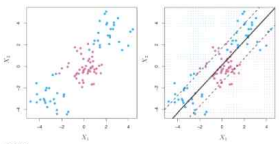
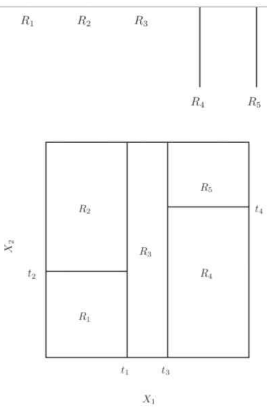
27. [9장] 서포트 벡터 분류기는 2개의 클래스 분류와 클래스 간의 경계가 선형인 경우에 적용된다.
그러나 아래의 그림과 같이 2개 클래스의 경계가 선형이 아닐 경우, 서포트 벡터 분류기를 적용하여 사용할 수 없다. 이러한 비선형 결정 경계에 적용하기 위해 서포트 벡터 분류기의 개념을 확장한 분류기의 이름은?
(단답형이며, 답에 띄어쓰기하지 말것)



(10점)

※해당 문제는 복사/붙여넣기 기능을 사용할 수 없습니다.

< 이전 다음 >



(10점)

□1) R_3이 더 이상 분할되지 않은 이유는 R_3 내부의 점의 개수가 특정 값보다 작거나 같기 때문이다.

□2) 가장 먼저 분할된 공간은 $X_1 \leq t_1$와 $X_1 > t_1$ 이다.

□3) 전체의 RSS를 최소화 하는데 두번째로 중요한 요소는 t_2이다.

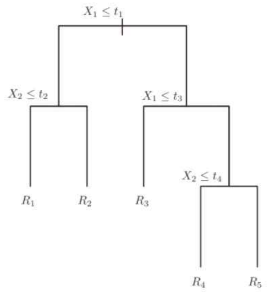
□4) 가장 마지막으로 분할된 공간은 R_4와 R_5이다.

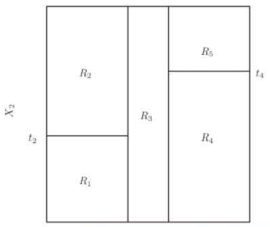
□5) 전체의 RSS를 최소화 하는데 가장 중요한 결정 요소는 t_1이다.

< 이전 다음 >

26. [8장] 다음의 그림은 재귀 이진 분할 알고리즘을 이용하여 구성된 결정 트리이다.

이 결정 트리에 대한 설명으로 틀린 것을 모두 고르시오.





25. [6장] 다음의 능형회귀(ridge regression)와 최소 제곱에 대한 비교 설명 중 옳은 것을 모두 고르시오. (10점)

☐1) 능형회귀는 최소 제곱에 비해 유연성이 높기 때문에, 분산의 증가가 편향 감소보다 작을 경우, 예측의 정확도가 향상된다.

☐2) 능형회귀는 최소 제곱에 비해 유연성이 낮기 때문에, 편향의 증가가 분산의 감소보다 작을 경우 예측의 정확도가 향상된다.

☐3) 능형회귀는 최소 제곱에 비해 유연성이 낮기 때문에, 분산의 증가가 편향의 감소보다 작을 경우 예측의 정확도가 향상된다.

☐4) 능형회귀는 최소 제곱에 비해 유연성이 높기 때문에, 편향의 증가가 분산의 감소보다 작을 경우, 예측의 정확도가 향상된다.

< 이전 다음 >

24. [9장] 다음의 보기들은 최대 마진 분류기에 대한 설명들이다. 옳은 것들을 모두 선택하시오.
(10점)

☐1) 최대 마진 분류기의 결정 경계는 소수의 관측 데이터들로만 결정된다.

☐2) 최대 마진 분류기의 결정 경계는 모든 훈련 데이터를 기반으로 결정된다.

☐3) 관측된 훈련 데이터와 조평면 사이의 최소 거리를 마진이라 할 때, 최대 마진 분류기는 이 마진이 가장 큰 조평면을 결정 경계로 정한다.

☐4) 조평면을 사용하여 관측된 훈련 데이터가 완벽하게 분류되는 경우에만 적용 가능하다.

