CH 09. 서포트 벡터 머신

- [9.1] 최대 마진 분류기: "선형" 경계라는 한계점
- [9.2] 서포트 벡터 분류기: 최대 마진 분류기의 확장
- [9.3] 서포트 벡터 머신: 비선형 경계 수용, 2진 분류
- [9.4] 서포트 벡터 머신의 확장: 클래스 수가 3개 이상
- [9.5] 다른 통계 방법 사이의 관련성

9.1 최대 마진 분류기

> 1. 초평면(hyperplane)
> 2. 분리 초평면(separating hyperplane)
> 3. 최대 마진 분류기

9.1.1 초평면은 무엇인가?

- **초평면(hyperplane)**: p차원 공간에서 차원이 p-1인 평평한 아핀(affine) 부분공간
  - 2차원에서는 선, 3차원에서는 평면
  - 부분공간(subspace): n차원 공간에 포함되면서, n차원 벡터들에 대해 선형 변환이 성립하는 공간
  - subspace = linear manifold, flat, affine space = w/ "affine" property
  - 아핀(affine)이란 부분공간이 원점을 지날 필요가 없다는 것을 의미한다.
- 초평면의 정의
  - $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0$
  - 초평면을 "정의한다" ?
    - X는 p차원 공간의 점 또는 길이가 p인 벡터이다.
    - 식이 성립하는 임의의 $X = (X_1, X_2, \ldots, X_p)^T$ 는 초평면 상의 점이다.
  - X가 식을 만족하지 않는다면?
    - 식의 부호에 따라 초평면의 어느 쪽에 있는지를 결정할 수 있다.
    - 즉, 초평면은 p차원 공간을 두 개의 부분으로 이등분한다.

9.1.2 분리 초평면(Separating Hyperplane)을 사용한 분류

- 분류기를 개발하자.
  - p차원 공간에서 n개의 훈련 관측치로 구성되는 n x p 데이터 행렬 X가 있다.
  - 이들 관측치는 두 개의 클래스에 포함된다. 즉, $y_1, \ldots, y_n \in (-1, 1)$이다.
  - 훈련 데이터를 기반으로 검정 관측치인 p-벡터를 분류하자.
  - 4장의 LDA, 로지스틱 회귀와 8장의 트리, 배깅, 부스팅이 그러한 기법들이다.
- **분리 초평면**: 관측치들을 클래스 라벨에 따라 분리하는 초평면
- $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$ 값에 따라 $y_i = 1$ 또는 $y_i = -1$이라는 라벨을 붙일 수 있다.

- $y_i = 1$이면 $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} > 0$
- $y_i = -1$이면 $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} < 0$
- 즉, 모든 $i = 1, \ldots, n$에 대하여 분리 초평면은 다음을 만족한다.

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) > 0$$

- 검정 관측치 x^*는 $f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \ldots + \beta_p x_p^*$를 기준으로 분류된다.
  - 부호: $f(x^*)$가 양수이면 클래스 "1"에, 음수이면 클래스 "-1"에 할당된다.
  - 크기: $f(x^*)$가 0에서 멀리 떨어져 있으면 확신이 크고, 가까우면 덜하다.
- 분리 초평면에 기초한 분류기는 선형결정경계를 낳는다.

9.1.3 최대 마진 분류기

- 분리 초평면을 기반으로 분류기를 구성하려면?
  - 초평면을 조금씩 이동하거나 회전할 수 있다는 점에서, 무한 개 초평면이 존재할 것이다.
  - 그렇다면 그 중 어느 것을 사용할지 결정하여야 한다.
- **최대 마진 초평면**
  - 훈련 관측치들로부터 가장 멀리 떨어진 분리 초평면을 선택하는 것이 자연스럽다.
  - 마진(margin): 관측치들에서 초평면까지의 가장 짧은 거리
  - 즉, 훈련 관측치들까지의 최소 거리가 가장 먼 초평면이다.
- **최대 마진 분류기**는 $f(x^*)$의 부호를 기반으로 검정 관측치 x^*를 분류한다.
- **서포트 벡터(support vector)**: 최대 마진 초평면에서 동일하게 마진만큼 떨어져 있는 관측치들
  - WHY support? 이 점들이 이동하면 최대 마진 초평면도 이동할 것이다.
  - 최대 마진 초평면은 서포트 벡터에만 직접적으로 의존적이고, 다른 관측치들에는 의존적이지 않다.

