

워크북

교과목명 : 머신 러닝

차시명: 7차시

◆ 담당교수: 장 필 훈

● 세부목차

- soft margin SVM
- multiclass SVM
- SVM을 이용한 회귀
- 상관벡터머신
- 그래프모델
- 베이지안 네트워크
- 조건부독립

학습에 앞서

■ 학습개요

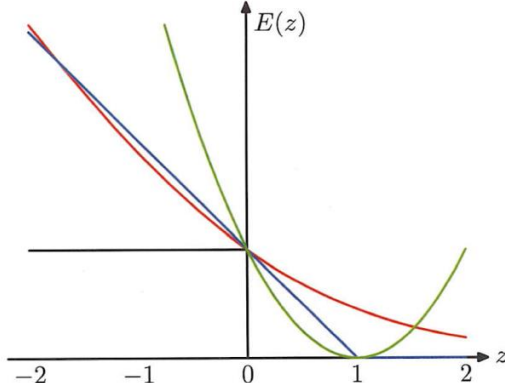
이전시간에 배웠던 SVM을 마무리한다. softmargin일때, 식을 전개해보고 로지스틱회귀와의 관계를 힌지오류 함수와 함께 이해한다. SVM의 단점과 함께, SVM의 베이지안 버전이라고 할수 있는 RVM에 대해 배우고 맨 첫시간에 배웠던 sine함수회귀에 응용한 결과를 본다.

그래프모델에서는 다른 모델을 그래프의 관점에서 이해할 수 있는 기초를 다진다. 그래프 자체로 계산하는 과정은 많지 않으므로, 조건부 독립등을 그래프로 어떻게 나타내는지 자세히 살펴본다. 수식으로 이해한 것을 그래프를 통해 더 직관적으로 이해할 수 있다.

■ 학습목표

1	SVM에서 softmargin을 허용할 때 어떻게 식이 전개되는지 이해한다.
2	RVM의 대략을 SVM과 비교하여 이해한다.
3	그래프를 나타내는 기본적인 방법을 이해한다.

■ 주요용어

용어	해설
slack variable	SVM의 기본적인 형태는 margin을 최대화 하는 것인데, 이는 선형분리 가능한것을 전제로 한다. 선형분리 불가능할 경우이거나 그렇지 않더라도 마진이 작은 경우, 어느정도 오분류를 허용하도록 SVM을 디자인할 수 있는데, 그 허용하는 정도를 나타내는 변수를 slack variable 이라고 하고 ξ 로 주로 나타낸다.
힌지오류함수	<p>결과값과 예측값의 차이를 나타내는 함수를 오류함수라고 하고, 지금까지는 제곱오류함수를 가장 많이 써왔다. 힙지오류함수는 식으로 나타내면 $[1 - y_n t_n]_+ = \max(1 - y_n t_n, 0)$와 같고, 그림으로 나타내면 아래그림의 파란색 선과 같다.</p>  <p style="text-align: right;">Bishop, Fig7.5</p>
상관벡터머신	SVM의 베이지언 버전이라고 생각할 수 있다. 출력값이 확률로 나오며, 다수의 클래스에 관해 확장 가능하고 커널의 조건에도 엄격한 제한이 없다. 입력벡터를 조건으로 타겟변수에 대한 조건부분포를 가정하고 식을 전개하여 가능도함수를 최대화 한다. SVM의 support vector에 해당하는 역할을 하는 벡터를 relevance vector라고 하기 때문에 상관벡터머신이라는 이름이 붙었다. SVM에 비해 훈련시간이 긴 단점이 있고, 서포트벡터와 달리 경계지점에 상관벡터들이 놓일 필요가 없다.
방향성그래프모델	그래프는 노드와 엣지로 이루어지는데 엣지가 방향성을 가지면 방향성 그래프 모델이라고 한다. 여기에 순환이 있느냐 없느냐로 다시 구분할 수 있고, 주로 순환이 없는 방향성 그래프를 많이 다룬다(DAG). 조건부 확률을 모델링할 수 있기 때문에 베이지언 네트워크라고 불리기도 한다.
조건부 독립	특정 변수가 주어진 상황에서 확률변수 a와 b가 독립이면 c가 주어진 상황에서 a는 b로부터 조건부 독립이라고 말할 수 있고 식으로는 $p(a b, c) = p(a c)$ 와같이 나타낼 수 있다.

