머신러닝 소개

허석진

강의계획

1. 개요와 로지스틱 회귀
2. 나이브 베이즈 분류기와 kNN
3. 추천시스템
4. 의사결정나무
5. 앙상블 메서드 (Bagging, Random Forest, Boosting), 데이터 전처리
6. SVM (Support Vector Machine)

- 참고서적
  An Introduction to Statistical Learning (역서: 가볍게 시작하는 통계학습)
  -> http://www-bcf.usc.edu/~gareth/ISL/

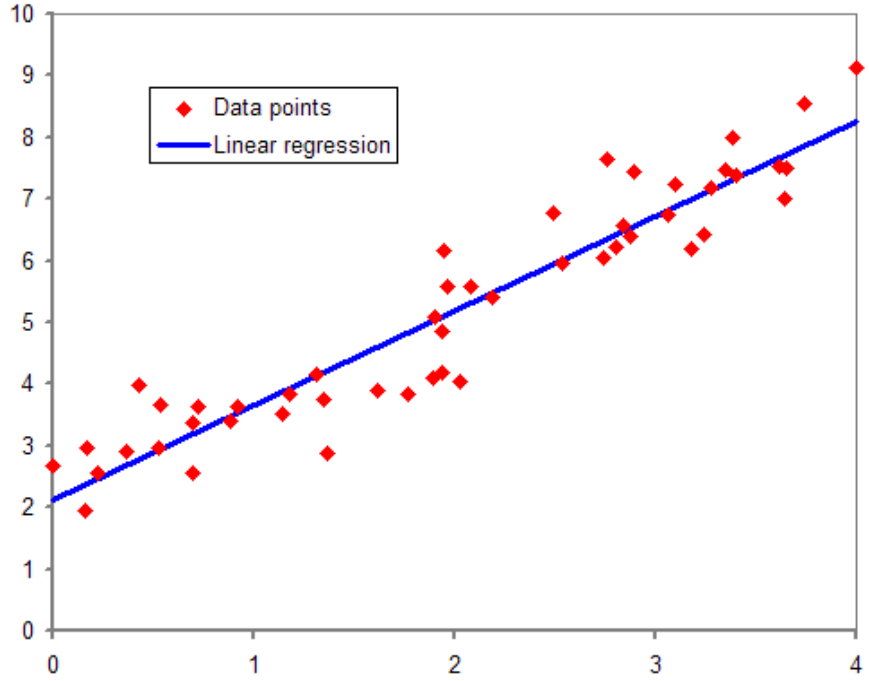
머신러닝/기계학습 (Machine Learning)

- 정의
  - 컴퓨터가 학습하여 정답에 가까운 행동을 하게 하는 알고리즘과 기술을 개발하는 분야
  - 여기서 정답은 훈련(training) 데이터의 형태로 주어짐
  - 모형을 만드는 일이라고도 할 수 있음
- 데이터 마이닝과의 차이
  - 머신러닝은 예측이 목표
  - 데이터 마이닝은 속성 발견에 중점

머신러닝의 분류

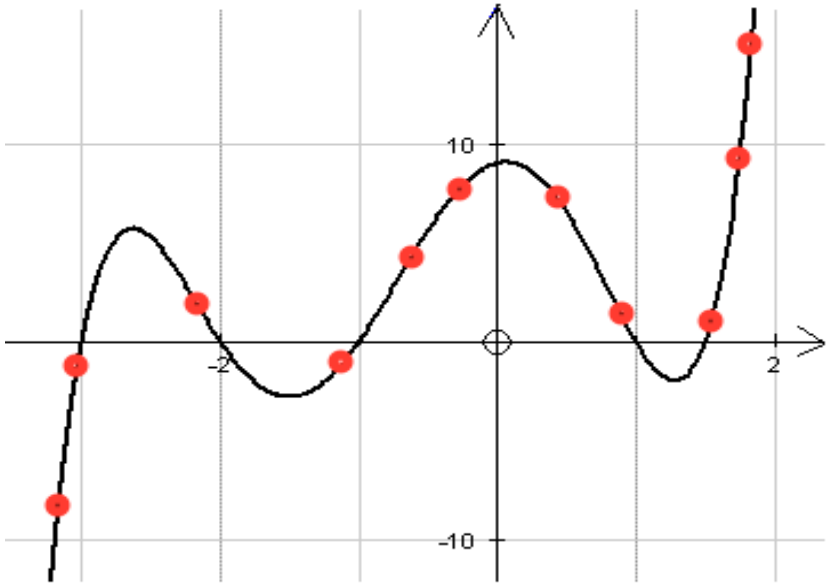
- 지도학습 (Supervised Learning)
  - 입력에 대한 속성과 목표값 명시
  - 예) 회귀분석, k-NN, SVM, 의사결정나무
- 자율 학습/비지도학습 (Unsupervised Learning)
  - 입력에 대한 목표값이 없음
  - 예) 군집화(clustering)
- 강화학습 (Reinforcement Learning)
  - 상태와 행동에 따른 보상을 최적화

과적합 (1)



- 훈련데이터(빨간 점)로부터 예측 모형(파란 직선) 생성
- 특징변수: 가로축

과적합 (2)



