www.daishin.com

Quant Active Strategy

SEP 2008

리서치센터 투자전략부

Quant. Analyst 박주환
☎ 769-2764 / hanl@daishin.com

Quant. Analyst, Ph.D. 정진혁
☎ 769-3845 / jjung@daishin.com

내일 주가는 상승할까, 하락할까?

– SVM(Support Vector Machine)을 활용한 KOSPI 예측

이 보고서에서는 마켓타이밍에 대한 연구의 일환으로서 Support Vector Machine(이하 SVM)이라는 학습기법을 사용해서 주가의 상승과 하락시기의 특징과 패턴을 탐색함으로써 향후 주가를 통계적으로 유의한 수준으로 예측할 수 있다는 것을 보이고자 한다. SVM이란 패턴 분류기(pattern classifier)의 기능을 하는 기계학습기법으로, 많은 양의 입력변수와 이에 따르는 결과를 학습함으로써 새로운 입력변수에서는 어떠한 결과가 나올지를 예측해주는 AI(Artificial Intelligence)의 일종이다.

KOSPI의 시가, 고가, 저가, 종가 4가지의 데이터로 구성된 16개의 기술적 지표를 입력변수로 취하고 주가의 상승/하락 여부와 함께 학습시킨 후 내일, 그리고 다음주의 주가를 예측하는 SVM모델을 각각 구성한 후 시뮬레이션 작업을 했다. 결과는 일간모델의 경우 56.6%, 주간 모델의 경우 56.3%의 적중률을 나타냈으며, 이는 시도횟수(일간은 4600번, 주간은 870번)를 감안해서 볼 때 모델이 주가예측력이 있다고 할 수 있는 충분한 정도의 수치이다. 모델의 예측에 맞추어서 매수/매도 포지션을 취한다고 가정했을 때 수익률 시뮬레이션 결과는 일간 모델(4600영업일, 약 17년)의 경우 세금을 감안하더라도 무려 3000%에 달하는 수익률을 기록하였다. 56%의 예측가능성은 초과 수익을 달성하기 위해 필요한 수치를 넘어서는 수치이며, 모델의 주가 예측력을 보여주는 또 하나의 증거라고 할 수 있겠다.

SVM모델의 KOSPI의 상승/하락 예측 확률

	일간	일간2	주간	주간2
전체추정횟수	4600	1077	870	156
상승추정횟수	2398	527	457	88
하락추정횟수	2202	550	413	66
적중확률(%)	**56.6**	**59.6**	**56.3**	**64.1**
상승적중확률(%)	57.7	59.5	57.9	68.2
하락적중확률(%)	55.3	59.6	54.5	58.8
P-value	0	1.025E-10	8.227E-05	1.436E-04

주: 1. 일간2와 주간2의 모델에서는 SVM수치가 +1이상일 때 상승, -1이하일 때 하락신호로 간주하고, 그 사이의 값은 예측 표본에서 제외했을 경우임
2. P-value는 적중확률은 0.5라는 귀무가설에 대한 유의수준임. 모두 유의수준 5% 및 1%에서 편측 검정으로 귀무가설 기각 가능

자료: 대신증권 리서치센터

1. SVM 개요

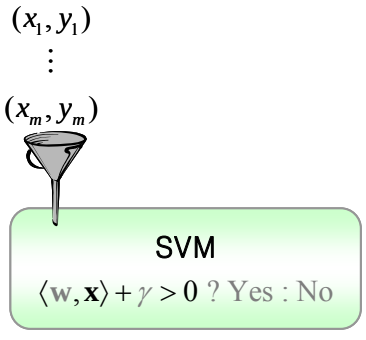
아이를 키워보면 알겠지만 아이들은 정말 빨리 배운다. 스스로 주변환경을 관찰해 터득하기도 하고 부모나 선생님의 지도를 통해 새로운 사실을 배우기도 한다. 마치 학습을 위해 태어난 기계 같다는 생각이 드는 것은 과장일까? 이런 아이와 같은 인공지능 기계를 만들 수 있을까? 완벽하지는 않지만 그와 같은 기계를 만들려는 시도는 많이 있으며 많은 분야에서 우수한 성과를 거두고 있고 사람이 판단하는 경우보다 적중률이 높기도 하다.

우리가 앓고 있는 병이나 증상을 인공지능 기계가 진단하기도 한다. 과거 종양 환자의 진단을 통해 얻은 종양의 크기 및 모양, 색깔, 혈액에서 발견되는 특이 단백질 등의 정보를 이용해 인공지능 기계를 학습시켜, 새로운 환자의 종양이 악성인지 양성인지 판단하는 작업을 할 수 있다. 뿐만 아니라 신용카드 부정사용이나 대출 한도 심사 등 많은 곳에서 알게 모르게 인공지능 기계들이 활동하고 있다.

학습 방법에는 크게 두 가지가 있다. 첫째는 선생님의 지도하에 지식을 습득하는 방법이 있고 둘째로는 스스로 관찰을 통하여 결론을 도출하는 방법이 있다. 전자를 지도학습(supervised learning)이라 하고 후자를 자율학습(unsupervised learning)이라 한다. SVM은 지도학습 방법의 한가지로 신경망과 같은 기존의 기계학습법과는 달리 통계적으로 검증된 예측능력을 지니고 있다. SVM은 마치 아이와 같아서 부모가 아기에게 옳고 그름을 반복해서 알려주면 아이는 패턴을 파악해 비슷한 상황이 왔을 때 스스로 옳고 그름을 판단하듯, 우리가 옳고 그름을 분류해 놓은 데이터를 대량으로 주어 SVM을 학습시키면 새로운 데이터에 대해 SVM이 스스로 판단을 하게 된다.

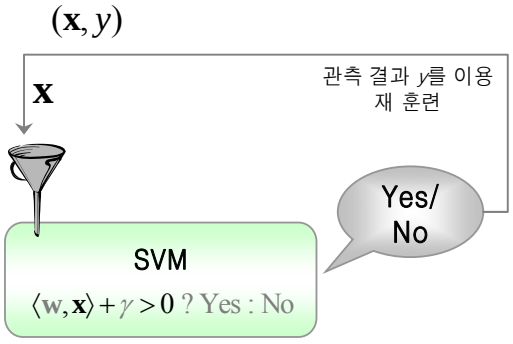
[그림 1]에 예시한 바와 같이 SVM의 내부에는 입력 데이터를 분류할 수 있는 분류경계면이 들어있으며 이를 초평면(hyperplane) $\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle - \gamma = 0\}$ 으로 정의한다.[1] SVM은 이 초평면에 의해 나눠지는 두 개의 반공간(half-space) 중에 입력 데이터가 어느 쪽에 속하느냐에 따라 예측을 수행하게 된다. SVM은 훈련과정을 통해 초평면을 정의하는 파라미터인 $\mathbf{w}$ 와 γ 의 최적 값을 찾아내고, [그림 2]와 같이 이 파라미터를 이용해 예측을 수행한다. 위 초평면 식을 보면 SVM이 선형 분류경계면만 찾아낸다고 생각할지도 모르나, 고차원으로의 맵핑(mapping)이나 커널트릭(kernel trick)을 사용하면 다양한 형태의 비선형 분류경계면을 찾아낼 수 있다. 더 자세한 내용은 부록을 참고하기를 바란다.