(전 시간에 이어 계속)

soft margin의 제약조건은 hardmargin의 경우와 마찬가지로 $a_n(t_n y(x_n) - 1 + \xi_n) = 0$ 이고, 따라서 $t_n y(x_n) = 1 - \xi_n$ 입니다. 여기서부터는 다시 2차계획법을 사용해서 a 를 찾습니다.

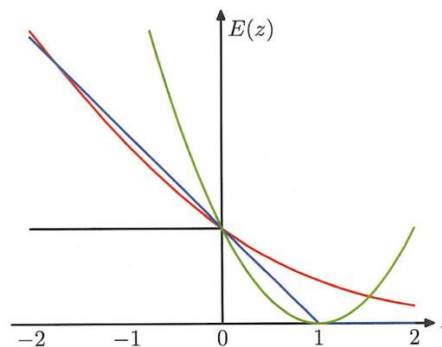
로지스틱회귀와의 관계를 보겠습니다. 소프트마진의 경우는 다음을 최소화 하는 것이 목표이고,

$$C \sum_{n=1}^N \xi_n + \frac{1}{2} \|\mathbf{w}\|^2$$

이것은 다음의 식과 같습니다.(상수배차이 무시)

$$\sum_{n=1}^N [1 - y_n t_n]_+ + \lambda \|\mathbf{w}\|^2$$

여기서 $[1 - y_n t_n]_+ = \max(1 - y_n t_n, 0)$ 이고 아래 그림에서 파란색 선입니다.



Bishop, Fig. 7.5

이 오류를 힌지오류라고 합니다. 로지스틱회귀의 경우 시그모이드 함수를 쓰고 그림에서는 빨간선입니다(상수배해서 힌지오류와 가장 비슷하게 맞춘 것입니다.) 빨간선과 파란선이 0에서 가까운 구간에서 일치 비슷한것을 볼 수 있습니다. 녹색선은 제곱오류입니다. 결정경계에서 멀리 떨어질수록 오류가 엄청나게 커지는 것을 볼 수 있습니다. 우리는 앞선 강의에서 제곱오류함수가 가지는 이런 성질때문에 오히려 분류기의 성능이 나빠지는것을 한번 보았습니다.

<다중클래스 SVM>

다중클래스의 분류는 앞서나왔던 것처럼 $y(x) = \max_k y_k(x)$ 를 이용합니다. 하지만 이 경우 여러 분류기들이 서로 다르게 훈련되었기 때문에 $y_k(x)$ 들의 절대치가 서로 비교가능하다는 보장이 없습니다. 그리고 다중클래스의 경우는 클래스의 n수가 서로 다를 경우 불균형에서 비롯되는 문제가 (분류기의 종류를 가리지 않고) 언제나 있습니다. (class imbalance문제)

<SVM을 이용한 회귀>

SVM을 이용해서 회귀를 할 수 도 있으나 실제로 자주 쓰이지는 않습니다. 기본 아이디어는 선형회귀와 동일하나 위에서 본것과 같이 오류함수를 교체한 후 식을 풉니다. 녹화강의에 조금 더 자세히 설명되어 있으니 참고 바랍니다.

<상관벡터머신>

SVM도 몇가지 한계점을 가지고 있습니다. 출력 값이 사후확률이 아니고 결정값이라는 점(이 점은 보기에 따라 분류기의 특성이고 단점이 아닐 수도 있습니다), 다중 class에 관해 확장이 어려운 점, 커널의 조건에 제약이 있다는 점 등입니다. 그 대안중 하나로 상관벡터머신이 있습니다. SVM과 비슷하고, 차이점이라면 SVM보다 더욱 희박한 모델을 준다는 점입니다.

