12주차:
모델의 성능 평가

ChulSoo Park

School of Computer Engineering & Information Technology

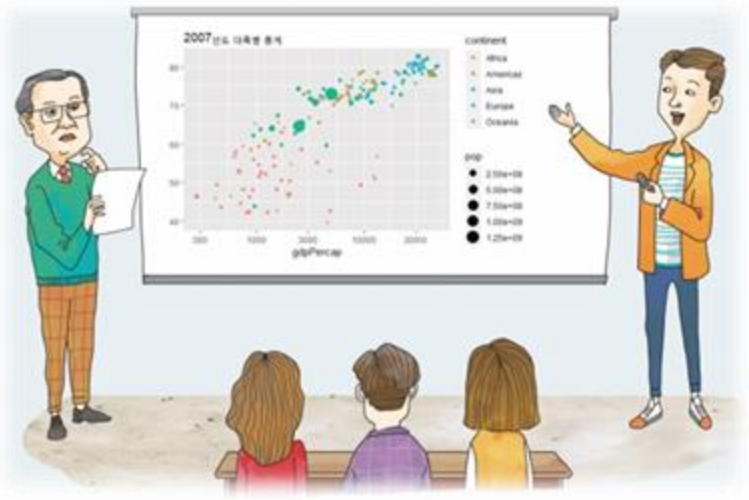
Korea National University of Transportation

학습목표 (12주차)

- 예측 오류 발생 이유 이해
- 정확률 계산 방법 학습
- 일반화 능력 측정 방법 습득
- 교차 검증 방법 이해
- 모델 선택 방법 학습
- 정밀도와 재현율 학습
- ROC 곡선과 AUC 이해

10 CHAPTER

모델의 성능 평가

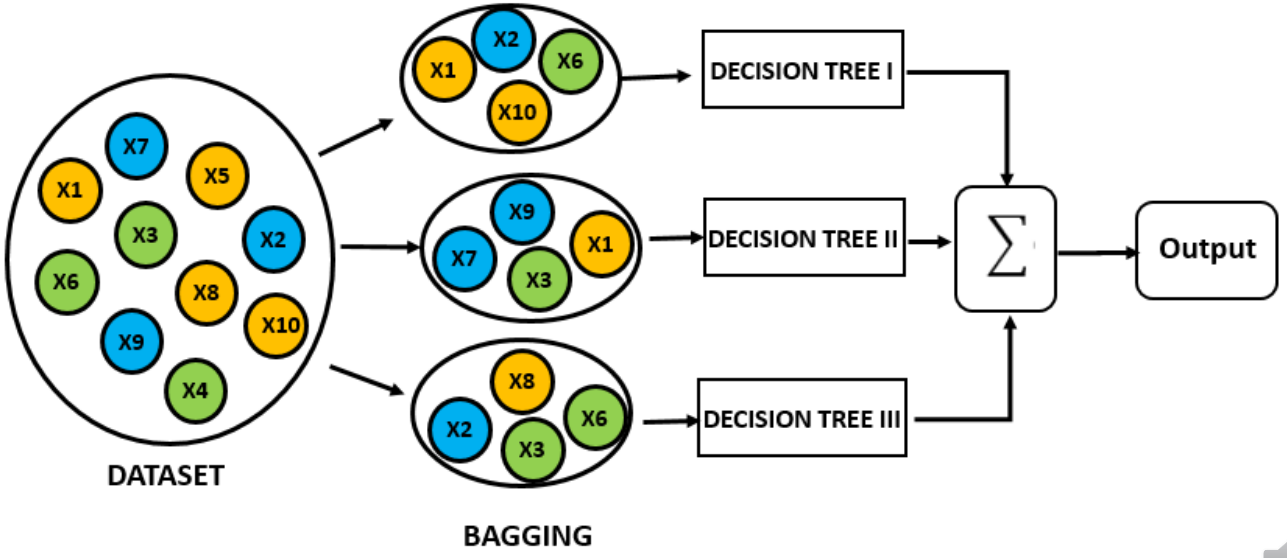


CONTENTS

Review

random-forest

As we see, there is **multiple decision trees** as base learners. Each decision tree is given a **subset of random samples from the data set** (Thus, the name 'Random'). The Random Forest algorithm uses **Bagging** (Bootstrap Aggregating) which we learned in **ensemble methods**. The general idea of the ensemble methods is that a combination of learning models **increases the overall result.**