9.1.4 최대 마진 분류기의 구성

- 최대 마진 초평면은 다음 최적화 문제에 대한 해이다.

$$\underset{\beta_0, \beta_1, \ldots, \beta_p}{\text{maximize}}\, M \tag{9.9}$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1, \tag{9.10}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M \;\; \forall\, i = 1, \ldots, n. \tag{9.11}$$

- Each observation will be on the correct side of the hyperplane, provided that M is positive.
  - Actually, for each observation to be on the correct side, we would simply need > 0
  - So, (9.11) in fact requires that each observation to be on the correct side wih some cushion.
- (9.11): (M이 양수일 때) 각 관측치가 초평면의 올바른 쪽에 있도록 보장한다.
- (9.10): i번째 관측치에서 초평면까지의 수직거리가 (9.11)의 좌변과 같다.
  - 초평면은 그대로 초평면이기에 제약조건은 아니지만 "의미"를 더한다.

    - Ensure at least a distance M from the hyperplane.
- 즉, 이 최적화 문제는 초평면의 마진인 M을 최대로 하는 $\beta_0, \beta_1, \ldots, \beta_p$를 선택하는 것이다.

9.1.5 분류 불가능한 경우

- 분리 초평면이 존재하지 않는 경우라면 최대 마진 분류기도 없다.
- 그렇다면 분리 초평면의 개념을 확장해보자.
- 소프트 마진(soft margin)을 사용하여 클래스들을 거의(almost) 분류하는 초평면을 생각해볼 수 있다.

9.2 서포트 벡터 분류기

> 1. 서포트 벡터 분류기

9.2.1 서포트 벡터 분류기의 개요

- 최대 마진 분류기는 완벽할 수 없다.
    - 분리 초평면이 존재하지 않을 수 있으며, 존재해도 바람직하지 않을 수 있다.
    - 모든 훈련 관측치들을 완벽하게 분류한다면 과적합 문제가 발생한다.
    - 단일 관측치 변화에 극도로 민감하거나, 마진이 너무 작아 신뢰성이 떨어지는 등 문제의 소지가 있다.
- **서포트 벡터 분류기**: "소프트" 마진 분류기 = 마진이 soft = 일부 마진이 위반된다.
    - 개별 관측치에 대해 더 robust(둔감)하다.
    - 대부분의 훈련 관측치들을 더 잘 분류한다.
    - 몇몇 훈련 관측치들을 잘못 분류하더라도 나머지를 더 잘 분류할 수 있다면 의미가 있을 수 있다.
    - 따라서 일부 관측치들은 마진, 심지어는 초평면의 옳지 않은 쪽에 있을 수 있다.

9.2.2 서포트 벡터 분류기의 세부 사항

- 서포트 벡터 분류기는 다음 최적화 문제에 대한 해를 찾는다.

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M \tag{9.12}$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1, \tag{9.13}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i), \tag{9.14}$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C, \tag{9.15}$$

    - 마찬가지로 검정 관측치 x^*는 $f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \ldots + \beta_p x_p^*$를 기준으로 분류된다.
- 슬랙변수(slack variable): $\epsilon_1, \ldots, \epsilon_n$은 관측치들이 마진 또는 초평면의 옳지 않은 쪽에 있도록 허용한다.
    - 슬랙변수 ϵ_i는 i번째 관측치가 초평면과 마진에 관해 어디에 위치하는지 알려준다.
    - $\epsilon_i = 0$이면 i번째 관측치는 마진의 옳은 쪽에 있다.