[그림 1] SVM 초기 훈련 과정



자료: 대신증권 리서치센터
주: 대량의 데이터를 이용해 SVM을 훈련시키면 SVM은 내부적으로 분류경계면 파라미터인 $\mathbf{w}$ 와 γ 를 결정한다.

[그림 2] SVM 훈련 후 예측(판단) 및 재훈련



자료: 대신증권 리서치센터
주: 훈련된 SVM은 새로운 입력 데이터에 대해 결과를 예측 할 수 있다. 예측 결과가 실현된 관측 결과와 다르면 관측 결과를 다시 입력으로 넣어 재훈련 시키는 작업을 하여 성능을 개선할 수 있다.

[1] $\langle \mathbf{w}, \mathbf{x} \rangle$ 은 벡터 $\mathbf{w}$ 와 $\mathbf{x}$ 의 내적이며 n 차원 유클리드 공간에서의 표준 내적은 $\langle \mathbf{w}, \mathbf{x} \rangle := \sum_{i=1}^{n} w_i x_i$ 로 정의된다.

2. SVM을 이용한 KOSPI예측

1) 예측을 위한 입력 데이터 선정

경험적으로 Macro/Micro 지표는 예측 확률을 높이지 못함

주가의 기본적 분석에 사용되는 지표들 중에서 환율, 금리, 경기선행지수, 각종 상품가격 등의 매크로 데이터와 PER, 실적 추정치 추이 등의 기업 실적 데이타가 주가와 관련이 있다고 알려져 있지만, SVM의 입력 변수로 이들 지표를 사용했을 경우 주가의 상승, 하락 패턴을 효율적으로 분류해 내지 못했다.[2] 따라서 이번 연구에서는 기본적 분석보다는 기술적 분석을 위한 지표들을 입력변수로 사용한 기존의 연구[3,4,5]를 토대로 입력 데이터를 구성했다.

KOSPI의 시가, 고가, 저가, 종가로만 지표를 구성

기본적으로 많은 양의 데이터를 사용하지 않는 것을 원칙으로 했다. 이는 결과 데이터의 분석에 용이하고 향후 실전적인 매매 전략에 활용하기 쉽다는 장점이 있을 것으로 판단된다. 따라서 거래량/거래대금, 상승/하락 종목 수, 매매주체별 순매수 등의 데이터들은 배제하고 오로지 KOSPI의 시가, 고가, 저가, 종가 데이터만을 이용해서 다음 **[표 1]**과 같은 기술적 지표 16개를 만들어서 입력변수로 사용하였다.

---

2 이러한 사실이 기본적 분석 지표가 주가에 무관하다는 것을 의미하는 것은 아니다. 이들 변수들은 시계열의 흐름이 중요하지만 SVM은 모든 변수를 독립 변수로 간주하기 때문에 변수들이 담고 있는 정보를 효과적으로 사용하지 못하는 것으로 풀이된다. 변수들을 시계열의 추세와 반전 등을 효과적으로 나타낼 수 있도록 변경하는 작업이 가미된다면 주가와의 관련성이 나타날 수 있다.

3 김유일, 신은경, 홍태호, "*신경망과 SVM을 이용한 주가지수 예측의 비교*", 인터넷전자상거래연구 제4권 3호 2004년

4 Kyoung-jae Kim, "*Financial time series forecasting using support vector machines*", Neurocomputing, 2003

5 Rohit Choudhry, Kumkum Garg, "*A Hybrid Machine Learning System for Stock Market Forecasting*", Proceedings of World Academy of Science, Engineering and Technology Volume 29 May 2008

[표 1] SVM모델을 위한 지표 구성

지표	계산식	비고
KOSPI 수익률	CP(t) = [C(t) – C(t–1)]/C(t–1)	일일 종가 수익률
KOSPI 수익률의 Unexpected 정도	일간: CP(t) + CP(t–1) * 0.23 주간: CP(t) – CP(t–1) * 0.23	ARIMA(1,1)모델에서 착안 KOSPI의 당일 수익률은 전일 수익률의 약 0.23배 만큼 음의 방향(주간일 경우는 양의 방향)이고 나머지 수익률은 Unexpected 라고 가정
12일 KOSPI 수익률	[C(t) – C(t–12)]/C(t–12)	
시가대비 종가의 수익률	[C(t) – S(t)]/S(t)	
전일 종가대비 시가의 수익률	[S(t) – C(t–1)]/C(t–1)	
Linear Slope	회귀선 C = a + b*t + e, t=1,2,…,14로 a, b값을 추정한 후 t값에 15를 입력	14일간의 KOSPI 종가와 1~14 의 회귀분석 후, 15번째의 KOSPI 예측치
CCI	A = [H(t)+L(t)+C(t)]/3 B = A의 5일 이동평균 C = |A–B| CCI = (A–B)/(C의 5일 이동평균*0.015)	주가의 최근 평균치와 현 주가 사이의 편차를 나타내는 지표
MACD Oscillator	KOSPI의 12일 이동평균 – KOSPI의 26일 이동평균	
RSI	5일 주가상승폭 합/(5일 주가상승폭 합 + 5일 주가하락폭 합) * 100	
Momentum	C(t) – C(t–4)	
%K	(종가 – 5일 최저가)/(5일 최고가 – 5일 최저가) * 100	
%D	%K의 5일 이동평균	
SLOW %D	%D의 5일 이동평균	
DM+	[H(t)–H(t–1)] > 0 이고 [H(t)–H(t–1)] > [L(t–1)–L(t)] 이면 [H(t)–H(t–1)] 아니면 0	
DM–	[L(t–1)–L(t)] > 0 이고 [L(t–1)–L(t)] > [H(t)–H(t–1)] 이면 [L(t–1)–L(t)] 아니면 0	
TR	[H(t) – L(t)], |C(t–1)–H(t)|, |C(t–1)–L(t)| 중의 최대값	

주: 1. C(t) – t시점의 종가, S(t) – t시점의 시가, H(t) – t시점의 고가, L(t) – t시점의 저가
2. 주간 모델에서는 KOSPI의 시가, 고가, 저가, 종가 데이터를 주간 데이터로 전환했음
자료: 대신증권 리서치센터

