참고 설명

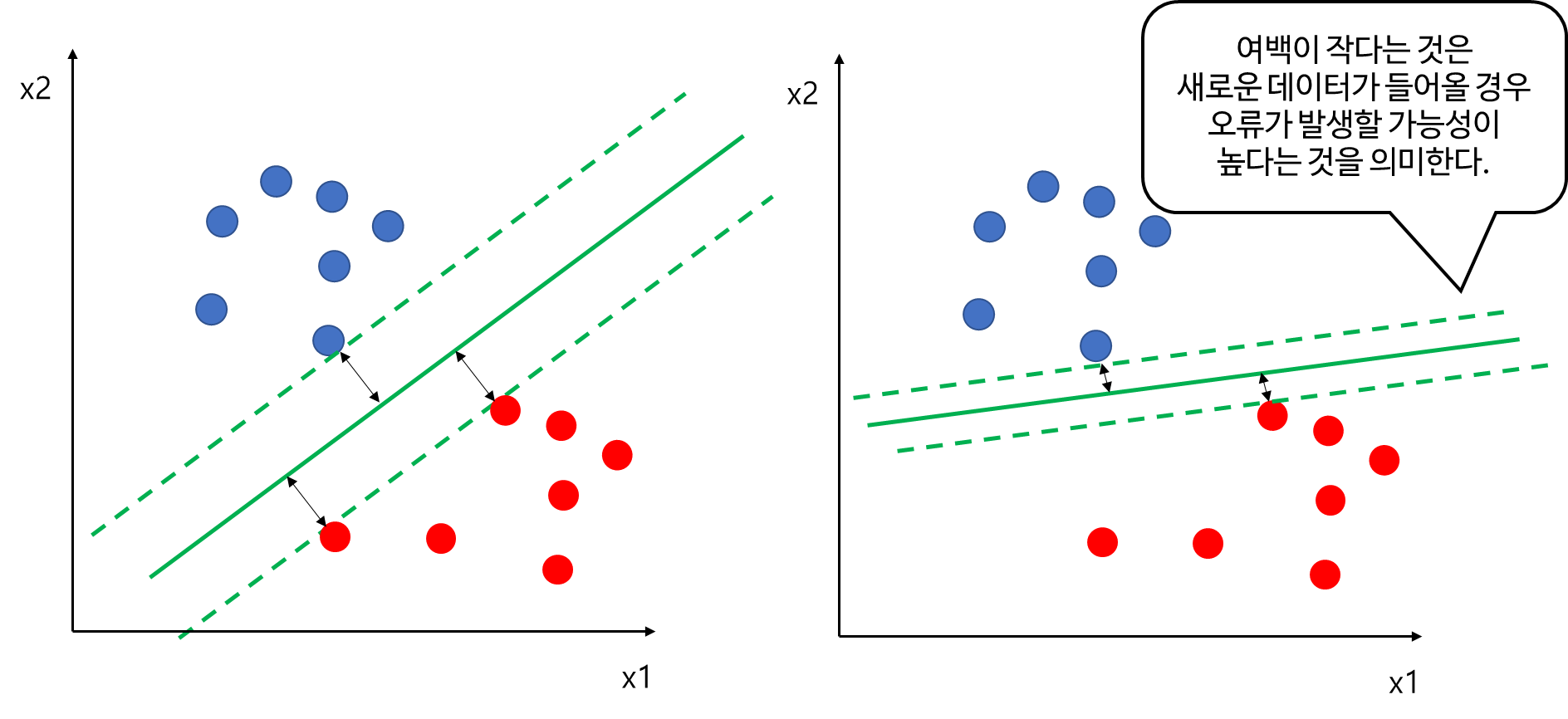
서포트 벡터 머신(SVM : Support Vector Machine)이란 두 클래스로부터 최대한 멀리 떨어져 있는 **결정 경계**를 찾는 분류기로 특정 조건을 만족하는 동시에 클래스를 분류하는 것이 목표

* 마진을 최대화하는 결정 경계를 찾는 것이 목표이기때문에 여러가지 결정 경계가 있다면 서포트벡터와의 거리를 최대화 하는 것이 중요하다. (단, 허용 가능한 오류 범위 내에서)

왜냐하면 분류 결과가 같더라도 서포트벡터와의 마진이 가장 큰 결정 경계를 찾는 것이 가장 일반화하는데 효율적이다

마진 : 데이터가 오류를 발생하지 않으면서 가질 수 있는 최대 공간

초평면or 결정경계 : 데이터들을 분류하는 기준선

서포트 벡터 : 결정 경계에 가장 가까운 각 클래스의 점들

발표 내용

다음은 서포트벡터머신입니다.