☐5) 최대 마진 분류기는 관측된 훈련데이터들의 RSS 합의 최소 제곱을 제공하는 조평면을 결정 경계로 정한다.

< 이전 다음 >

23. [6장] 아래의 수식에서 특정 값의 s에 대해 이 식을 최소화하여 선형회귀모델의 회귀 계수를 추정하려고 한다. s 값을 0에서 부터 증가시킴에 따라, 다음 보기의 설명 중 맞는 것을 모두 선택하시오.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \text{ subject to } \sum_{j=1}^{p}|\beta_j| \leq s$$

(10점)

☐1) 검정 RSS는 계속해서 감소한다.

☐2) 검정 RSS는 처음에는 증가하다가 결국 거꾸로 된 U자 형태로 감소하기 시작한다.

☐3) 검정 RSS는 계속해서 증가한다.

☐4) 검정 RSS는 처음에는 감소하다가 결국 U자 형태로 증가하기 시작한다.

☐5) 검정 RSS는 일정하게 유지된다.

< 이전 다음 >

22. [8장] 결정트리는 관측된 훈련 데이터에 쉽게 영향을 받기 때문에 기존의 회귀 모델에 대해 예측의 정확도를 떨어지는 문제점이 있다. 이를 해결하기 위해, 배깅을 활용한 다수의 결정 트리들을 구성하여 예측모델의 정확도를 향상시키고자 한다. 다음의 보기에서 배깅에 대해 옳은 설명들을 모두 고르시오.
(10점)

☐1) 배깅에서 양적 반응 변수의 예측은 각 결정트리들 예측값의 평균값을 취한다.

☐2) 배깅 트리들을 모델링은 결정트리를 기반으로 하기 때문에 이해, 해석, 설명이 용이하다.

☐3) 배깅에서는 하나의 관측된 훈련 데이터를 브트스트랩하여 많은 수의 훈련셋을 생성한다.

☐4) 배깅에서 질적 반응 변수의 예측은 각 분류트리들의 분류 결과의 다수결을 선택한다.

☐5) 하나의 결정트리는 높은 분산과 낮은 편향을 가지므로 배깅을 통해 분산을 축소한다.

< 이전 다음 >

/test2_question_form.acl#link_1

21. [7장] 다음의 수식으로 표현되는 계단 함수는 변수의 범위를 K개 영역으로 구분하여 질적 변수를 생성하여 조각별 상수 함수를 적합하게 된다.

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \ldots + \beta_K C_K(x_i) + \epsilon_i$$

다음의 계단함수의 특성 설명 중 틀린 것을 모두 선택하라.
(10점)

☐1) 구간의 개수 K가 작을 수록 검정 MSE가 감소한다.

☐2) 구간의 개수 K가 클 수록 훈련 MSE가 감소한다.

☐3) 계단 함수 모델의 각 변수들의 계수들은 각 구간에 속하는 반응변수들의 평균값이다.

☐4) 구간의 개수 K가 클수록 검정 MSE가 증가한다.

☐5) 구간의 개수 K 가 작을 수록 훈련 MSE가 증가한다.

< 이전 다음 >

/test2_question_form.acl#link_1

20. [6점] 선형모델은 최소제곱오류를 통해 반응변수 Y와 설명 변수 X1, .., Xp의 상관 관계를 모델링한다. 이때 이 모델의 정확도는 관측 데이터의 개수 N과 설명 변수의 개수 p의 관계에 많은 영향을 받게 된다. 다음의 보기 중 그 설명이 틀린 것을 모두 고르시오.
(10점)

☐1) $N \simeq p$ 인 경우, 최소 제곱 추정치 $\beta_0, \beta_1, \ldots, \beta_p$의 분산은 작고 검정 관측치에 대한 예측의 정확도는 낮다.

☐2) N >> p인 경우, 최소 제곱 추정치, $\beta_0, \beta_1, \ldots, \beta_p$들의 분산은 크고, 검정 관측치에 대한 예측의 정확도는 높다.