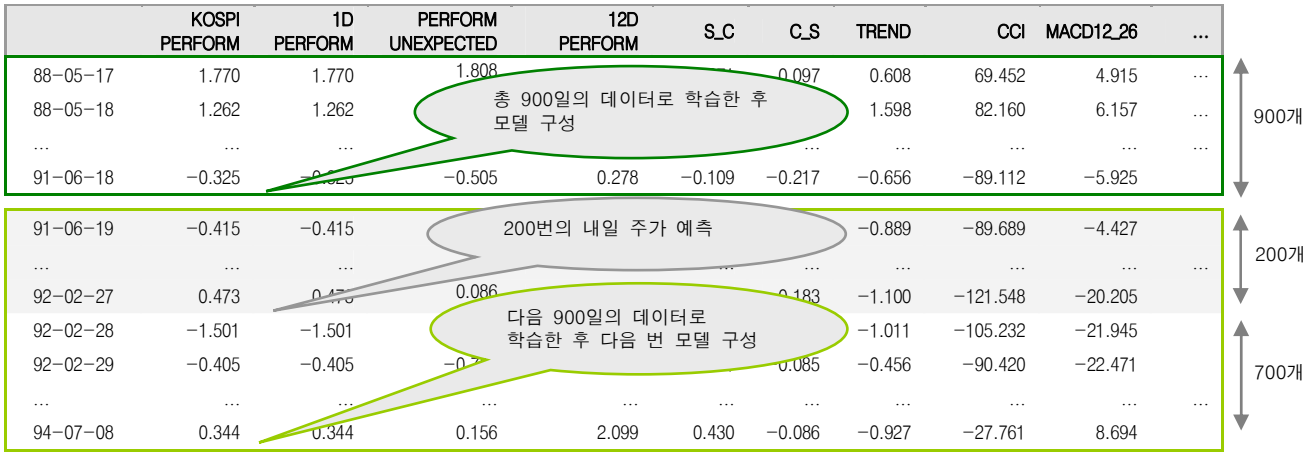
2) Daily 주가 예측

900개의 학습데이터로 모델 구성, 내일의 상승/하락 예측을 200번 반복

SVM의 학습데이터의 개수는 900개로 고정했다. 즉, 이전 900일의 학습데이터로 모델을 구성한 후, 향후 200일 동안 해당 시점에서 내일의 KOSPI의 상승/하락을 예측했다. 이렇게 한 개의 모델로 200번의 예측을 한 후, 다시 해당 시점에서 이전 900일의 데이터로 모델을 구성하는 방식으로 진행하였다.

KOSPI 1991년 6월 19일부터 2008년 9월 5일까지 총 4600번의 예측

1991년 6월 18일 이전의 900일의 학습데이터를 시작으로 1991년 6월 19일부터 KOSPI의 예측을 시작했으며, 2008년 9월 5일까지 이러한 작업을 총 23번 반복해서 4600일을 예측했다. 각각의 23개의 모델에서 nu, sigma등의 파라미터와 커널함수는 동일하고 오직 학습데이터만을 최근 900일 동안의 데이터로 변경시켰다.



일간 상승/하락 적중확률 56.6%

4600번의 예측 중에서 2603번이 적중해서 56.6%의 적중확률을 보였다. P-value는 0으로써, 우연히 이러한 결과가 나올 확률은 0에 가깝다. 모델의 유의함을 보이기 위해 SVM수치가 +1 이상인 경우와 -1이하를 나타냈을 경우(즉, 모델이 상승/하락을 확연하게 구분하는 경우)만을 고려해 보았는데, 이 경우(아래 **[표 2]**에서 일간2)에는 적중확률이 59.6%로 상승했다.

[표 2] 일간 SVM모델의 KOSPI의 상승/하락 예측 확률

	일간	일간2
전체추정횟수	4600	1077
상승추정횟수	2398	527
하락추정횟수	2202	550
적중확률(%)	56.6	59.6
상승적중확률(%)	57.7	59.5
하락적중확률(%)	55.3	59.6
P-value	0	1.025E-10

주: 1. 일간2 모델에서는 SVM수치가 +1이상일 때 상승, -1이하일 때 하락신호로 간주하고, 그 사이의 값은 예측 표본에서 제외했을 경우
2. P-value는 적중확률은 0.5라는 귀무가설에 대한 유의수준임

자료: 대신증권 리서치센터

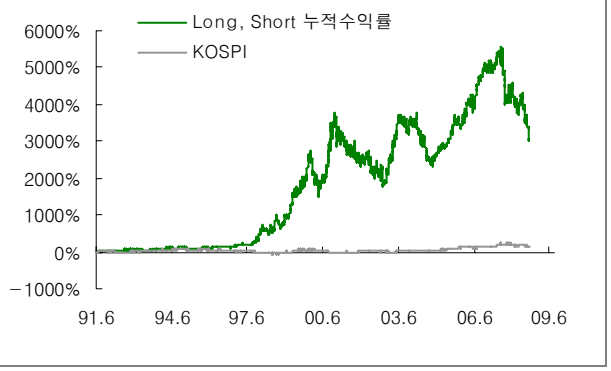
일간 SVM모델의 성과모사 – 56.6% 의 적중확률은 충분히 초과 수익을 낼수있는 수준

본 연구는 애초에 매매 전략의 관점에서 접근했다기 보다는 향후 주가 예측에 있어서 각종 지표를 입력변수로 하는 SVM기법의 활용 가능성에 초점을 맞추었다. 실전적 매매 전략의 각도에서 연구하고자 한다면 아래에서 수행된 성과 모사보다는 실시간 매매 가능성, 매매 수수료 및 시장 충격 비용 등을 고려한 좀 더 정교한 성과 분석이 이루어져야 할 것이다. 다만 SVM모델이 나타내는 56% 정도의 적중률이 얼마만큼의 성과를 내는지 정성적으로 판단하기 위해 시뮬레이션을 간단하게 시도했음을 알려두는 바이다.

모델이 내일 주가의 상승과 하락을 예측했을 때 당일 종가로 매매가 가능하다고 가정한다. 즉, 내일 주가가 상승할 것을 예측하면 당일 종가로 매수(Long), 하락할 것을 예측하면 당일 종가로 매도(Short) 한다고 했을 경우 성과모사를 해 보았다. 포지션이 바뀌는 날은 0.3%의 세금 효과를 감안하고 수수료는 없다고 가정했을 때, 1991년 6월 19일부터 2008년 9월 5일까지 총 4600일 동안 2999%의 누적 성과를 보였다. 동기간 KOSPI는 131% 상승했다**[그림 3]**.

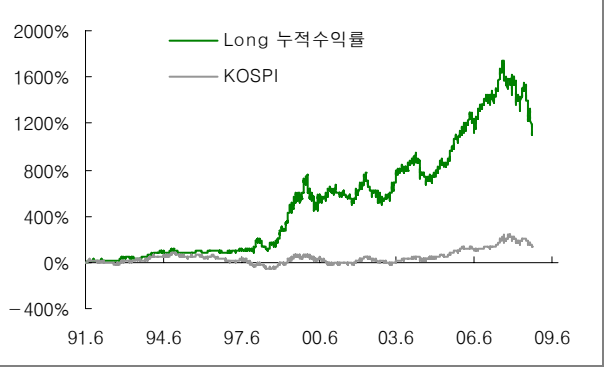
매수만 가능하다고 했을 때, 즉, 상승 예측 시 매수하고 하락 예측 시 청산한다고 가정했을 경우에는 동기간 1096%의 누적 성과를 보였다**[그림 4]**.