RVM 회귀모델에서는 입력벡터 \mathbf{x} 에 대한 타겟변수 t 의 조건부 분포를 가정합니다.

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}), \beta^{-1})$$

여기서 $y(\mathbf{x})$ 는 기저함수들의 선형모델을 가정합니다. 기저함수를 커널로 준다면 구조가 SVM과 동일하게 됩니다.

N 개의 데이터를 관측했을 때 가능도 함수는 (조건부 분포에 따라) 다음과 같고,

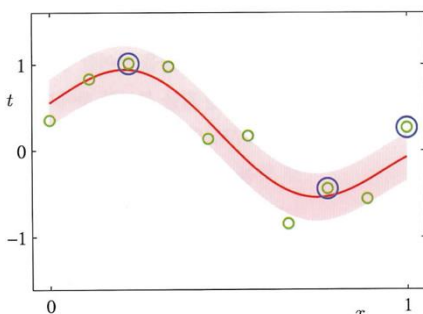
$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^N p(t_n|x_n, \mathbf{w}, \beta)$$

\mathbf{w} 에 대한 사전분포로 가우시안을 가정합니다. 이 가우시안은 평균이 0이고 분산을 매개변수로 가집니다. \mathbf{w} 는 벡터이므로 여러개의 변수로 이루어져 있고, 각 변수마다 새로운 매개변수가 도입됩니다.(이 매개변수를 α_i 로 두겠습니다). 그러면 \mathbf{w} 에 대한 가중치의 사전분포는 다음과 같게 됩니다.

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^M \mathcal{N}(w_i|0, \alpha_i^{-1})$$

이제 $\boldsymbol{\alpha}$ 와 β 를 근사하면 됩니다. 근사하는 과정은 중요도가 떨어진다고 판단되어 강의에서도 다루지 않았습니다. 더 깊이 공부하고 싶으신 분은 비숍책을 먼저 참고하시고, 추가적으로 필요하다면 비숍책에 있는 참고문헌부터 보시기를 추천합니다. 인터넷에 RVM관련해서는 자료가 많지 않습니다.

결과적으로 α_i 중 일부가 매우 큰 값을 가지고 이에 해당하는 w_i 들의 분산이 모두 0이 됩니다.(평균은 처음부터 0이였습니다) 그러면 이에 해당하는 기저함수는 모델에서 빠지게 됩니다. 0이 아닌 나머지 가중치에 해당하는 입력값들을 연관벡터(relevance vector)라고 부르고, SVM의 서포트벡터에 해당합니다. SVM과는 달리 RVM은 시각적으로 경계에 바로 놓이지 않습니다. 베이지안적 접근의 자연스러운 결과라고 할 수 있습니다. 아래 그림을 보면 회귀의 예와 연관벡터를 보여줍니다.



녹색점이 주어진 데이터, 검은 색 동그라미가 연관벡터, 원래 곡선(데이터를 추출한 곡선)이 빨간색입니다. 빨간색 음영은 표준편차 1만큼의 예측분포입니다.(베이지안 접근은 언제나 이렇게 범위로 결과가 나옵니다) [Bishop. Fig.7.9]

<그래프모델>

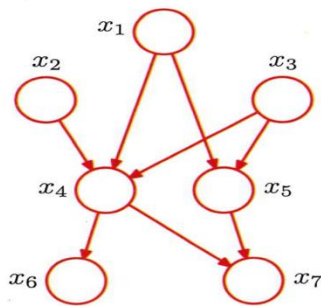
그래프는 노드와 링크로 이루어집니다. 확률변수와 연결해보면, 노드는 확률변수, 링크는 변수들간 확률적인 관계가 있음을 보여줍니다.

방향의 유무에 따라서도 다른데 방향이 있는 그래프의 경우 베이지안 네트워크가 대표적이고 비방향성 그래프모델은 마르코프 무작위장이 대표적입니다.

