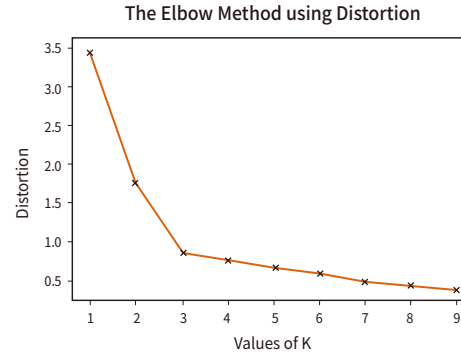
■ 군집 분석 평가 방법

<table>
<tr><td rowspan="2">외부 평가</td><td>자카드계수 평가</td><td>두 데이터 군집 간의 유사도를 계산
$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP+FP+FN}$</td></tr>
<tr><td>분류 모형 평가 방법을 응용</td><td>▪ 혼동행렬(confusion matrix)
▪ ROC 곡선(군집 분석 평가에 분류 평가 방법을 사용)</td></tr>
<tr><td rowspan="3">내부 평가</td><td>단순 계산법</td><td>전체 데이터의 개수가 n개인 경우, 군집의 개수인 K값은 $\sqrt{\frac{n}{2}}$로 계산</td></tr>
<tr><td>군집 간의 거리를 계산하여 평가</td><td>▪ 유클리드 거리(Euclidean)
▪ 맨해튼 거리(Manhattan)
▪ 민코프스키 거리(Minkowski)
▪ 표준화 거리(Standardized)
▪ 마할라노비스 거리(Mahalanobis)
▪ 캔버라 거리(Canberra)
▪ 체비셰프 거리(Chebychev)</td></tr>
<tr><td>엘보 메소드</td><td>▪ K-평균 분석 시각화
 
</td></tr>
</table>

■ 분류분석 평가지표

• 혼동행렬

실제

예측	클래스	Positive	Negative
	Positive	True Positive(TP)	False Positive(FP)
	Negative	False Negative(FN)	True Negative(TN)

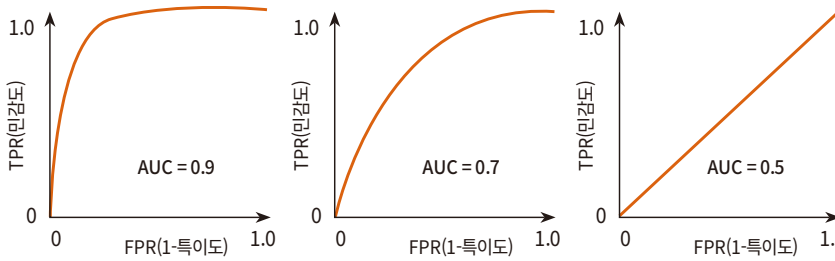
구분	의미
TP(True Positive)	예측한 값이 Positive이고 실제 값도 Positive인 경우
FP(False Positive)	예측한 값이 Positive이고 실제 값은 Negative인 경우
TN(True Negative)	예측한 값이 Negative이고 실제 값도 Negative인 경우
FN(False Negative)	예측한 값이 Negative이고 실제 값은 Positive인 경우

• 혼동행렬 평가 지표

평가지표	계산식	의미
정확도(Accuracy)	$\frac{TP+TN}{TP+TN+FP+FN}$	전체 데이터에서 올바르게 분류한 데이터의 비율
정밀도(Precision)	$\frac{TP}{TP+FP}$	Positive로 예측한 것 중에서 실제 값이 Positive인 비율
재현율(Recall), 민감도(Sensitivity), 참 긍정율(TPR, True Positive Rate)	$\frac{TP}{TP+FN}$	실제 Positive인 값 중 Positive로 분류한 비율
특이도(Specificity), 참 부정율(TNR, True Negative Rate)	$\frac{TN}{TN+FP}$	실제 Negative인 값 중 Negative로 분류한 비율
거짓 긍정률 FPR(False Positive Rate)	$1-\frac{TN}{TN+FP}=\frac{FP}{TN+FP}$	실제 Negative인 값 중 Positive로 잘못 분류한 비율(1–Specificity)
F1–스코어	$2\times\frac{Precision\times Recall}{Precision+Recall}$	정밀도와 재현율의 조화평균으로, 정밀도와 재현율 중 한쪽만 클 때보다 두 값이 골고루 클 때 큰 값이 된다.