[그림 3] 주가 상승 예측 시 Long, 하락 예측 시 Short했을 경우의 누적수익률 2999%



주: SVM모델 상의 예측에 따라 포지션을 변경하는 경우, 즉 Long->Short, 혹은 Short->Long인 경우 0.3%의 세금을 고려함, 세금을 고려하지 않을 경우는 250,000%가 넘는 누적수익률을 기록함
자료: Fnguide, 대신증권 리서치센터

[그림 4] 주가 상승 예측 시 Long만 하는 경우의 누적수익률 1096%



주: 매도 시 0.3%의 세금을 고려함, 세금을 고려하지 않을 경우에는 10,654%의 누적수익률을 기록함.
자료: Fnguide, 대신증권 리서치센터

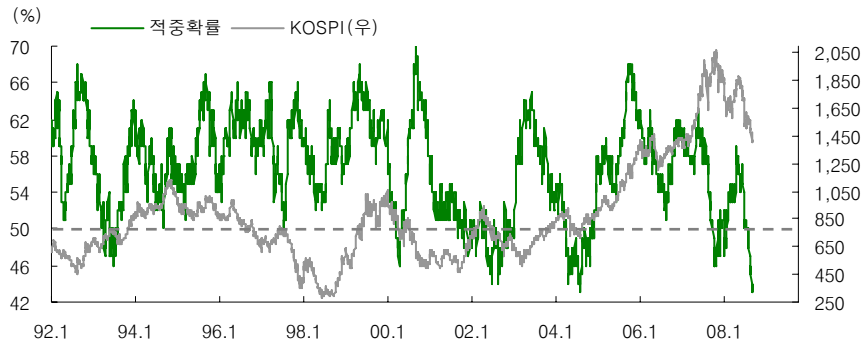
[표 3] 일간 SVM모델을 이용한 매매 전략의 수익률 통계

	KOSPI	SVM Long/Short	SVM Long
누적수익률(%)	131.4	2999.4	1096.5
일일 평균수익률(%)	0.034	0.091	0.063
표준편차(%)	1.777	1.801	1.309
Sharp Ratio	0.019	0.050	0.048
상승횟수	2370(51.54%)	2457(53.41%)	1326(55.30%)
하락횟수	2228(48.46%)	2143(46.59%)	1072(44.70%)
Skewness	-0.041	-0.062	0.003
Kurtosis	3.278	3.205	8.896

자료: 대신증권 리서치센터
주: Sharp Ratio는 연율화 수치가 아닌 [일일 평균수익률/표준편차]로 계산함

마켓 타이밍을 56%정도로만 예측해도 그 누적 성과는 상상 외로 엄청난 성과를 나타낸다는 것을 알 수 있다. 달리 말하면 56%의 적중이 수치상으로 느껴지는 것 보다는 수월하지 않다는 것이다. 4600일 동안 전체적으로 56.6%의 적중 확률을 나타내긴 했지만, 다음 **[그림 5]**에서 보듯이 그 정도가 항상 일정하지는 않다. 모델의 예측력이 지속적이며 안정적이라고는 단정지을 수 없는 부분이다. 최근 들어서는 50%를 밑도는 적중 확률을 나타내고 있다.

[그림 5] 적중확률의 시계열 – 적중하는 정도가 일정하지 않다



주: 적중확률은 해당 시점의 이전 100일 동안의 적중 횟수로 계산함
자료: 대신증권 리서치센터

3) Weekly 주가 예측

600개의 학습 데이터로 모델 구성, 다음주의 KOSPI 상승/하락 예측을 10번 반복

주간 SVM의 입력변수인 학습 데이터의 개수는 600개로 고정했다. 즉, 이전 600주의 주간 학습 데이터로 모델을 구성한 후, 향후 10주 동안 해당 시점에서 다음 주의 KOSPI의 상승/하락을 예측했다. 이렇게 한 개의 모델로 10번의 예측을 한 후, 다시 해당 시점에서 이전 600주의 데이터로 모델을 구성하는 방식으로 진행하였다.

KOSPI 1992년 1월 11일부터 2008년 9월 5일까지 총 870번의 예측

1992년 1월 11일 이전의 600주의 주간 학습 데이터를 시작으로 1992년 1월 11일부터 KOSPI의 예측을 시작했으며, 2008년 9월 5일까지 이러한 작업을 총 87번 반복해서 870주를 예측했다. 각각의 87개의 모델에서 nu, sigma등의 파라미터와 커널함수는 동일하고 오직 학습 데이터만을 최근 600주 동안의 데이터로 변경시켰다.

	KOSPI PERFORM	1D PERFORM	PERFORM UNEXPECTED	12D PERFORM	S_C	C_S	TREND	CCI	MACD12_26	...	
80-07-12	-0.813	-0.813	-0.569				-0.912	-82.263	4.740	...	600개
80-07-19	-0.117	-0.117					-0.708	-113.828	3.755	...	
...	...	...	...				...	...	...	...	
92-01-04	6.448		6.068	-7.362	5.375	1.018	6.429	51.343	-19.873		
92-01-11	-4.981	-4.981					-1.106	17.507	-25.851		10개
...	...	...				...	...	...	...	...	
92-03-14	-2.408		-2.859	1.404	-2.446	0.038	0.800	-53.622	-16.947		
92-03-21	1.288	1.288	1.842				1.079	-25.873	-13.656		590개
92-03-28	-2.372	-2.372					-0.155	-24.932	-14.201		
...	...	...				...	...	...	...	...	
03-07-15	2.358		2.658	18.917	2.908	-0.534	0.944	48.826	28.646		

총 600주의 데이터로 학습한 후 모델 구성

10번의 다음주 주가 예측

다음 600주의 데이터로 학습한 후 다음 번 모델 구성

주간 상승/하락 적중확률 56.3%

870번의 예측 중에서 490개가 적중해서 56.3%의 적중확률을 보였다. P-value는 8.227E-05으로써, 우연히 이러한 결과가 나올 확률은 0.01% 미만이다. 일간의 경우와 마찬가지로 모델의 유의함을 보이기 위해 SVM수치가 +1이상인 경우와 -1이하를 나타냈을 경우(즉, 모델이 상승/하락을 확연하게 구분하는 경우)만을 고려해 보았는데, 이 경우(아래 [표 4]에서 주간2)에는 적중확률이 64.1%로 상승했다.