☐3) N < p인 경우, 최소 제곱 추정치 $\beta_0, \beta_1, \ldots, \beta_p$들의 유일한 값이 존재하지 않아, 선형모델을 하나로 특정할 수 없다.

☐4) $N \simeq p$ 인 경우, 최소 제곱 추정치 $\beta_0, \beta_1, \ldots, \beta_p$들의 분산은 크고, 검정 관측치에 대한 예측의 정확도는 낮다.

☐5) N >> p인 경우, 최소 제곱 추정치, $\beta_0, \beta_1, \ldots, \beta_p$들의 분산은 작고, 검정 관측치에 대한 예측의 정확도는 높다.

< 이전 다음 >

2_question_form.acl#link_1

19. [7점] 국소 회귀를 유연한 비선형 함수들은 적합하는 또 다른 기법으로 목표점 x_0에서 그 주변의 훈련 관측치들만을 이용하여 적합한다. 아래의 알고리즘은 $X = x_0$에서의 국소 회귀 알고리즘을 설명하고 있다. 다음 보기의 설명 중 이 알고리즘에 대한 설명으로 틀린 것을 모두 선택하시오.

1. 훈련 포인트들의 x_i가 x_0에 가장 가까운 일부 $s = k/n$을 모은다.
2. 이 이웃의 각 점에 가중치 $K_{i0} = K(x_i, x_0)$을 할당한다. x_0에서 가장 먼 점은 가중치가 영이고 가장 가까운 점은 가장 높은 가중치를 가진다. k개의 최근접이웃 이외의 모든 점은 가중치가 영이다.
3. 앞의 가중치를 사용하여 식 (7.14)를 최소로 하는 $\hat{\beta}_0$와 $\hat{\beta}_1$을 찾음으로써 x_i에 y_i의 가중 최소 제곱회귀를 적합한다

$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2 \qquad (7.14)$$

4. x_0에서 적합된 값은 $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$로 주어진다.

(10점)

☐1) s값이 작을수록 훈련 MSE가 작아진다.

☐2) s값이 작을수록 훈련 MSE가 커진다.

☐3) s의 값이 작을수록 검정 MSE가 작아진다.

☐4) s는 가까운 이웃 점들의 비율을 의미한다.

☐5) s값이 커질수록 검정 MSE가 커진다.

< 이전 다음 >

est2_question_form.acl#link_1

18. [8장] 다음의 보기 중 결정트리의 장점에 해당되는 보기를 모두 선택하라.
(10점)

☐1) 결정 트리의 예측 정확도는 기존의 회귀 및 분류 모델보다 높다.
☐2) 결정 트리는 비선형적 관측 데이터에 적합하다.
☐3) 결정 트리는 질적 설명 변수를 더미(dummy) 변수를 사용하지 않고도 쉽게 모델에 포함할 수 있다.
☐4) 결정트리의 예측과정은 사람의 의사 결정 과정과 유사하여 선형회귀에 비해 설명하기 쉽다.
☐5) 결정 트리는 선형적 관측 데이터에 적합하다.

< 이전 다음 >

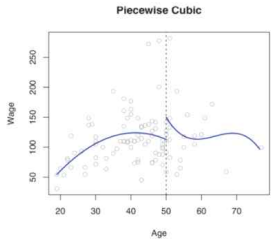
utest2_question_from_artlink_1

17. [7장] 다음의 기법들 중 선형모델의 해석력을 가능한 유지하면서 예측의 정확도를 향상시키기 위한 비선형적 확장성의 특징을 가지는 모델이 아닌 것은?
(10점)

○1) 평활스플라인
○2) 회귀스플라인
○3) 주성분회귀
○4) 다항식회귀
○5) 일반화가법모델

< 이전 다음 >

16. [7장] 아래의 그림은 조각별 3차 다항식 회귀 스플라인의 문제점을 보여준다.



이 문제를 해결하기 위해 1개의 매듭을 가지는 3차 스플라인 함수가 아래와 같이 정의되었다.