- **ROC 곡선**

ROC 곡선(Receiver Operating Characteristic Curve)은 임곗값을 다양하게 조절해 분류 모형의 성능을 비교할 수 있는 그래프로, 서로 반비례 관계에 있는 민감도(TPR)를 y축에 두고, 거짓 긍정률(FPR)을 x축에 두어 시각화한 것이다.



■ 회귀 평가지표

평가지표	수식	오차 상쇄 처리	이상치
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert$	절댓값	유리
MSE	$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	제곱	불리
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$	제곱	불리
MAPE	$100\cdot\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i-\hat{y}_i}{y_i}\right\rvert$	절댓값	유리

■ 분석기법별로 활용되는 성능 평가 기준

데이터 마이닝	시뮬레이션
정확도(Accuracy)	throughput
정밀도(Precision)	average waiting time
검출률(detect rate) 재현율(Recall), 민감도(Sensitivity)	average queue length
향상도(lift)	time in system

■ 교차검증

학습 데이터셋과 테스트 데이터셋으로 단순히 한 번 나눈다면 테스트 셋이 어떻게 샘플링되느냐에 따라 점수가 크게 달라질 수 있다. 교차검증(cross validation)은 데이터를 나누고 학습하는 과정을 여러 차례 반복함으로써 일반화 성능을 평가한다.

- **k-폴드 교차검증(k-fold cross validation)**: 데이터를 k개의 폴드(fold)라고 불리는 파티션으로 나누어 k-1개의 폴드는 학습용으로 나머지는 검증용으로 k번 학습하여 얻은 결과들의 평균값으로 일반화 성능을 평가하는 방법이다.
- **홀드아웃(Holdout)**: 가장 단순한 종류의 교차검증 방법으로, 데이터를 랜덤으로 추출해 학습 데이터와 테스트 데이터로 나눈다.
- **리브-p-아웃 교차 검증(Leave-p-out cross validation, LpOCV)**: p개의 관측치만 검증용으로 사용되고 나머지 관측치는 모두 모형을 학습하는 데 사용한다.

부트스트랩(Bootstrap)

부트스트랩은 한 번 추출한 표본을 다시 모집단에 넣어 또 다른 표본을 추출하는 '단순 랜덤 복원추출법'을 활용하여 동일한 크기의 표본을 여러 개 생성하는 샘플링 방법의 하나다. 표본의 중복을 허용하는 방법으로 '무작위 추출 방법'이라고도 부른다.

■ 모수 검정과 비모수 검정 비교

	모수 검정	비모수 검정
가설 설정	가정된 분포의 모수(모평균, 모분산)에 대해 가설 설정	분포의 형태에 대한 가설 설정 (분포의 형태가 동일함, 분포의 형태가 동일하지 않음)
검정 방법	표본평균, 표본분산	(관측값의 절대적 크기에 의존하지 않고 이상치의 영향이 적은) 순위, 부호
검정력	비교적 강함	비교적 약함

비모수 검정 예시

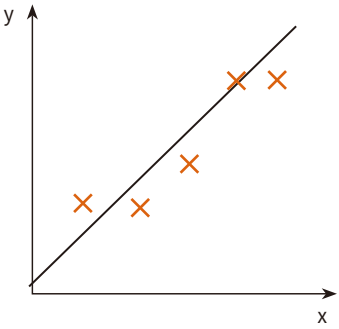
부호검정(sign test), 윌콕슨의 부호순위합검정(Wilcoxon signed rank test), 윌콕슨의 순위합검정(rank sum test), 스피어만의 순위상관계수, 런검정(run test), 만–위트니의 U검정

■ 적합도 검정

적합도(Goodness–of–Fit)는 실험에서 얻은 결과가 이론 분포와 일치하는 정도를 의미한다. 즉, 적합도 검정(Goodness of fit test)은 데이터가 특정 이론 분포를 따르는지를 검정하는 것이다.