[표 4] 주간 SVM모델의 KOSPI의 상승/하락 예측 확률

	주간	주간2
전체추정횟수	870	156
상승추정횟수	457	88
하락추정횟수	413	66
적중확률(%)	56.3	64.1
상승적중확률(%)	57.9	68.2
하락적중확률(%)	54.5	58.8
P-value	8.227E-05	1.436E-04

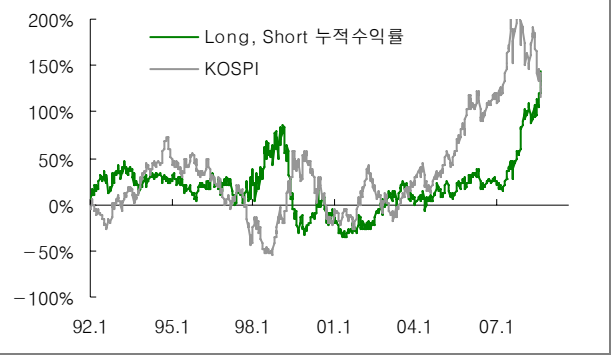
주: 1. 주간2의 모델에서는 SVM수치가 +1이상일 때 상승, -1이하일 때 하락신호로 간주하고, 그 사이의 값은 예측 표본에서 제외했을 경우임
2. P-value는 적중확률은 0.5라는 귀무가설에 대한 유의수준임
자료: 대신증권 리서치센터

주간 SVM모델의 성과모사 – KOSPI 대비 28%의 초과 수익

일간 SVM의 경우와 마찬가지로 다음주 주가가 상승할 것을 예측하면 당일 종가로 매수(Long), 하락할 것을 예측하면 당일 종가로 매도(Short) 한다고 했을 경우의 성과모사를 해보았다. 포지션이 바뀌는 경우 0.3%의 세금효과를 감안하고 수수료는 없다고 가정했을 때, 1992년 1월 11일부터 2008년 9월 5일까지 총 870주 동안 KOSPI는 116% 상승하고 Long/Short 포트폴리오는 144%의 누적 성과를 보였다. KOSPI 대비 28%의 초과 수익을 달성했지만, 일간 모델의 경우처럼 크게 상회하지는 못했다**[그림 6]**.

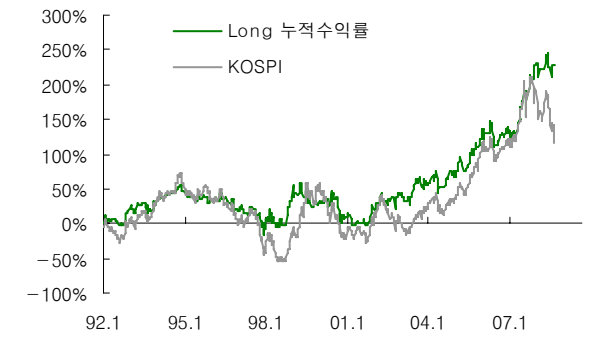
매수만 가능하다고 했을 때, 즉, 상승 예측 시 매수하고 하락 예측 시 청산한다고 가정했을 경우에는 동기간 229%의 누적 성과를 보였다**[그림 7]**.

[그림 6] 주가 상승 예측 시 Long, 하락 예측 시 Short했을 경우의 누적수익률 – KOSPI 대비 27.5% 초과수익



주: SVM모델 상의 예측에 따라 포지션을 변경하는 경우, 즉 Long->Short, 혹은 Short->Long인 경우 0.3%의 세금을 고려함, 세금을 고려하지 않을 경우는 584.5%의 누적수익률을 기록함.
자료: Fnguide, 대신증권 리서치센터

[그림 7] 주가 상승 예측 시 Long만 하는 경우의 누적수익률 – KOSPI 대비 113% 초과수익



주: 매도 시 0.3%의 세금을 고려함, 세금을 고려하지 않을 경우에는 448.9%의 누적수익률을 기록함.
자료: Fnguide, 대신증권 리서치센터

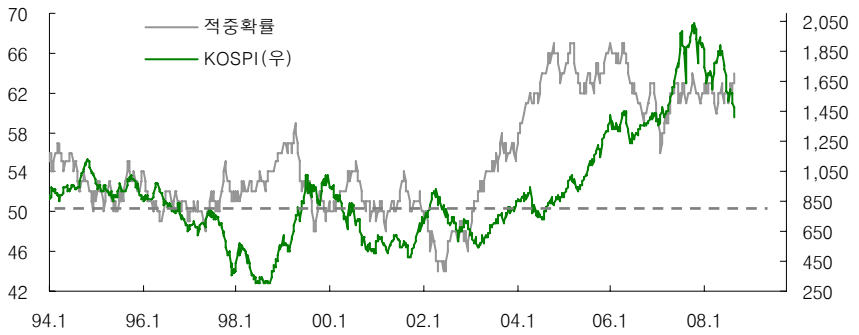
[표 5] 주간 SVM모델을 이용한 매매 전략의 수익률 통계

	KOSPI	SVM Long/Short	SVM Long
누적수익률(%)	116.0	143.5	228.7
주간 평균수익률(%)	0.175	0.188	0.182
표준편차(%)	4.157	4.136	3.003
Sharp Ratio	0.042	0.046	0.061
상승횟수	453(52.07%)	487(55.98%)	264(57.77%)
하락횟수	417(47.93%)	383(44.02%)	193(42.23%)
Skewness	0.072	-0.168	0.057
Kurtosis	1.699	1.714	6.208

주: Sharp Ratio는 연율화 수치가 아닌 [주간 평균수익률/표준편차]로 계산함
자료: 대신증권 리서치센터

적중 확률의 추이를 보면, 주간 모델의 경우에는 2000년 전후에서는 적중확률이 높지 않았지만, 2004년 이후 상승해서 최근에는 60%를 상회하는 적중률을 나타내고 있다**[그림 8]**.

[그림 8] 적중확률의 시계열 – 2004년 이후 적중확률이 높아짐



주: 적중확률은 해당 시점의 이전 100주 동안의 적중 횟수로 계산함
자료: 대신증권 리서치센터

3. 결론

1) SVM기법의 예측력과 한계점

과거데이터를 이용한 패턴 분석은 모든 분석가들의 기본적인 분석 방식이다

예측하려는 지표와 관련 있는 과거 데이터의 패턴을 분류하고, 현재 데이터를 바탕으로 향후 미래를 예측하려는 행위는 모든 분석가들이 수행하는 방법이다. SVM기법 역시 이러한 기능을 통계적인 기준에서 수행한다고 볼 수 있으며, 이진 의사결정에서 기존의 로지스틱 회귀분석, 인공신경망, 판별 분석 등의 기법보다 효과적이고 우수한 성능을 나타내면서도 수학적으로 간단한 기법이다. 이런 결과가 나오는 것은 SVM이 단순히 분류 평(곡)면을 찾는 다거나 표본 에러를 최소화하는 작업을 하는 것이 아니라 분류여백(Separation Margin)을 최대화함으로써 학습데이터가 아닌 새로운 데이터에 대해서도 올바르게 분류할 가능성을 높이기 때문이다.

평균적으로 56%의 적중률만 보여도 초과 수익은 보장된다 – 그 이상의 적중은 힘들다는 것을 시사

일간 SVM모델의 성과 모사의 경우에서 보듯이 56%의 적중만 가능하다면 초과 수익은 크게 걱정할 바가 되지 못한다. 이는 위험/수익 상충관계의 관점에서 보면 향후 56%의 적중 역시 보장하기 쉽지 않다는 것을 시사한다. 무작위적으로 1000번 이상 시도했을 때 56%이상의 적중률을 나타내는 것은 불가능에 가까우며, 랜덤 워크라는 단단한 갑옷에 둘러싸인 주가에 대한 예측확률로서는 유의한 수치라고 할 수 있다.

