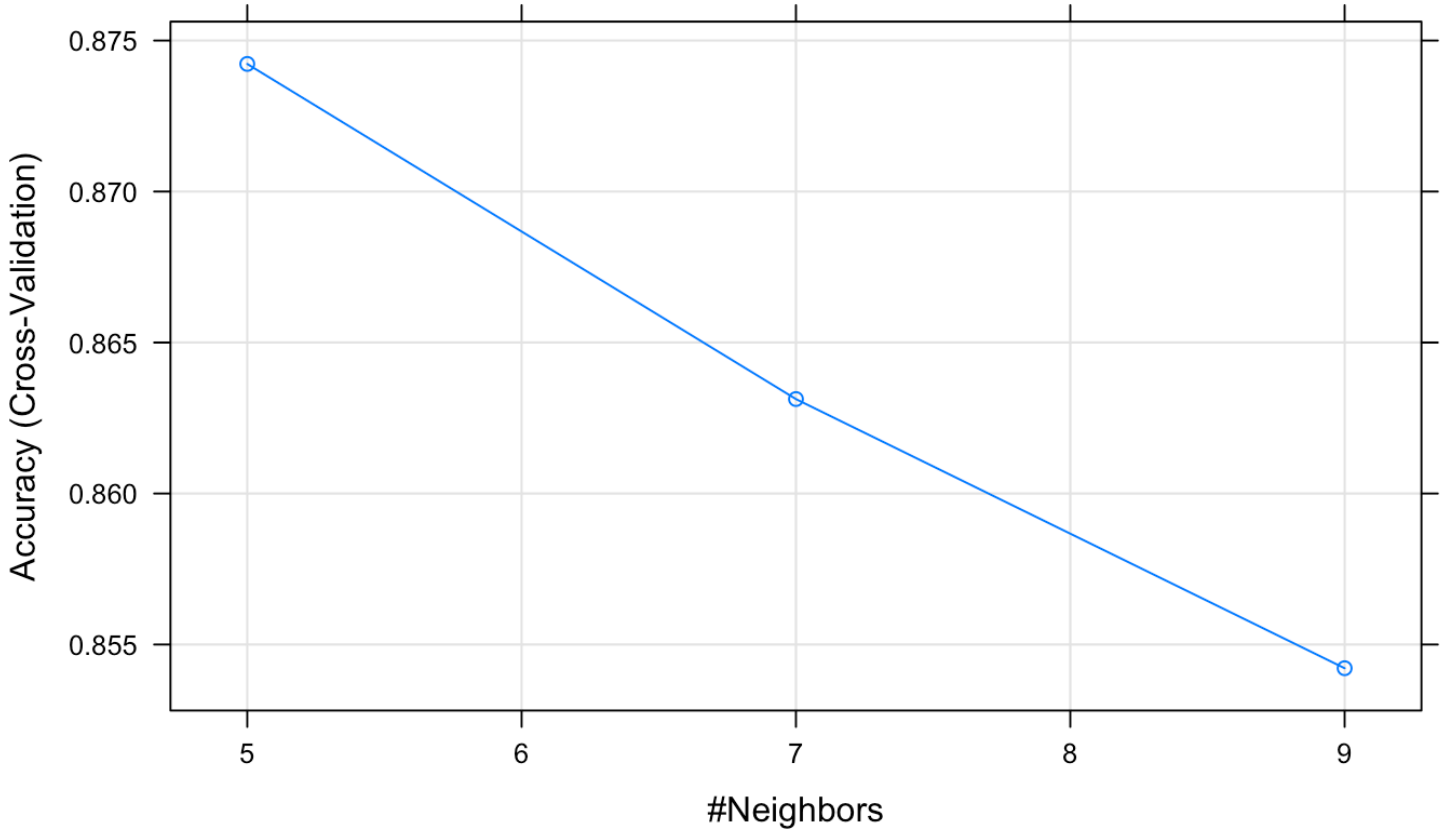
k-최근접 이웃 알고리즘 (k-NN)



- 몇 개의 k가 좋은지? k = 1, …., 훈련 데이터 전체 크기
- k가 증가함에 따라 과적합 감소, 잘못된 분류 증가
- 거리에 반비례하는 가중치를 주는 방법도 있음

k-최근접 이웃 알고리즘 (k-NN)

- 최적의 k 찾기



k-최근접 이웃 알고리즘 (k-NN)