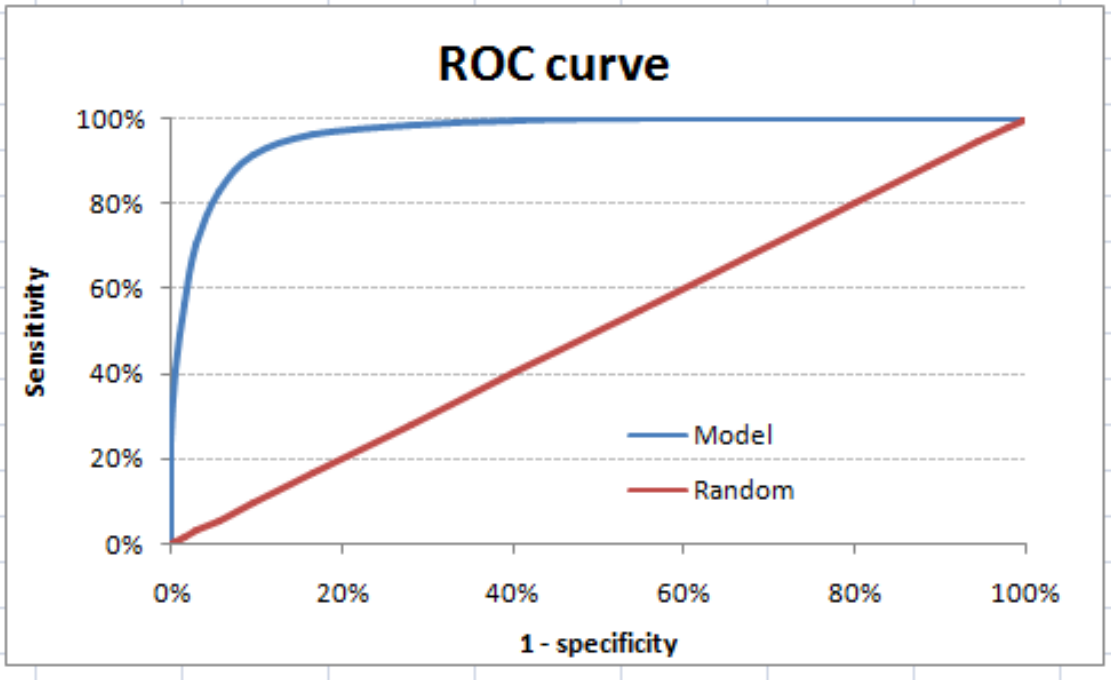
- 훈련데이터(빨간 점)로부터 예측 모형(곡선) 생성?

혼동행렬 (Confusion Matrix)

	긍정 예측	부정 예측
실제 분류	참 긍정	거짓 부정
실제 비분류	거짓 긍정	참 부정

- 정확도(Accuracy): 전체에서 올바른 예측 비율
- 정밀도(Precision): 긍정 예측 중에서 올바른 예측 비율
- Negative Predictive Value: 부정 예측 중에서 올바른 예측 비율
- 민감도(Sensitivity): 실제 분류 중 올바른 예측 비율 (Recall)
- 특이도(Specificity): 실제 비분류 중 올바른 예측 비율
- 어떤 값이 중요?

수신자 조작 곡선 (ROC)



- 가로축: 실제 관측되지 않은 경우 긍정으로 예측한 비율
- 세로축: 실제 관측된 경우 긍정으로 예측한 비율
- AUC: ROC 아래 영역의 넓이

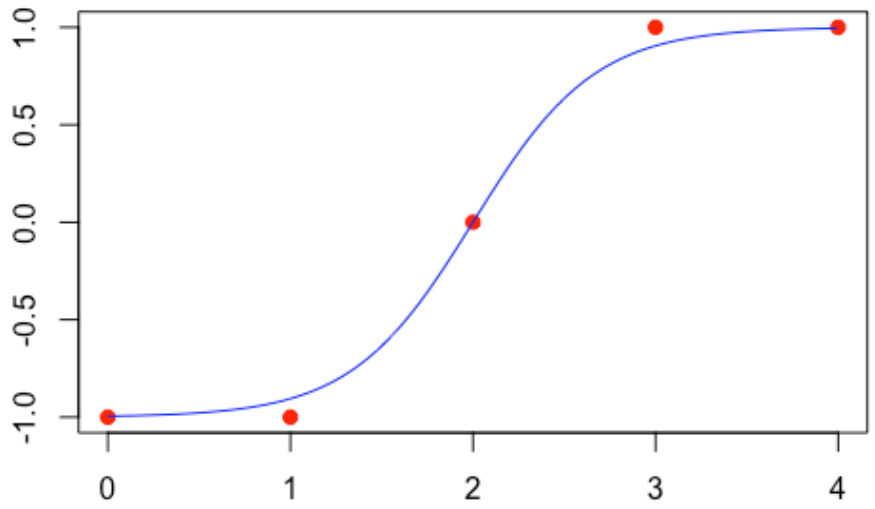
교차검증 (cross validation)

- 데이터를 모두 훈련에 사용하는 것이 아니라 일부를 남겨서 도출한 모형에 적용시켜 보는 것
- Leave-p-out cross-validation
  - p 개의 관측값을 검증 데이터로 사용하고 나머지 전부를 훈련 데이터로 사용하는 방식으로 반복
  - 전체 데이터에서 p 개의 관측값을 모든 가능한 방법으로 반복
  - 특수한 경우는 p=1인 Leave-one-out cross-validation
- k-fold cross validation
  - 전체 데이터를 k 부분으로 나누고 이 중 k-1 부분만 훈련에 사용하고 나머지를 검증에 사용하는 방식으로 k 번 반복

기타 모형 평가

- MAE (mean absolute error)
  - $\frac{\sum_{i=1}^{N} \left|\text{예측값}_i - \text{실제값}_i\right|}{N}$
- MSE (mean squared error)
  - $\frac{\sum_{i=1}^{N} \left(\text{예측값}_i - \text{실제값}_i\right)^2}{N}$
- RMSE (root mean squared error)
  - $\sqrt{MSE}$

로지스틱 회귀



- 데이터가 연속형이고 선형적인 경우 적용
- 범주형일 경우는?

로지스틱 회귀

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \leftrightarrow \pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

$$\text{logit: } \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

odds