SVM모델은 직관적인 설명도출이 힘들다

SVM기법의 단점 중 하나는 분류 작업을 가상의 고차원 공간에서 수행하기 때문에 입력 변수와 이에 대한 결과값에 대해 직관적인 설명을 도출하기가 쉽지 않다는 점이다. 모델이 주가 상승을 예측했을 경우, 어떠한 변수가 크게 작용을 했는지, 혹은 변수가 어떠한 방향으로 얼마만큼 변하면 결과치가 다르게 나올지 예측하기 힘들기 때문에, 모델의 유의성을 검증하려 할 때 과거 데이터의 적중률에 의존할 수밖에 없다.

입력 변수들이 적절한 변수인지에 대한 심층 연구 부족

본 연구에서 사용된 기술적 지표들은 현재 금융 시장에서 널리 알려진 지표들이긴 하지만 이러한 지표들이 SVM의 입력변수로서 적절한 변수인지를 판단할 때, 전체 시계열의 적중률에만 의존했는데 좀 더 정교한 판단 기준이 필요할 것으로 판단된다. 단적으로 과거에는 유의한 지표였지만, 지금은 그 유의성이 감소했을 수 있다. 일간 예측 모델에서 최근 1년여 동안 적중률이 현저히 낮아졌다는 사실도 이를 뒷받침한다.

2) SVM기법의 활용방안

국가별 지수 예측 및 종목 스크리닝에도 적용가능

SVM 기법은 패턴을 분류하는데 그 스타일을 가리지 않고 충분히 좋은 성능을 보이고 있으므로 다른 여타 국가의 상승/하락 예측이나, 종목 선정에도 얼마든지 활용 가능할 것으로 보인다. 다만, 적절한 입력변수의 선정과 SVM 커널 함수 및 파라미터를 선택하는 작업이 필요할 뿐이다.

특정 지표의 주가와의 관련성 여부 판단에 이용

무엇보다 SVM 기법은 과거 및 현재 특정 지표와 주가 수익률이 관련이 있는지 없는지를 판가름 할 수 있는 도구로서 활용 가능하다고 판단된다. 만약 어떠한 지표가 미래의 주가 수익률에 어떤 관련성 있다면 SVM기법은 어렵지 않게 그 관련성을 잡아 낼 수 있을 것이다. 지나치게 직관적이고 시장의 효율성을 고려하지 않은 상식적인 분석에 의존하기 보다는 이러한 통계적인 기법을 활용하면 조금 더 구체적이고 오랜 기간 지속 가능한 연구 성과를 얻을 수 있을 것으로 믿는다.

예측 성능 개선의 여지는 많아

본 리포트에서는 symmetric하고 positive definite(이하 SPD)인 커널 중 하나인 가우시언 커널만을 사용했으나, SPD 커널의 특성상 둘 이상의 서로 다른 SPD 커널의 양의 선형 조합도 SPD 커널이 되는 특징이 있다. 이를 활용한 기법을 커널 boosting이라 하며 이 기법을 이용하여 예측 성능을 높일 수 있다.[6] 또한 진화연산(evolutionary computing) 기법을 사용하여 최적의 커널 조합과 커널 파라미터 등을 찾아내 예측 성능을 개선할 수 있을 것으로 생각한다.[7] 56%의 예측 성능으로도 상당한 초과수익을 얻었음을 볼 때 약간의 성능 개선으로도 얻을 수 있는 초과수익은 클 것이다.

[6] Trevor Hastie, R. Tibshirani and J. Friedman. *The Element of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, USA. 2001.

[7] Frauke Friedrichs and Christian Igel. *Evolutionary Tuning of Multiple SVM Parameters*. Neurocomputing. 64:107–117, 2004.

A. 부록

1) SVM 훈련과정

SVM은 훈련과정을 통해 훈련데이터를 나눌 수 있는 초평면을 찾아낸다. 사실 훈련데이터를 잘 분류해낼 수 있는 초평면이 하나 이상 있을 수 있기 때문에, 어떤 초평면이 가장 좋은가 하는 문제에 대해서 생각해볼 필요가 있다. 먼저 훈련데이터부터 기호로 정의하고 이를 나눌 초평면에 대해 다루도록 하겠다.

입력데이터로 총 m 개의 훈련데이터가 주어지며, 각각의 훈련데이터 $\mathbf{x}_i \in \mathrm{R}^n$ 는 여러가지 속성(예측을 위해 사용했던 KOSPI 수익률, CCI, MACD oscillator 등) 별로 어떤 값을 모아놓은 유클리드 공간에서의 벡터로 볼 수 있다. 또한 각 훈련데이터에는 어떤 결과값(상승/하락)이 알려져 있으므로 이를 레이블 $y_i \in \{-1,+1\}$ 로 표현할 수 있다. 즉 모든 훈련데이터는 벡터/결과의 쌍으로 구성된다. 훈련데이터는 크게 두 가지로 분류할 수 있다. 즉, $y_i = +1$ 인 +훈련데이터가 있고 $y_i = -1$ 인 –훈련데이터가 있다.

이제 데이터를 잘 구분하는 초평면이 어떤 것일까에 대해 생각해보자. 우선 훈련데이터를 완전히 구분해낼 수 있는 초평면이 하나라도 존재하는 경우에 대해 다뤄보자. 이 경우 **[그림 9]**에서 보듯 훈련데이터를 정확하게 분류해낼 수 있는 초평면은 무한히 많다. 이 중에서 어떤 것이 좋은 평면일까? **[그림 9]**의 (a)를 사용해 예측할 경우를 생각해보자. 예측하라고 준 입력 데이터에는 관측오류가 있을 수 있다. 예를 들어 [그림 9]의 데이터 ⓓ가 예측용 데이터로 주어졌다고 하자. 실질적으로 ⓓ가 관측 에러 때문에 회색 점선 안 어디에선가 관측될 수도 있다. 운이 나쁘면 (a)가 오답을 줄 수도 있다. 반면 (b)는 ⓓ가 어디서 관측되건 상관없이 정답을 줄 것이다. 초평면 (a)와 (b)의 차이는 무엇일까?

[그림 9]를 보면 평면 (a)나 (c)는 몇몇 훈련데이터에 너무 가까운 반면 (b)는 모든 데이터에 대해 적절한 거리를 두고 있는 것을 알 수 있다. 즉 (a)와 (c)같이 훈련데이터에 너무 가까우면 관측 오류로 인해 잘못된 예측 결과를 내놓을 확률이 높아진다. 따라서 최적의 분류 평면은 가장 가까운 + 혹은 –데이터까지의 거리를 최대로 하는 평면이라 할 수 있다. 가장 가까운 +/– 훈련데이터까지의 거리의 합을 분류여백(separation margin)이라 한다. 분류여백을 최대화 함으로서 훈련데이터에 대한 분류 능력(실험오류, empirical error)으로부터 예측 능력에 대해 추정할 수 있다.[8]