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)_+^3$$

$$(x-\xi)_+^3 = \begin{cases} (x-\xi)^3 & \text{if } x > \xi \\ 0 & \text{otherwise} \end{cases}$$

관측 데이터 집합에 대해, 위의 3차 스플라인 함수의 β_j들을 찾아내기 위한 적합을 수행하는 조건 또는 설명으로 틀린 것은 모두 선택하시오.

(10점)

☐1) 이 함수의 자유도는 5이다.

☐2) 매듭 ξ 에서 이 함수의 1차 도함수는 연속이나 2차 도함수는 연속일 필요 없다.

☐3) 매듭 ξ에서 이 함수는 불연속이다.

☐4) 매듭 ξ 에서 이 함수의 1차 도함수는 연속일 필요가 없으나, 2차 도함수는 연속이여야 한다.

☐5) 매듭 ξ 에서 이 함수는 연속이이여야 한다.

< 이전 다음 >

15. [6장] 다음의 Lasso와 최소 제곱에 대한 비교 설명 중 옳은 것을 모두 고르시오.
(10점)

○1) Lasso는 최소 제곱에 비해 유연성이 높기 때문에, 분산의 증가가 편향 감소보다 작을 경우, 예측의 정확도가 향상된다.

○2) Lasso는 최소 제곱에 비해 유연성이 낮기 때문에, 편향의 증가가 분산의 감소보다 작을 경우 예측의 정확도가 향상된다.

○3) Lasso는 최소 제곱에 비해 유연성이 낮기 때문에, 분산의 증가가 편향의 감소보다 작을 경우 예측의 정확도가 향상된다.

○4) Lasso는 최소 제곱에 비해 유연성이 높기 때문에, 편향의 증가가 분산의 감소보다 작을 경우, 예측의 정확도가 향상된다.

< 이전 다음 >

이름 : 박병준

남은 시간 : 22:26

14. [6장] 아래의 수식에서 특정 값의 λ에 대해 이 식을 최소화하여 선형회귀모델의 회귀 계수를 추정하려고 한다. λ 값을 0에서 부터 증가시킴에 따라, 다음 보기의 설명 중 맞는 것을 모두 선택하시오.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

(10점)

☐1) λ값을 0에서 증가시킴에 따라, 훈련 RSS는 계속해서 증가한다.

☐2) λ값을 0에서 증가시킴에 따라, 훈련 RSS는 처음에는 증가하다가 결국 거꾸로 된 U자 형태로 감소하기 시작한다.

☐3) λ값을 0에서 증가시킴에 따라, 훈련 RSS는 처음에는 감소하다가 결국 U자 형태로 증가하기 시작한다.

☐4) λ값을 0에서 증가시킴에 따라, 훈련 RSS는 계속해서 감소한다.

☐5) λ값을 0에서 증가시킴에 따라, 훈련 RSS는 일정하게 유지된다.

< 이전 다음 >

test2_question_form.acl#link_1

13. [8장] 다음의 보기들은 배깅 모델의 문제점과 랜덤 포레스트(Random forest)에 대한 설명들이다. 바르게 설명한 것들을 모두 선택하시오.
(10점)

☐1) 랜덤 포레스트는 배깅 트리들 간의 상관성을 제거하기 위해 p개의 설명 변수 중, m개의 설명 변수들을 랜덤하게 선택하여 분할 후보로 사용한다.

☐2) 배깅 트리들의 높은 상관성이 있는 값들을 평균하는 것은 상관되지 않은 값들을 평균하는 것 만큼 크게 분산을 줄일 수 없다.

☐3) 배깅이 단일 트리에 비해 예측값의 분산을 크게 줄여 예측의 정확도를 향상하였다.

☐4) 배깅된 트리들은 모두 서로 상당히 유사하여 배깅된 트리들에서 얻은 관측치들은 서로 높게 상관되어 있다.

☐5) 배깅된 결정 트리들은 대부분 또는 모든 트리들이 상관성이 강한 설명 변수를 맨 위의 분할(top split)에서 사용한다.