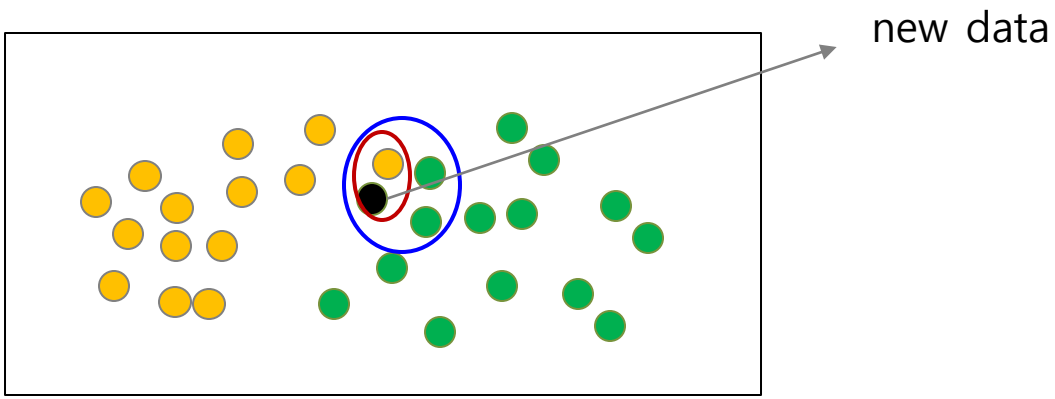


https://medium.com/greyatom/a-trip-to-random-forest-5c30d8250d6a

Review

K-최근접이웃(k-NN, K-Nearest Neighbor)

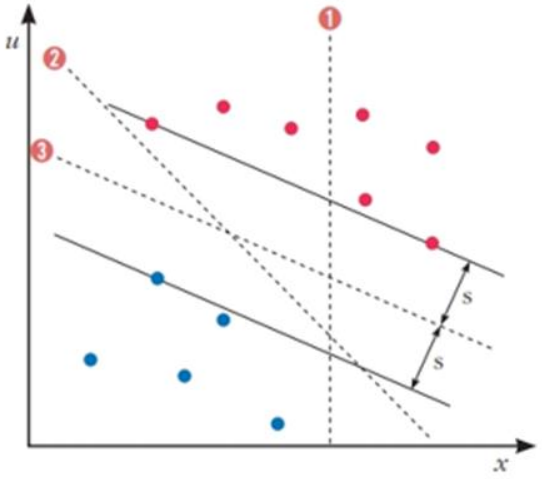
- k-NN은 새로운 데이터가 주어졌을 때 기존 데이터 가운데 가장 가까운 k 개 이웃의 정보로 새로운 데이터를 예측하는 방법론.
- 아래 그림처럼 검은색 점의 범주 정보는 주변 이웃들을 가지고 추론해낼 수 있음.
- 만약 k가 1이라면 오렌지색, k가 3이라면 녹색으로 분류(classification)하는 것.
- 만약 회귀(regression) 문제라면 이웃들 종속변수(y)의 평균이 예측값이 됨.



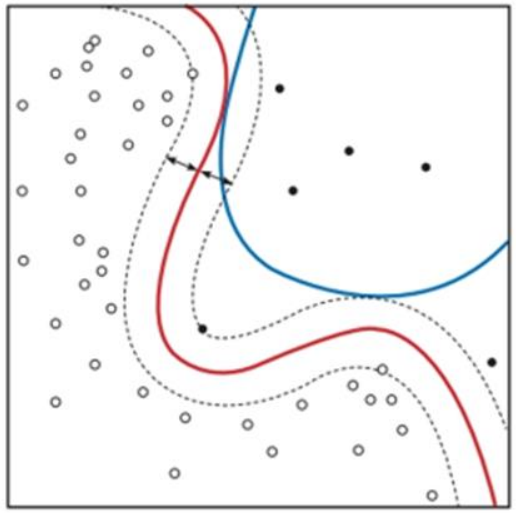
출처 : 데이터 사이언스 개론(김화종,홍릉과학출판사), p.175

Review

- SVM의 원리
  - 모델 ①은 빨간 샘플 3개를 잘못 분류 → 오류율 3/12=25%
  - ②와 ③은 오류율 0% ← 둘의 성능은 같나? SVM의 원리는 이 질문에서 출발
  - ②는 빨간 샘플에 조금만 변형이 발생해도 경계를 넘어 오류 발생할 가능성 높음
  - ③은 두 부류 모두에 대해 멀리 떨어져 있어 변형이 발생해도 경계를 넘을 가능성 낮음
  - 일반화 측면에서 모델 ③이 더 뛰어남. SVM 학습은 여백(margin) [아래 그림]에서 $2s$)을 최대로 하는 최적 모델을 찾아 줌