    - $\epsilon_i > 0$이면 i번째 관측치는 마진의 옳지 않은 쪽에 있다. = 마진을 "위반"한다
    - $\epsilon_i > 1$이면 i번째 관측치는 초평면의 옳지 않은 쪽에 있다.
- 조율 파라미터 C: ϵ_i의 합을 한정하여 허용될 위반의 수와 정도를 결정한다.
    - C = 0이면 최대 마진 분류기의 최적화 문제와 동일해진다.
    - C > 0인 경우 C개 이하의 관측치들이 초평면의 옳지 않은 쪽에 있을 수 있다.($\epsilon_i > 1$)
    - 편향-분산 절충: C가 증가함에 따라 마진의 폭이 넓어질 것 = 편향은 높지만 분산이 낮음
- ***마진상에 높이거나 마진을 위반하는 관측치들만이 초평면에, 또 분류기에 영향을 준다!***
    - 즉, 서포트 벡터들만이 서포트 벡터 분류기에 영향을 준다.
        - C가 편향-분산 절충을 제어한다는 주장과 일맥상통, C가 클수록 "서포트 벡터"들이 많아지기 때문.
    - 초평면에서 멀리 떨어진 관측치들에는 robust하다.
        - LDA 분류기가 모든 관측치들의 평균 및 공분산 행렬을 사용하는 것과는 반대
        - 로지스틱 회귀가 결정경계에서 멀리 떨어진 관측치들에 대한 민감도가 매우 낮다는 것과 비슷

9.3 서포트 벡터 머신

> 1. 서포트 벡터 머신

9.3.1 비선형 결정경계를 가진 분류

- 서포트 벡터 분류기는 클래스가 2개이고 경계가 선형일 때 효과적이다.
    - 그러나 변수들 사이에 비선형적 관계가 있다면 이 분류기는 아무 쓸모가 없다.
- 그렇다면 더 높은 차수의 다항식 항, 상호작용 항을 가지고 무한히 확장시킬 수 있다.
    - 이때 복잡해지는 계산을 서포트 벡터 머신이 효율적인 방식으로 변환시켜준다.

9.3.2 서포트 벡터 머신

- **서포트 벡터 머신(SVM)**: 서포트 벡터 분류기의 확장
    - 비선형 경계를 수용하기 위해 "커널 기법"을 사용하여 변수공간을 확장한다.
    - 서포트 벡터 분류기 문제에 대한 해는 관측치들, 그 중에서도 서포트 벡터의 내적만이 관련된다.
        - $f(x) = \beta_0 + \sum_{i \in S} \alpha_i < x, x_i >$ where $< x, x_i >= \sum_{j=1}^{p} x_{ij} x_{i'j}$
        - 즉, 훈련 관측치가 서포트 벡터가 아니면 각 관측치의 계수 α_i는 0이다.
    - 여기서 내적을 일반화된 형태인 $K(x_i, x_{i'})$로 바꾼다고 해보자.
        - $f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i)$
        - K = kernel = 두 관측치들의 유사성을 수량화하는 함수
        - 단순히 (표준) 선형 커널이거나, 차수가 d인 다항식 커널이거나, 방사 커널일 수 있다.
- 방사 커널(radial kernel)

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2). \qquad (9.24)$$

    - 검정 관측치 x_i^*가 훈련 관측치 x_i로부터 유클리디안 거리가 크면, K는 아주 작은 값이 될 것이다.

    - 따라서 x^*에 대해 예측된 클래스 라벨에 아무 역할도 하지 않을 것이다.
    - 즉, 주변 관측치들만이 클래스 라벨에 영향을 준다는 점에서 방사 커널은 국소적(local)
- 원래 변수들의 함수들을 이용하는 대신 커널을 사용하는 장점?
    - nC2개의 쌍에 대해 단지 $K(x_i, x_{i'})$만 계산하면 된다.
    - 확장된 변수공간에서 명시적으로 계산하지 않고도 얻을 수 있다.
    - 변수공간이 명시적(implicit)이지 않고 차원이 무한한 경우라면, 계산 자체가 불가능하다.

9.3.3 심장질환 자료에 적용

- ROC 그래프 기준, LDA와 서포트 벡터 분류기 둘 다 잘 작동하지만 후자가 약간 더 낫다.
- 방사커널의 γ값이 커질수록 훈련오차율은 작아지지만, 검정 데이터에서도 반드시 성능이 더 좋진 않다.