< 이전 다음 >

test2_question_form.acl#link_1

12. [7장] 일반화 가법 모델 (GAM)은 여러 개의 설명 변수 x_i를 기반으로 반응 변수 Y를 유연하게 예측하기 위해 다중 선형회귀를 확장한 것이다. 이 모델은 가산성은 유지하면서 각 설명 변수들이 반응 변수 Y에 대한 기여분을 비선형 함수들로 모델링하여 표준 선형 모델을 확장하는 일반적인 체계를 제공한다. 다음의 보기는 GAM에 대한 장점을 설명한 것인다. 옳은 것을 모두 고르시오.
(10점)

☐1) GAM의 결과가 정확하기 위해서는 각 설명 변수간의 상호작용이 없어야 한다.

☐2) GAM에서 허용되는 각 설명 변수 x_i의 종류는 양적변수이거나 질적 변수 중 한가지만 허용된다.

☐3) GAM은 각각의 설명 변수 x_i에 대해 반응 변수 Y에 대한 기여분을 비선형 함수로 모델링하기 때문에 선형회귀에 적합하지 않은 비선형 관계를 표현할 수 있다.

☐4) GAM은 가산적이기 때문에 반응 변수 Y에 대한 각 설명 변수 x_i의 영향을 개별적으로 표현하므로 추론에 용이하다.

< 이전 다음 >

11. [8장] 다음의 보기들은 부스팅과 배깅의 차이를 설명하고 있다. 바르게 설명된 것들을 모두 고르시오.
(10점)

☐1) 부스팅에서 트리의 수 B가 크면 훈련 MSE가 감소한다.

☐2) 수축 파라메터 λ는 부스팅의 학습 속도를 제어한다. λ가 작은 값일 수록 좋은 성능을 달성하기 위해 작은 값의 B(트리의 수)를 사용한다.

☐3) 각 트리의 분할 수 d는 부스팅 구성의 복잡도 제어한다. 보통 d = 1이면, 부스팅 구성에서 각 항이 하나의 변수만 포함한다.

☐4) 부스팅은 현재의 잔차들을 반응 변수로 사용하여 주어진 의사 결정 트리를 적합한다.

☐5) 배깅으로 생성되는 트리들은 상호 간에 의존적이지 않으나, 부스팅의 각 트리는 이전에 만들어진 트리들을 기반으로 생성된다.

< 이전 다음 >

10. [8장] 질적 반응 변수에 대해 분류트리를 구성할 경우, RSS를 이진 분할의 기준값으로 사용할 수 없어 분류 오차율과 같은 기준값을 사용한다. 그러나 심장병 진단과 같이 분류의 정확도 보다 분류의 신뢰도가 더 중요할 경우, 단말 노드에 대한 순도를 기준으로 이진분할을 수행해야 한다. 이 때 사용되는 기준값을 무엇이라 하나? 이 기준값이 0에 가까운 작은 경우, 해당 단말 노드는 주로 단일 클래스의 관측치를 포함하게 된다.

(10점)

※ 해당 문제는 복사/붙여넣기 기능을 사용할 수 없습니다.

< 이전 다음 >

9. [7장] 평활 스플라인 모델링의 목적식은 아래와 같이 정의된다. 평활 스플라인 모델링에 대한 설명으로 틀린 것을 모두 선택하시오.

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

(10점)

- ☐1) $\lambda = \infty$ 일 때, 함수 $g(x_i)$는 훈련 데이터들의 점을 가능한 가깝게 통과하는 직선이다.
- ☐2) $\lambda = 0$ 일 때, 함수 $g(x_i)$는 y_i를 완벽하게 적합한다.
- ☐3) λ는 평활 스플라인의 편향-분산을 절충하는 파라메터이다.
- ☐4) λ가 0에서 ∞로 증가함에 따라 평활 스플라인의 유효 자유도는 n에서 2로 감소
- ☐5) λ가 적절한 값일 때, 평활 스플라인의 결과는 회귀 스플라인과 동일하다.

< 이전 다음 >

7. [6장] 아래의 수식에서 특정 값의 λ에 대해 이 식을 최소화하여 선형회귀모델의 회귀 계수를 추정하려고 한다. λ값을 0에서 부터 증가시킴에 따라, 다음 보기의 설명 중 맞는 것을 모두 선택하시오.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

(10점)

☐1) λ값을 0에서 증가시킴에 따라, 검정 RSS는 계속해서 감소한다.