베이지안 네트워크를 그림으로 나타내 보겠습니다. 일단 결합분포를 다음과 같이 분해해 보겠습니다.

$$p(a, b, c) = p(c|a, b)p(a, b) = p(c|a, b)p(b|a)p(a)$$

분해의 방법이 유일하지 않음에 주의하세요. 위의 경우는 모든 노드쌍 사이에 연결이 있게 됩니다. 좀 더 복잡한 예를 하나 더 보겠습니다.



$$p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

하나하나 찬찬히 보면 쉽게 이해할 수 있습니다. 변수간 의존관계만 모두 표현해주면 됩니다. 보통 단말(terminal)노드들이 관측변수, ancestor쪽 노드들(위 그림에서 x_1, x_2, x_3)이 잠재변수에 해당합니다.

단 베이지안 네트워크의 경우 순환이 있으면 안됩니다.

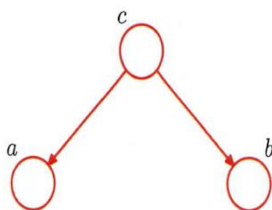
그럼 독립인 경우는 어떻게 나타낼까요? 중간 링크를 그리지 않으면 됩니다.

베이지안 네트워크의 경우 계산에 관해서는 녹화강의에 조금 더 자세히 다루었습니다.

<조건부독립>

그러면 그래프를 통해 조건부독립을 확인하는 방법을 알아보겠습니다. 이부분은 앞으로도 몇번 나오는 것이므로 자세히 봐주세요.

아래 그림은 $p(a, b, c) = p(a|c)p(b|c)p(c)$ 입니다.



Bishop 8.16

이때 아무런 변수도 관측되지 않은 상태라면

$p(a, b) = \sum_c p(a|c)p(b|c)p(c)$ 이고, 이것은 $p(a)p(b)$ 로 인수분해되지 않습니다. 따라서 조건부독립이 아닙니다. 하지만 만약 이 상태에서 c가 관측되었다면,

$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = p(a|c)p(b|c)$$

즉 조건부독립입니다.

tail-to-tail(화살표가 모두 아래를 향하고 있습니다)노드에서 c가 관측되면 a, b가 조건부 독립이라고 기억해두시면 됩니다.

그러면 head-to-tail노드는 어떨까요. (a) -> (c) -> (b) 형태라고 가정하고 c에 대해 주변화해보면,

$$p(a) \sum_c p(c|a)p(b|c) = p(a)p(b|a) \neq p(a)p(b)$$

따라서 조건부독립이 아닙니다. 이때도 다시 c가 관측되었다고 가정하면,

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} = p(a|c)p(b|c)$$

즉 조건부 독립입니다

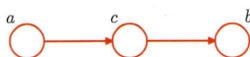
(a)->(c)<-(b)와 같은 형태는 어떨까요. $p(a,b,c)=p(a)p(b)p(c|a,b)$ 로 나타낼 수 있고, 양변을 c에 대해 주변화 하면 $p(a,b)=p(a)p(b)$ 를 얻습니다. 즉, 아무변수도 관측되지 않은 상태로 head-to-head는 조건부 독립이 성립합니다. a와 b는 서로 영향을 주지 않는다는 뜻입니다. 이경우는 c의 값이 주어지면 조건부 독립이 아니게 됩니다. 즉,

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(b)p(c|a,b)}{p(c)} \neq p(a|c)p(b|c)$$

입니다.

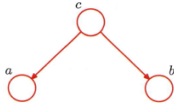
연습문제

- 선형분리 불가능한 데이터집합을 분류하는데는 SVM을 쓸 수 없다.
 - X
 - slack variable넣어도 되고, 커널을 사용하는 비선형 SVM을 사용할 수도 있다.
- RVM은 커널조건이 없다.
 - X
 - SVM에 비해 느슨한 것이지 커널조건이 아예 없을수는 없다.
- RVM에서 상관벡터도 SVM의 서포트벡터처럼 경계에서 구분의 기준이 되는 점으로 주어진다.
 - X
 - RVM의 상관벡터는 경계에서 주어질 필요가 없고, 그렇게 주어지지 않는 경향을 보인다
- 베이지안 네트워크를 그래프로 나타내기 위해서는 비방향성으로 나타내는 것이 기본이다.
 - X
 - 방향성으로 나타내야 사전분포를 명확히 나타낼수 있다.
- 아래와 같은 구조에서 C가 관측되었다면 a와 b는 조건부 독립이다.