(a) 선형 SVM



(b) 비선형 SVM

Preview

평가의 중요성

- 회사가 신입사원을 뽑을 때, 실기와 필기 중 어느 것에 가중치를 두느냐에 따라 합격자가 달라짐 → 업무에 적합한 사람을 선택하려면 평가 기준이 중요함

Preview

- 데이터 과학에서의 성능 평가
  - 모델 선택에 필수
    - ✓ 예) 1000개 샘플 중 랜덤 포리스트가 850개, SVM이 910개를 맞춘다면 SVM을 선택
    - ✓ 정확률 대신 정밀도와 재현율(10.6절)을 평가 기준으로 사용한다면 다른 결과일 수 있음
  - 상황에 적합한 성능 평가 기준을 골라 사용하는 일의 중요성
- 학습을 마친 모델, 즉 예측 시스템을 현장 설치할지 결정할 때 중요
  - 예) colon 데이터의 경우 성능 평가 결과가 어느 정도 이상을 넘어야 현장 설치 가능

10.1 예측 오류는 왜 발생하나?

- 첫째, 세상은 불확실성 투성이(코로나바이러스감염증-19(COVID-19)
  - 목소리를 보고 성별을 구분하는 경우, 목소리가 가는 남성은 여성의 음성과 흡사. 같은 남성이라도 평소 굵은 목소리를 냈는데 무척 피곤한 경우 가는 목소리
  - 기상청의 날씨 예측 오류, 프로모션은 매출 변화에 중요한 영향 요인

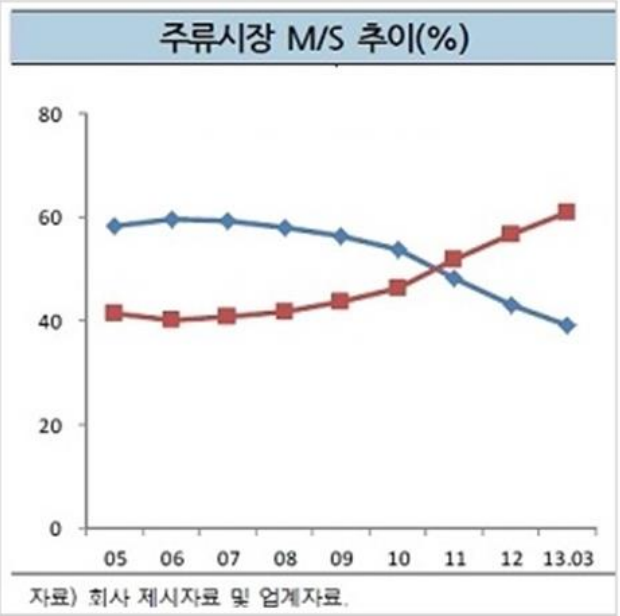
- 둘째, 데이터의 불완전성
  - 데이터를 측정할 때 기구의 불완전성이나 사람의 불완전성
  - 데이터 양이 작아 현장을 완전히 대표하지 못함. 예) colon 데이터의 경우 아무리 많은 대장암 환자 데이터를 모으더라도 데이터에 없는 특수 체질이 새로 발생함

- 데이터 과학이 할 수 있는 일
  - 엄정한 성능 평가 기준을 세우고, 여러 모델을 성능 평가하여 가장 뛰어난 모델을 선택하고, 성능이 일정 수준 이상이 되면 현장 설치

10.1 예측 오류는 왜 발생하나?

수요예측 기법 및 고려사항

2014.10.14(화)



10.1 예측 오류는 왜 발생하나?

수요예측의 필요성

- 미래 계획 수립에 불확실성의 감소
- 변화를 예상하고 관리
- 원활한 의사소통
- 재고수준, 생산 능력 및 리드타임의 관리
- 원가 추정
- 고객 만족 및 생산성 향상

10.1 예측 오류는 왜 발생하나?

예측 영향 요인들

- 데이터 소스(Data Sources)
  - Published data, Original Data and Time series
- 예측 기법(Forecasting Methods)
  - 정성적 기법 및 정량적 기법
- 통합의 정도(Grouping)
  - 시간, 지리적 위치 및 제품 그룹
- 데이터 관리 주기(Time Dimensions)
- 예측 범위(Forecast Ranges)
  - 장기, 중기 및 단기 예측
- 데이터 품질 및 정확도(Data Quality and Accuracy)

프로모션 (매출 2배 이상 증가)
경쟁 기업의 프로모션 정보 ?