분류여백을 최대로 하는 초평면은 어떻게 찾을 수 있을까? 분류 초평면으로 데이터를 직접 나누는 대신 **[그림 10]**과 같이 분류 초평면에 평행한 두 경계 초평면을 만들어 훈련데이터를 분류하는 방법을 사용하면 된다. 이러면 경계간의 거리가 분류 여백의 폭과 같아진다. 다시 말해 + 경계 $\langle \mathbf{w}, \mathbf{x} \rangle - \gamma = +1$ 는 +훈련데이터를 잘 분류하고, 즉

$$\langle \mathbf{w}, \mathbf{x} \rangle - \gamma \geq +1, \; y_i = +1 \text{ 인 경우,} \tag{1}$$

–경계 $\langle \mathbf{w}, \mathbf{x} \rangle - \gamma = -1$ 는 –훈련데이터를 잘 분류하도록, 즉

$$\langle \mathbf{w}, \mathbf{x} \rangle - \gamma \leq -1, \; y_i = -1 \text{ 인 경우,} \tag{2}$$

와 같이 설정한다. 수식 (1)과 (2)를 하나로 정리하면

$$y_i \left(\langle \mathbf{w}, \mathbf{x} \rangle - \gamma \right) \geq 1 \tag{3}$$

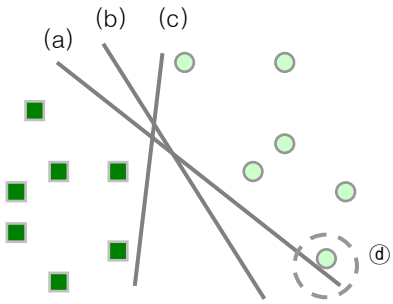
[8] Olivier Bousquet and Andre Elisseeff. *Stability and Generalization*. Journal of Machine Learning Research. 2:499–526, 2002.

로 표현할 수 있다. 이제 최적의 초평면을 찾는 문제는 (3)을 만족하면서 +경계와 –경계간의 거리를 최대로 하는 $\mathbf{w}$ 와 γ 를 찾는 것으로 바뀐다. 결국 분류 초평면은 두 경계 평면의 정확히 가운데 위치한다. 두 경계 평면 사이의 거리는 $2/\|\mathbf{w}\|$ 다. 따라서 최적의 평면을 찾는 문제는

$$\min_{\mathbf{w},\gamma} \frac{1}{2}\|\mathbf{w}\|^2$$
$$s.t.\ y_i(\langle \mathbf{w},\mathbf{x}\rangle - \gamma) \geq 1,\ \text{for } i = 1,\ldots,m \qquad (4)$$

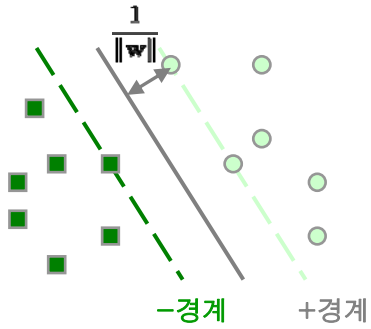
와 같은 최적화 문제로 표현된다.[9]

[그림 9] 훈련데이터를 나눌 수 있는 초평면



주: 훈련데이터를 나눌 수 있는 초평면은 무한히 많이 존재한다. 따라서 이 중 최적의 평면을 찾아낼 필요가 있다
자료: 대신증권 리서치센터

[그림 10] 최적의 초평면



주: 최적의 초평면은 + 경계와 – 경계간의 거리를 최대로 하는 것이다.
자료: 대신증권 리서치센터

훈련데이터를 완벽하게 구분해내는 평면이 없을 수도 있다. 이런 경우 추가적인 변수 $\xi_i \geq 0$ 를 도입하여 수식 (4)의 제약조건을

$$y_i(\langle \mathbf{w},\mathbf{x}\rangle - \gamma) + \xi_i \geq 1,\ \text{for } i = 1,\ldots,m$$

와 같이 완화하여 구분해 내지 못하는 데이터를 허용한다. 다만 이 경우 모든 훈련데이터에 대해 관대해질 수 있으므로, 목적 함수를

$$\min_{\mathbf{w},\gamma,\xi_1,\ldots,\xi_m} \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{m} C_i\xi_i$$

와 같이 고쳐 구분해 내지 못하는 데이터에 대해 페널티(penalty)를 주도록 한다. 즉 가급적이면 훈련데이터를 잘 분류하는 초평면을 찾도록 하는 것이다. 위 식에서 C_i 는 i 번째 데이터를 +/− 경계 평면으로 구분하지 못할 경우 부여할 페널티의 강도를 나타내며 반드시 모든 i 에 대해 $C_i > 0$ 여야 한다. 이렇게 유도한 식을 정리하면

$$\min_{\mathbf{w},\gamma,\xi_1,\ldots,\xi_m} \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{m} C_i\xi_i$$
$$s.t.\ y_i(\langle \mathbf{w},\mathbf{x}\rangle - \gamma) + \xi_i \geq 1,\ \text{for } i = 1,\ldots,m \qquad (5)$$
$$\xi_i \geq 0,\ \text{for } i = 1,\ldots,m$$

[9] $\|\mathbf{w}\| := \sqrt{\langle \mathbf{w},\mathbf{w}\rangle}$

가 된다. 이 문제를 해결해 훈련시킨 SVM을 soft-margin SVM이라 한다. 이 외에도 다른 식을 통해 soft-margin SVM을 훈련시킬 수도 있는데 이와 구별하기 위해 (5)를 풀어 훈련시킨 SVM을 C-SVM이라 부른다. 문제 (5)는 가장 쉽게 직관적으로 도출할 수 있는 식인 반면 파라미터 C_i 를 적절히 정하기 힘든 면이 있다.

이를 보완하여 설계된 것이 본 리포트에서 사용한 ν-SVM으로서 최적화 문제,

$$\min_{\mathbf{w},\gamma,\xi_1,\ldots,\xi_m,\rho} \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{N}\sum_{i=1}^{m}\xi_i$$
$$s.t.\ y_i(\langle \mathbf{w},\mathbf{x}\rangle - \gamma) + \xi_i \geq \rho,\ \text{for } i = 1,\ldots,m$$
$$\xi_i \geq 0,\ \text{for } i = 1,\ldots,m,$$

를 풀어 훈련시킨다. ν-SVM의 특징은 파라미터인 ν 를 사용해 서포트 벡터(support vector, 이하 SV)의 비율 상한을 정할 수 있다는 점이다.[10] 서포트 벡터란 +/- 경계 평면 위에 있거나 +/- 경계 평면이 잘못 구분한 훈련데이터를 말한다. 서포트 벡터의 비율은 훈련 오류의 상한이기도 한다.