9.4 클래스가 2개보다 많은 SVM

- SVM을 2진 분류에서 K-클래스의 경우로 확장한다.
- 가장 널리 사용되는 방법으로 일대일(one-versus-one) 기법과 일대전부(one-versus-all) 기법이 있다.

9.4.1 일대일 분류

- 일대일 또는 모든 쌍(all-pairs) 기법이라고 한다.
- K > 2 클래스가 있을 때, 한 검정 관측치에 대해 KC2개의 SVM을 구성하고 각각은 한 쌍의 클래스를 비교한다.
- 이 검정 관측치가 각각의 클래스($k = 1, \ldots, K$)에 할당되는 횟수를 기록한다.
- 마지막으로 가장 자주 할당된 클래스에 할당함으로써 분류가 완료된다.

9.4.2 일대전부(One-Versus-All) 분류

- 매번 K개 클래스 중 하나를 나머지 K-1개 클래스와 비교한다.
- $\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk}$는 k번째 클래스를 다른 클래스들과 비교하는 SVM을 적합한 결과로 얻는 파라미터들이다.
- 검정 관측치 x^*는 $\beta_{0k} + \beta_{1k}x_1^* + \beta_{2k}x_2^* + \ldots + \beta_{pk}x_p^*$가 가장 큰 클래스에 할당된다.

9.5 로지스틱 회귀에 대한 상관관계

- SVM을 "손실 + 페널티" 형태로 나타낼 수 있다.

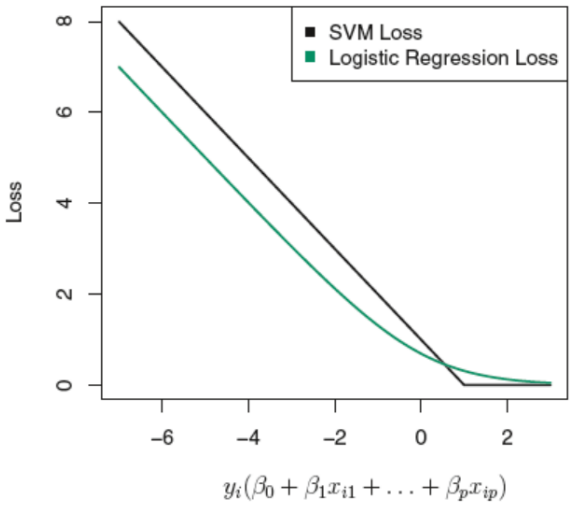
$$\underset{\beta_0, \beta_1, \ldots, \beta_p}{\text{minimize}} \left\{ \sum_{i=1}^{n} \max\left[0, 1 - y_i f(x_i)\right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \qquad (9.25)$$

$$\underset{\beta_0, \beta_1, \ldots, \beta_p}{\text{minimize}} \left\{ L(\mathbf{X}, \mathbf{y}, \beta) + \lambda P(\beta) \right\}. \qquad (9.26)$$

- 페널티항은 능형(ridge) 페널티항이고, 손실 함수는 힌지 손실(hinge loss)이다.

$$L(\mathbf{X}, \mathbf{y}, \beta) = \sum_{i=1}^{n} \max\left[0, 1 - y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})\right].$$

- 서포트 벡터 분류기의 손실 함수는 로지스틱 회귀의 경우와 상당히 유사하다.
  - 마진의 올바른 쪽 / 결정경계에서 멀리 떨어진 관측치들에 대한 손실함수의 값이 아주 작기 때문이다.
  - 그러나 전자가 0인 반면, 로지스틱은 0에 가까워지기 때문에 정확히 같진 않다.
  - 클래스들이 잘 분리되어 있으면 SVM, 겹치는 경우에는 로지스틱 회귀가 선호된다.



- 고전적 통계방법들과의 상관성
  - 조율 파라미터는 nuisance가 아니며, 편향-분산 절충 또는 과적합 여부를 결정한다.
  - 비선형 클래스 경계를 수용하기 위해 커널을 사용하는 것은 "특별하지" 않다.
  - 서포트 벡터 회귀: 마진 개념을 회귀로 확장시켜, 일정값보다 큰 잔차들만이 손실함수에 영향을 준다.