10.1 예측 오류는 왜 발생하나?

확산 모형 분석 사례

대형할인점 Big3 확산모형 추정 파라메터

	bass 추정	로지스틱
M	441.2	430.8
p or a	0.0035	4.7380
q or b	0.3123	-0.3342
R^2	0.9952	0.9946
수정 R^2	0.9942	0.9936
exit flag	1	1

Big3 대형 할인점(이마트,홈플러스,롯데마트) 점포 누계 예측치

Open 년도	실 data	Bass	로지스틱	Open 년도	실 data	Bass	로지스틱
1993년	1	1.8	5.2	2006년	2(		209.0
1994년	2	4.2	7.2	2007년	228	243.9	244.8
1995년	4	7.5	10.0	2008년	294	277.7	279.0
1996년	6	11.9	13.9	2009년	310	308.9	310.1
1997년	9	17.8	19.2	2010년	345	336.3	336.9
1998년	17	25.7	26.3	2011년	361	359.5	359.1
1999년	29	36.0	35.9	2012년	370	378.5	376.9
2000년	51	49.3	48.5	2013년		393.7	390.8
2001년	79	66.2	64.9	2014년		405.5	401.4
2002년	103	87.1	85.5	2015년		414.6	409.3
2003년	121	112.3	110.7	2016년		421.5	415.2
2004년	136	141.6	140.3	2017년		426.7	419.5
2005년	160	174.1	173.6				
				SSE		7.4	7.5

- ✓ 잠재 수요 : Bass모형의 잠재수요가 4410이고 로지스틱 모형이 431로 Bass모형이 조금 높게 나타남.
- ✓ Bass 모형의 환산 계수를 살펴보면 혁신계수 p는 0.0035로 일반적인 내구재의 혁신 계수 정도이며 모방계수 p도 0.3123으로 일반적인 내구재의 모방계수 수준임.
- ✓ R^2의 값이 0.99로 독립 변수에 의한 종속 변수의 설명력이 좋으며 수정 R^2의 값도 각각 0.99로 0.7 이상으로 나타나 좋은 결과이며 exit flag값이 1로 확인 되었다.
- ✓ 추정된 모형의 예측치와 실제의 data는 거의 일치함을 보여주고 있다. 각 확산 모형과 실제 data의 적합도를 비교하는 에러 제곱합의 자연로그 값 비교도 거의 차이를 보이지 않고 있다.
- ✓ 예측결과 그래프는 다음 장에 표시함.

10.2 정확률

가장 널리 활용되는 정확률

- 전체 Sample 수 : n, 정답 수 : n_1, 오답 수 : n_2
- 정확률 : $\frac{n_1}{n}$, 오류율 : $\frac{n_2}{n}$

- 예) iris에 있는 150개 샘플
  - ✓ 정확률 : 105개 중에 102개를 맞추었다면 102/150=68%
  - ✓ 오류율은 48/150=32%

10.2 정확률

- 기각(rejection) 기능이 있는 경우

  - 전체 Sample 수 : n, 정답 수 : n_1, 오답 수 : n_2 , 기각한 샘플 수 : n_3

    (n = $n_1 + n_2 + n_3$)

  - 정확률 : $\frac{n_1}{n}$, 오류율 : $\frac{n_2}{n}$, 기각률 : $\frac{n_3}{n}$

  - 예) 새로운 환자에 대해 0.6 확률을 예측했다면 확신이 서지 않기 때문에 기각 전문가인 의사에게 판정을 미룸

10.2 정확률

결정 트리 사례

- 전체 Sample 수 : 150, 정답 수 : n_1= 50+49+45 = 144, 오답 수 : n_2 = 0+1+5=6
- 정확률 : $\frac{144}{150}$ =96%, 오류율 : $\frac{6}{150}$ = 4%

Console C:/RSources/

```
> r = rpart(Species~., data = iris)                    # 결정 트리 모델
> f = randomForest(Species~., data = iris, ntree = 3)  # 랜덤 포리스트 모델
> r_pred = predict(r, iris, type = 'class')
> confusionMatrix(r_pred, iris$Species)
Confusion Matrix and Statistics

            Reference
Prediction   setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         49         5
  virginica       0          1        45

Overall Statistics

               Accuracy : 0.96
                 95% CI : (0.915, 0.9852)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.94
```