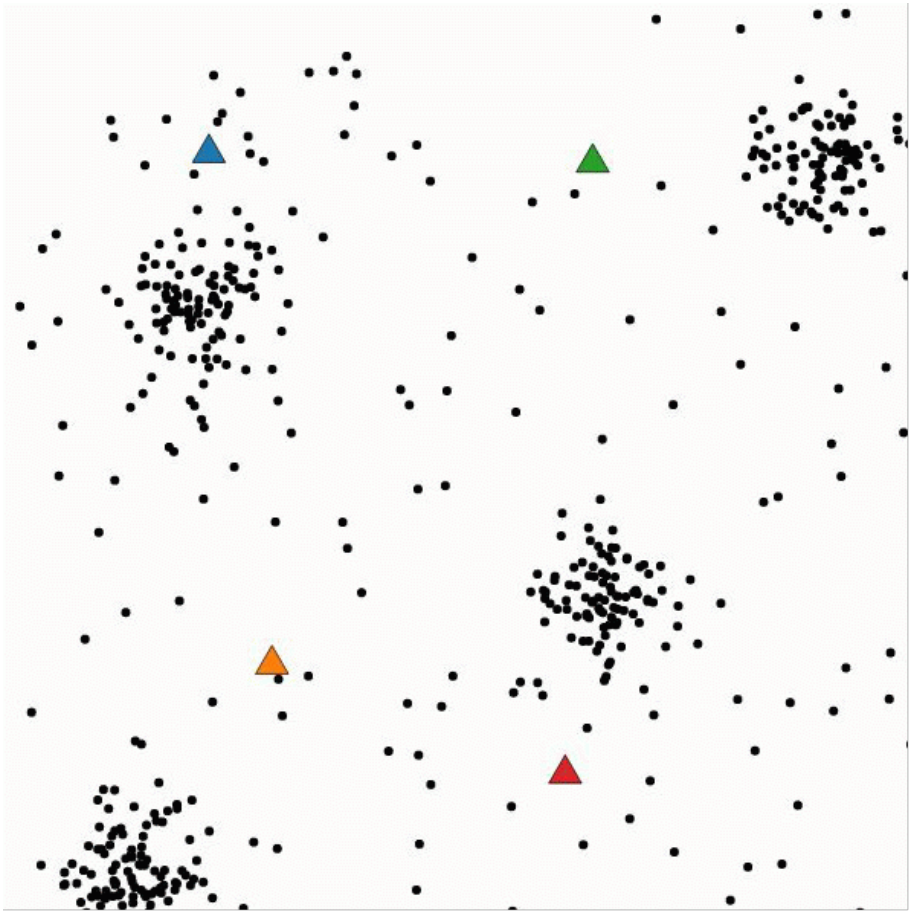
- 선형회귀와 비교
  - 모수적 <-> 비모수적 방법
  - 선형적인 모형을 위해서는 선형 회귀가 유리
  - 오차의 분포 / 선형성을 가정할 수 없는 기하학적인 분류에는 kNN
  - kNN의 계산량이 많음

k-평균 알고리듬 (k-means)

- 알고리듬
  1. 무작위로 k 개의 점을 선택
  2. 모든 데이터 각각을 1번의 k 개 점 중 가장 가까운 점에 할당하여 클러스터 구성
  3. 각 클러스터들의 평균점 계산 (k 개 존재)
  4. 3번의 평균점을 기준으로 2번부터 반복
  5. 클러스터들이 변하지 않으면 종료

k-평균 알고리듬 (k-means)

- 알고리듬 시각화



k-평균 알고리듬 (k-means)

- 특징
  - 비지도 학습
  - 알고리듬 단순
  - k와 최초에 선택한 중심점에 민감함
  - 클러스터의 모양에 따라 잘 작동하지 않을 수 있음

베이즈 정리 (Bayes Theorem)

- 정리
  - $P(H|X) = \frac{P(X|H)P(H)}{P(X)} = \frac{P(X|H)P(H)}{\sum P(X \cap H_i)} = \frac{P(X|H)P(H)}{\sum P(X|H_i)P(H_i)}$
  - H: 확률을 판정하려는 사건, H_i: 모든 가능한 사건
  - X: 관측된 사건

베이즈 정리 (Bayes Theorem)

- 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 (*또는 Kernel Density Estimation*)
- 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 유사한 방법으로 계산하여 $P(S|W)$ 이 더 크면 스팸 (또는 $P(S|W)$이 특정 값보다 크면 스팸)

키와 몸무게로부터 성별 분류

- 사전 확률
  - $p(M)$: 전체에서 남성의 비율
  - $p(F) = 1 - p(M)$: 전체에서 여성의 비율
- 특정 키 h에 대해 그가 남성일 확률
  - $p(M|h) = \frac{p(h|M) \cdot p(M)}{p(h)}$
  - $p(h)$: 전체에서 키가 h인 확률 $= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(h-\mu)^2}{2\sigma^2}\right)$
  - $p(h|M)$: 남성 중 키가 h인 확률 $= \frac{1}{\sqrt{2\pi\sigma_M^2}} exp\left(\frac{-(h-\mu_M)^2}{2\sigma_M^2}\right)$

키와 몸무게로부터 성별 분류

- 나이브 베이즈 분류
  - 전체 키 데이터로부터 h의 확률 밀도 $p(h)$ 계산
  - 남성 키 h의 확률 밀도 $p(h|M)$ 계산
  - 같은 방법으로 여성 키 h의 확률 밀도 $p(h|F)$ 파악
  - $p(M|h) = \frac{p(h|M)\cdot p(M)}{p(h)} = \frac{p(h|M)\cdot p(M)}{p(h\cap M)+p(h\cap F)} = \frac{p(h|M)\cdot p(M)}{p(h|M)\cdot p(M)+p(h|F)\cdot p(F)}$
  - $p(M|h), p(M|w)$ 을 곱하여 $p(M|h,w)$ 계산
  - $p(F|h,w)$도 유사한 방법으로 계산하여 $p(M|h,w)$ 가 더 크면 남성

나이브 베이즈 장단점

- 장점
  - 우수한 성능
  - 모형이 단순
  - 필요한 훈련 데이터의 양이 적음
- 단점
  - 변수들이 독립적이지 않은 경우 적용이 어려움