- O
- $p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} = p(a|c)p(b|c)$ 이므로 관측되면 독립이다.

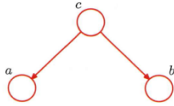
6. 아래와 같은 구조에서 C가 관측되었다면 a와 b는 조건부 독립이다.



a. O

b. $p(a, b|c) = \frac{p(a, b, c)}{p(c)} = p(a|c)p(b|c)$ 이므로 조건부 독립이다.

7. 아래와 같은 구조에서 C노드를 tail-to-head라고 한다.

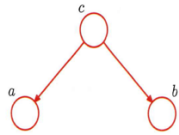


a. X

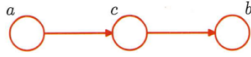
b. tail-to-head노드라고 한다.

정리하기

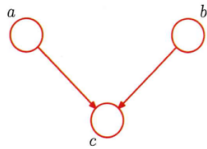
- SVM에 slack variable을 도입해서 오분류를 허용하는 모델을 만들 수 있다.
 - 그 경우 $\sum_{n=1}^N [1 - y_n t_n]_+ + \lambda \|w\|^2$ 를 최소화하는 것이 된다.
 - slack variable을 넣는 것은 오류로 힌지함수를 쓰는 것과 동일하게 된다.
- multiclass SVM도 가능한 하나 여러가지 문제로 많이 쓰이지는 않는다.
 - multiclass로 써야 할 때는 앞서 나온 함수를 쓴다. $y(x) = \max_k y_k(x)$
- SVM은 출력이 결정값(확률값X)이고 멀티클래스 확장이 어려우며 커널조건이 있다.
- RVM은 입력벡터를 조건으로 출력에 대한 조건부 분포를 가정하고 식을 전개한다.
 - SVM보다 더 희박한 모델을 결과로 준다
 - 일반식은 SVM과 동일하게 된다. $y(x) = \sum_{n=1}^N w_n k(x, x_n) + b$
 - SVM에 비해 훈련시간이 길다.
- 그래프는 노드와 링크로 이루어진다.
- 그래프모델에서 노드는 확률변수를 나타낸다.
- 방향이 있는것의 예는 베이지안 네트워크, 방향이 없는것의 예는 마르코프 무작위장
- 결합분포를 분해하는 방법에는 여러가지가 있을 수 있다.
 - 그래서 그래프로 나타내는 방법도 하나가 아니다.
 - 그래프로 나타내면 한눈에 들어온다.
- 베이지안 네트워크는 순환이 없다.
 - 보통 단말이 관측변수, 조상이 잠재변수.
- 방향성 그래프에서 구조에 따라, 그리고 관측여부에 따라 조건부 독립이나 아니냐를 알아볼 수 있다.



- a. 여기서 아무변수도 관측되지 않았다면 a,b는 조건부 독립이 아니다.(영향을 준다) 하지만 c가 주어지면, a와 b는 조건부 독립이 된다.



- b. 여기서도 동일하게 c가 관측되어야 a,b가 조건부 독립이 된다.



- c. 여기서는 반대로 c가 관측이 안되어 있어야 조건부 독립이다.

참고하기

Bishop, C. M. "Bishop-Pattern Recognition and Machine Learning-Springer 2006." Antimicrob. Agents Chemother (2014): 03728-14.

다음 차시 예고

- 조건부독립 예제
- d분리
- 마르코프 무작위장
- 트리
- 혼합모델
 - o K-means
 - o 혼합 가우시안
 - o EM