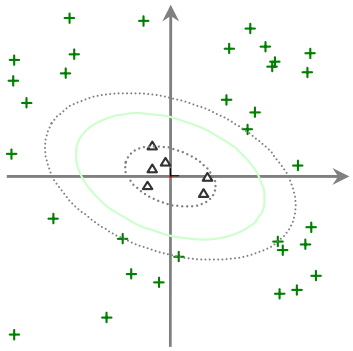
2) 맵핑과 커널트릭

데이터를 분류할 때 단순한 선형 초평면만으로는 분류가 힘들 때가 많다. 예를 들어 **[그림 11]**과 같이 2차원 입력공간(input space)에서 +데이터는 원점에서 충분히 멀리 떨어져 있고 -데이터는 원점 근처에 분포하고 있는 경우를 생각해보자. 이 경우 2차원 공간에서의 선형 초평면인 직선으로는 훈련데이터를 분류할 수 없다. 하지만 훈련데이터를 다음과 같은 매핑(사상, mapping)

$$\Phi(\mathbf{x}) : (x_1, x_2) \mapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2) \qquad (6)$$

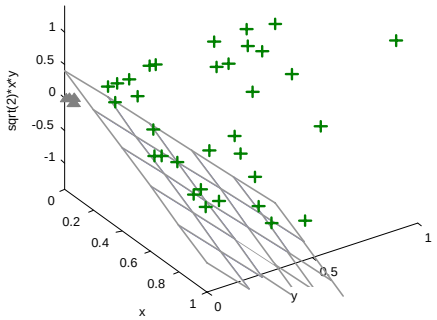
을 사용하여 3차원 특징공간(feature space)으로 이동시키면 **[그림 12]**와 같이 선형 분류 초평면(3차원에서는 평면)을 찾아낼 수 있다.

[그림 11] 비선형 분류경계



주: 녹색 +는 +데이터이고 흑색 세모는 -데이터이다. 연두색 타원이 분류경계이고 외측 타원은 +경계, 내측 타원은 -경계이다. 이 데이터는 2차원에서의 초평면인 직선으로는 분류할 수 없다.
자료: 대신증권 리서치센터

[그림 12] 고차원에서 선형 초평면으로 분류



주: [그림 11]의 데이터를 매핑(6)를 사용하여 3차원으로 대응시켜 분류하면 선형 분류경계를 찾아낼 수 있다. 이 평면은 [그림 11]에서는 연두색 타원에 대응된다.
자료: 대신증권 리서치센터

[10] Bernhard Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

원래의 입력공간에서의 비선형 경계를 찾기 위해 이와 같이 데이터를 특징공간으로 이동시킬 경우 특징공간의 차원이 너무 커지는 문제가 발생할 수 있다. 예를 들어 n 차원 입력데이터를 위와 같이 2차항으로만 이뤄진 특징공간으로 대응시키면 ${}_nC_2$ 차원에서 선형분류를 찾게 되며, k 차 항으로만 이뤄진 특징공간으로 대응시키면 상황은 더욱 나빠져 ${}_nC_k$ 차원에서 작업하게 된다. 이와 같이 매핑을 사용하여 특징공간에서 직접 작업하게 되면 작업시간과 데이터 처리량이 급격히 증가하는 차원의 저주(curse of dimension)에 빠질 수 있다. 다행히도 커널트릭이라는 기법을 사용해 특징공간에서 간접적으로 작업할 수 있는 방법이 있다. 커널트릭을 사용하기 위해서는 원래의 최적화 문제 (5)의 쌍대문제를 풀어 해결한다.

모든 최적화 문제는 그와 연결된 쌍대문제(dual problem)가 있다. 쌍대문제는 원본문제(primal problem)의 Lagrangian 함수의 안장점(saddle point)을 찾아 구할 수 있다. 원본문제 (5)의 쌍대문제는

$$\begin{aligned} &\max_{\alpha_1,\ldots,\alpha_m} \; -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i y_i k(\mathbf{x}_i,\mathbf{x}_j) y_j \alpha_j + \sum_{i=1}^{m}\alpha_i \\ &s.t. \sum_{i=1}^{m} y_i \alpha_i = 0 \\ &\quad 0 \le \alpha_i \le C_i, \text{ for } i = 1,\ldots,m \end{aligned} \tag{7}$$

로서, $k(\mathbf{x}_i,\mathbf{x}_j) \coloneqq \langle \mathbf{x}_i,\mathbf{x}_j \rangle$ 로 정의된다. 여기서 주목할 점은 훈련데이터인 $\mathbf{x}_i$ 가 내적 형태로만 나타난다는 점이다. 이 때문에 $k(\mathbf{x}_i,\mathbf{x}_j)$ 는 유클리드 공간에서의 내적뿐만 아니라 다른 (고차원 혹은 무한 차원) 공간에서의 내적일 수 있다. 다시 말해 두 입력 데이터 $\mathbf{x}$ 와 $\bar{\mathbf{x}}$ 에 대해 커널(kernel) 연산을 수행하는 것이 어떤 매핑 Φ 를 이용해 입력 데이터를 어떤 특징공간으로 이동해서 데이터를 $\Phi(\mathbf{x})$ 와 $\Phi(\bar{\mathbf{x}})$ 로 변환해 그곳에서 어떤 내적을 수행하는 것과 같을 수 있다는 말이다. 즉 간략히 수식으로 설명하면

$$k(\mathbf{x},\bar{\mathbf{x}}) = \langle \Phi(\mathbf{x}), \Phi(\bar{\mathbf{x}}) \rangle_H$$

이다.[11] 이 경우 특징공간에서의 선형 결정 초평면이 원래의 입력 데이터 공간에서는 비선형 결정 곡면으로 나타나게 된다.

이와 같이 커널을 사용하고 쌍대문제 (7)을 풀어 비선형 결정곡면을 찾아내는 작업을 커널트릭이라 한다. 물론 직접적인 매핑과 연결된 커널도 있으며, 예를 들어

$$\left\langle \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right), \left(\bar{x}_1^2, \bar{x}_2^2, \sqrt{2}\bar{x}_1\bar{x}_2\right) \right\rangle = (x_1\bar{x}_1 + x_2\bar{x}_2)^2 = \langle \mathbf{x},\bar{\mathbf{x}} \rangle^2$$

이므로 매핑 (6)은 2차 homogeneous 다항 커널인 $k(\mathbf{x},\bar{\mathbf{x}}) \coloneqq \langle \mathbf{x},\bar{\mathbf{x}} \rangle^2$ 와 연결되어 있다.

커널 $k(\mathbf{x}_i,\mathbf{x}_j)$ 가 SPD이면 이 커널은 이 커널에 의해 유도되는 reproducing kernel Hilbert space에서의 내적임이 알려져 있다.[12] 또한 커널이 SPD인 경우 쌍대문제인 (7)의 목적함수의 이차 미분계수라 할 수 있는 헤시안(Hessian) 행렬이 SPD 행렬이 되어 목적함수는 위로볼록(Concave) 함수가 되며, 최적화 문제의 해를 구하기 쉽다. 이러한 성질을 갖는 커널은 여러 가지가 있으며, 대표적인 것 중 하나가 본 리포트에서 사용한 가우시언(Gaussian) 커널로 $k(\mathbf{x},\bar{\mathbf{x}}) = \exp(-\sigma\|\mathbf{x}-\bar{\mathbf{x}}\|^2)$ 와 같이 정의된다.

[11] 아래첨자 H 를 내적기호에 붙인 것은 커널과 연관된 어떤 공간에서의 내적임을 표시하기 위해서다.

[12] Chris Burges. *A tutorial on support vector machines for pattern recognition*. Data Mining and Knowledge Discovery, 2(2):121-167, 1998.