10.2 정확률

랜덤 포스트 사례

- (1)전체 Sample 수 : 150, 정답 수 : n_1= 50+50+48 = 148, 오답 수 : n_2 = 0+0+2=2
- (2)전체 Sample 수 : 150, 정답 수 : n_1= 50+48+49 = 147, 오답 수 : n_2 = 0+2+1=3
- 정확률1 : $\frac{148}{150}$ =98.67% , 오류율1 : $\frac{2}{150}$ = 1.33% ; 정확률2 : $\frac{147}{150}$ =98% , 오류율2 : $\frac{3}{150}$ = 2%

Console C:/RSources/

```
> f_pred = predict(f, iris)
> confusionMatrix(f_pred, iris$Species)
Confusion Matrix and Statistics

            Reference
Prediction    setosa versicolor virginica
  setosa          50          0         0
  versicolor       0         50         2
  virginica        0          0        48

Overall Statistics

               Accuracy : 0.9867
                 95% CI : (0.9527, 0.9984)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.98

 Mcnemar's Test P-Value : NA
```

ces/

```
rest(Species~., data = iris, ntree = 3)    # 랜덤 포리스트 모델
dict(f, iris)
rix(f_pred, iris$Species)
ix and Statistics

ference
etosa versicolor virginica
   50          0         0
    0         48         1
    0          2        49

tics

 Accuracy : 0.98
   95% CI : (0.9427, 0.9959)
tion Rate : 0.3333
tc > NIR] : < 2.2e-16

    Kappa : 0.97
```

10.2 정확률

결정 트리 사례

- 전체 Sample 수 : 150, 정답 수 : n_1= 50+49+45 = 144, 오답 수 : n_2 = 0+1+5=6
- 정확률 : $\frac{144}{150}$ =96% , 오류율 : $\frac{6}{150}$ = 4%

랜덤 포스트 사례

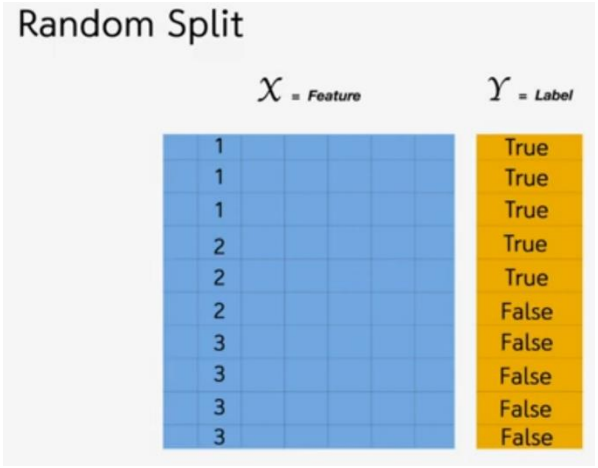
- (1)전체 Sample 수 : 150, 정답 수 : n_1= 50+50+48 = 148, 오답 수 : n_2 = 0+0+2=2
- (2)전체 Sample 수 : 150, 정답 수 : n_1= 50+48+49 = 147, 오답 수 : n_2 = 0+2+1=3
- 정확률1 : $\frac{148}{150}$ =98.67% , 오류율1 : $\frac{2}{150}$ = 1.33% ; 정확률2 : $\frac{147}{150}$ =98% , 오류율2 : $\frac{3}{150}$ = 2%

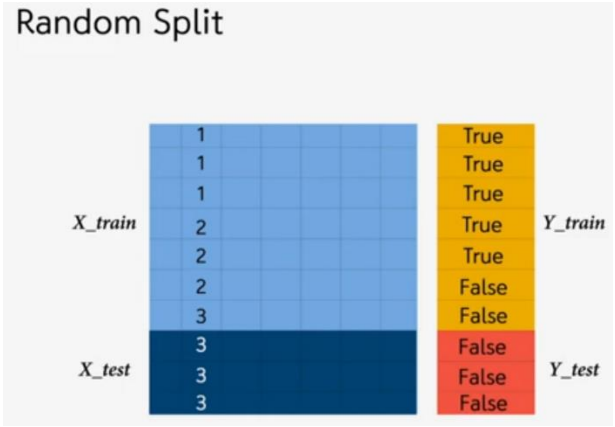
따라서 iris 데이터를 위한 모델로는 랜덤 포스트를 선택한다.

10.2 정확률

랜덤 포스트

data split (70 : 30)





label(정답) : 합격과 불합격, 이번 예는 타이타닉호 침몰 머신러닝 모델

Thank you


Africa
Americas
Asia
Europe
Oceania
pop
2.50e+08
5.00e+08
7.50e+08
1.00e+09
1.25e+09
gdpPercap