☐2) λ값을 0에서 증가시킴에 따라, 검정 RSS는 계속해서 증가한다.

☐3) λ값을 0에서 증가시킴에 따라, 검정 RSS는 처음에는 감소하다가 결국 U자 형태로 증가하기 시작한다.

☐4) λ값을 0에서 증가시킴에 따라, 검정 RSS는 처음에는 증가하다가 결국 거꾸로 된 U자 형태로 감소하기 시작한다.

☐5) λ값을 0에서 증가시킴에 따라, 검정 RSS는 일정하게 유지된다.

< 이전 다음 >

/test2_question_form.acl#link_1

6. [7장] 설명 변수 X의 전체 범위에 걸쳐 고차원 다항식을 적합하는 대신, X의 범위를 K개로 구분하여 각 범위에서 저차원의 다항식을 적합하는 것이 조각 별 다항식 회귀이며 회귀 스플라인의 여러 형태 중 하나이다. 다음의 설명 중에서 조각 별 다항식 회귀에 대한 설명 중 틀린 것을 모두 고르시오.
(10점)

☐1) 구간의 개수 K가 클 수록 훈련 MSE가 감소한다.

☐2) 구간의 개수 K가 작을 수록 훈련 MSE가 감소한다.

☐3) 구간의 개수 K가 클수록 각 구간에 낮은 차수의 수식을 이용하여 회귀하는 것이 유리하다.

☐4) 구간의 개수 K가 클수록 각 구간에 높은 차수의 수식을 이용하여 회귀하는 것이 유리하다.

☐5) 구간의 개수 K가 클수록 검정 MSE가 증가한다.

< 이전 다음 >

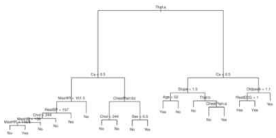
5. [6장] 관찰 데이터 N개에 대해, p개의 설명 변수 중, 가장 중요한 변수들의 집합을 선택하기 위한 방법으로 최상의 서브셋 선택, 전진 단계적 선택, 후진 단계적 선택을 모두 수행한 뒤, 각 방법을 비교한 설명이다. 틀린 것을 모두 선택하시오.
(10점)

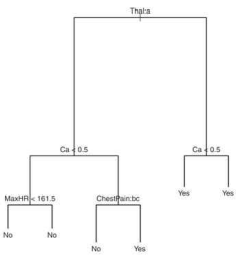
- [x] 1) 전진 단계적으로 얻어진 k개의 변수 모델의 설명 변수들은 후진 단계적 선택법에 의한 (k+1)개의 설명 변수 집합의 부분집합니다.
- [ ] 2) k개의 설명 변수를 갖는 세 모델 중 최상의 서브셋 선택이 가장 작은 훈련 RSS를 가진다.
- [ ] 3) 전진 단계적으로 얻어낸 k 개 변수 모델의 설명 변수들은 전진 단계적 선택법에 의해 선택된 (k+1)개의 설명 변수들의 부분 집합이다.
- [x] 4) 후진 단계적으로 얻어진 k개 변수 모델의 설명 변수들은 후진 단계적 선택법에 의한 (k+1)개의 설명 변수들의 부분 집합이다.
- [ ] 5) k개의 설명 변수를 갖는 세 모델 중 최상의 서브셋 선택이 가장 작은 검정 RSS를 가진다.

< 이전 다음 >

test2_question_form.acl#link_1

4. [8장] 다음의 그림은 Heart 자료에 대해 가지치기(pruning)을 하지 않은 분류트리와 가지치기를 한 분류트리 그림을 비교한 것이다. 이와 관련된 다음의 설명 중 옳은 것을 모두 선택하시오.