**서포트벡터머신 소개page**

서포트 벡터머신은 결정경계 또는 초평면이라 불리는 기준선을 경계로 분류를 수행하는 모델입니다.

사진에 보이는 것처럼 서포트벡터머신 모델링 전에 스케일링은 인스턴스간의 사이를 넓혀 결정경계를 더 뚜렷하게 만들기 때문에 성능 개선을 위해선 스케일링이 거의 필수적입니다.

결정경계를 설정하는 하이퍼파라미터를 말씀드리겠습니다.

kernel은 분류방법을 결정하는 하이퍼 파라미터로 linear,poly,RBF로 설정할 수 있습니다.

C는 결정경계와 서포트벡터 사이의 마진을 설정하는 값이고, 높게 설정하면 오류를 거의 허용하지 않는 하드마진, 낮게 설정하면 오류를 허용하는 소프트마진이 됩니다.

gamma는 결정경계의 곡률을 설정하는 값이고, 수치값이나 auto, scale로 설정할 수 있습니다.

주요 장점은 회귀와 분류 모델에 모두 사용가능하고, 과적합 가능성이 낮다는 것입니다.

주요 단점은 데이터 스케일링에 민감하고, 피처 수와 데이터 수가 많으면 학습속도가 매우 느리다는 점입니다.

**스케일링 별 기본 모델링 page**

다음은 scaling하지 않은 원본데이터와standard, min-max, log scaling을 진행한 데이터들을 디폴트 하이퍼파라미터(C=1,Gamma=scale, kernel=RBF)로 진행한 모델의 성능입니다. 재현율을 제외하고는 standard scaling의 모델이 가장 좋은 성능을 냈습니다. 그러므로 이 데이터에서 서포트벡터 사이의 거리를 넓혀주는 가장 적합한scaling방법은 standard scaling으로 보입니다.

**최적 하이퍼파라미터 및 예측 page**

최적 하이퍼파라미터를 탐색한 결과

**standard scaling의 하이퍼 파라미터는**(C=3, gamma=auto, kernel=RBF),

**Min-Max scaling의 하이퍼 파라미터는** (C=50, gamma=scale, kernel=RBF),

**Log** **Scaling의 하이퍼 파라미터는**(C=50,gamma=auto,kernel=RBF)였고,

각 모델을 최적 하이퍼파라미터로 진행한 결과 또한 대체로 standard scaling이 높았으며, 테스트데이터를 예측한 결과 과적합되지 않았음을 확인했습니다.

**모델 별 평가지표 시각화page**

그래프에 보이시는 대로 **정확도, 정밀도, F1 Score, ROCAUC score**는 최적 하이퍼파라미터를 적용한 **standard scaling**의 모델이 가장 성능이 좋았고, **재현율**은 디폴트 하이퍼파라미터를 적용한 **Log scaling** 모델이 가장 높았습니다.

**최종 모델 시각화page**

**첫번째**는base모델인 원본데이터에서 학습을 진행한 결과와 최종모델(Standard scaling) 모델의 성능지표 비교 표입니다. 스케일링과 최적 하이퍼파라미터로 학습을 진행한 결과 전체적으로 평가지표가 상승하였고, 그 중 재현율의 증가율은 21.1%로 크게 상승했습니다.

**두번째**는 최종모델을 이용해 테스트데이터의 confusion matrix를 시각화한 자료입니다.

총 11139개 데이터 중 8824개의 데이터를 제대로 예측해 **정확도**는 79.2%

흡연자로 예측한 데이터 4414개 중 실제 흡연자인 데이터는 3101개로 **정밀도**는 70.2%

실제 흡연자 4103개의 데이터 중 흡연자로 예측한 데이터는 3101개로 **재현율**은 75.6%가 나왔습니다.