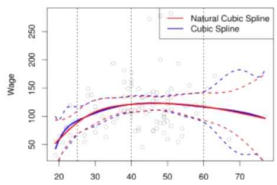




(10점)

- [ ] 1) 트리의 크기가 작아질 수록 훈련 MSE는 증가한다.
- [x] 2) 트리의 크기가 작아질 수록 결정트리 모델의 편향이 증가한다.
- [x] 3) 트리의 크기가 증가할 수록 훈련 MSE는 감소한다.
- [ ] 4) 트리의 크기가 작아질 수록 결정트리 모델의 분산이 작아진다.
- [ ] 5) 트리의 크기가 증가할 수록 검정 MSE는 증가한다.

3. [7장] 아래의 그림은 3차 회귀 스플라인과 자연 회귀 스플라인 차이를 보여주는 그림이다.



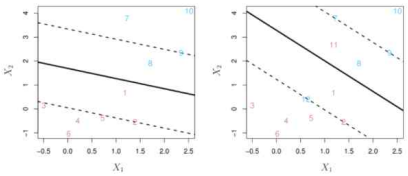
설명 변수의 외측 범위(x가 매우 작거나 매우 큰 값)에서 3차 스플라인이 가지는 문제점을 해결하기 위해, 자연 스플라인은 설명 변수의 외측범위(x <= 25, x >= 60)에서 몇 차 함수를 이용하여 모델링하는가? 차수를 숫자로 적으시오

(10점)

※ 해당 문제는 복사/붙여넣기 기능을 사용할 수 없습니다.

< 이전 다음 >

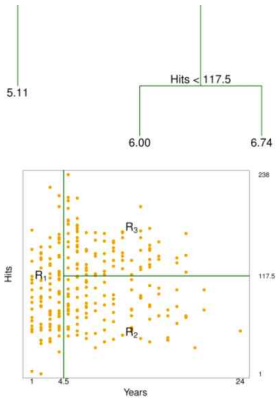
2. [9장] 최대 마진 분류기는 주어진 관측 데이터가 완벽하게 분류될 수 있는 경우에 적용가능하다. 그러나 실제 관측데이터는 완벽하게 분류되지 않는다. (아래의 그림 참조) 이런 경우의 문제를 해결하기 위해 서포트 벡터 분류기가 제안되었고, 이 분류기를 찾기 위한 문제는 아래의 최적화 문제로 기술된다.
이 최적화 문제의 파라메터 C에 대한 설명으로 옳은 것을 모두 선택하시오.



$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \; M \quad \text{subject to} \; \sum_{j=1}^{p} \beta_j^2 = 1,$$
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$
$$\epsilon_i \geq 0, \; \sum_{i=1}^{n} \epsilon_i \leq C,$$

(10점)

- [x] 1) C가 클 경우 넓은 마진을 선택되어 더 많은 넓은 마진 위반을 허용하여, 데이터에 덜 엄격하게 적합하고 편향은 더 높지만 분산이 더 낮은 분류기가 된다.
- [ ] 2) C가 줄어 들면 마진 위반 허용 정도가 커져, 마진의 폭이 감소한다.
- [x] 3) C가 증가함에 따라 마진 위반 허용 정도가 작아져 마진의 폭이 증가한다.
- [ ] 4) C는 마진에 대한 허용될 위반의 수와 그 정도를 결정한다.
- [ ] 5) C가 작을 때는 좁은 마진을 선택되어 훈련 관측치에 적합된 분류기가 되어, 편향은 낮지만 분산이 커진다.



(10점)

☐1) 주어진 경력 연수(Years)와 안타수(Hits)를 기반으로하는 이 트리 모델을 통해 연봉을 정확히 예측할 수 있다.

☐2) 설명 변수 Hits가 연봉을 예측하는데 가장 중요한 요소이다.

☐3) 설명 변수 Years 가 연봉을 예측하는데 가장 중요한 요소이다.

☐4) 터미널 노드의 숫자들은 각 영역에 속하는 연봉 데이터의 평균 값이다.

☐5) 이 트리 모델을 통해 연봉을 결정하는 중요 요소들을 쉽게 이해할 수 있다.

< 이전 다음 >

1. [8장] 아래의 2개 그림은 Hitters 연봉 자료에 대한 결정 트리와 이 결정 트리에 따른 설명 변수들의 영역 분할를 보여준다. 이 결정 트리에 대한 설명으로 틀린 것을 모두 고르시오.