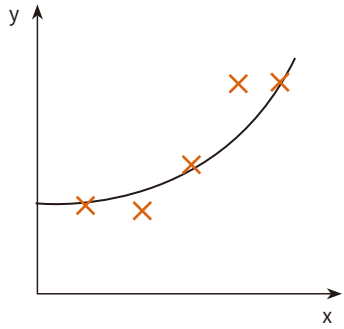
- **카이제곱 검정(Chi–Squared Test)**: 범주형 데이터를 대상으로 관측된 값들의 빈도수와 기대 빈도수가 의미 있게 다른지를 비교한다.
- **샤피로 윌크 검정(shapiro–wilk normality test)**: 데이터가 정규분포로부터 추출된 표본인지 검정한다.
- **콜모고로프 스미르노프 검정(Kolmogorov–Smirnov Test)**: 데이터의 누적분포함수와 임의 분포의 누적분포함수 간의 최대 차이 D를 검정통계량으로 하는 비모수 검정 기법이다.
- **Q–Q 플롯**: 그래픽적으로 데이터의 정규성을 확인하는 가장 간단한 방법이다. 대각선 참조선을 따라 값들이 분포하면 정규성을 만족한다고 판단할 수 있다.

■ 과소적합, 일반화된 모형, 과적합

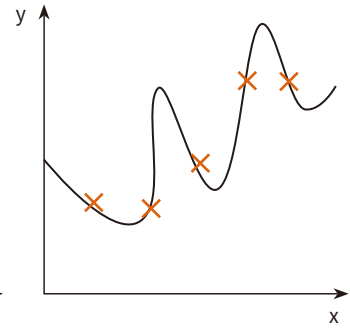


과소적합

지나치게 단순하고
설명력이 부족한 모형



일반화된 모형



과적합

지나치게 복잡한 새로운
데이터에 대해 일반화가
어려운 모형

■ 과적합을 방지하는 방법

① **학습 데이터를 더 많이 확보**: 더 많은 학습 데이터를 학습해 일반화 성능을 높일 수 있다.

② **교차검증**: 데이터를 나누고 학습하는 과정을 여러 차례 반복하는 교차검증은 검증용 데이터셋이 고정되어 있지 않아 일반화 성능을 높일 수 있다.

③ **피처의 수를 줄인다**: 중요도가 낮은 피처를 제거해 일반화 성능을 높일 수 있다. 하지만 제거된 피처의 정보가 손실된다는 단점이 있다.

④ **정규화**: 손실함수에 페널티를 부과해 과하게 적합하지 않게 규제를 가해 각각의 영향력을 줄이고 모형을 단순화해 일반화 성능을 높일 수 있다. 가중치 규제에는 모든 가중치의 절댓값 합계를 비용함수에 추가하는 L1 규제와 모든 가중치의 제곱의 합을 비용함수에 추가하는 L2 규제가 있다.

■ 매개변수 최적화와 경사하강법

최적화(optimization)는 일반적으로 손실함수(loss function)의 값을 최소화하는 분석 모형의 매개변수를 찾는 것을 의미하며, 경사하강법(gradient descent)은 최적화를 수행하기 위해 사용되는 대표적인 알고리즘 중 하나다.

- **경사하강법**: 현재 위치에서 기울기(그레이디언트, gradient)를 구해 함수의 값이 급격히 감소하는 방향으로 매개변수 값을 조정하는 것을 반복하여 전역 최솟값(global minimum)을 찾아 나가는 것
- **확률적 경사하강법(SGD)**: 무작위로 샘플링된 하나의 샘플로 그레이디언트를 계산하고 매개변수를 업데이트한다. 즉, 배치(batch)의 크기가 1인 경사하강법으로 볼 수 있다.
- **미니 배치 확률적 경사하강법(BGD)**: 미니 배치 확률적 경사하강법: 한번 매개변수를 업데이트 할 때마다 전체 데이터셋을 사용하는 방식인 BGD와 SGD 알고리즘의 장점을 모두 취하면서 매개변수 업데이트를 수행하는 알고리즘이다.
- **모멘텀(Momentum)**: 물리학의 '운동량'에서 유래된 단어로 SGD가 가는 방향에 가속도를 부여해주는 알고리즘이다. 결과적으로 진동이 감소하고 더 빠르게 학습을 진행할 수 있다.
- **AdaGrad**: 매개변수별 적응 학습률(adaptive learning rate)을 사용하는 알고리즘으로 업데이트가 빈번히 수행된 매개변수들은 낮은 학습률로 조정하고 그렇지 않은 매개변수들은 학습률을 크게 조정한다.
- **Adam(Adaptive Moment Estimation)**: 오래된 기울기의 영향력을 지수적으로 줄여 AdaGrad를 개선한 RMSProp에 모멘텀을 더한 것으로, 최근 가장 많이 사용되는 알고리즘이다.

■ 앙상블 기법

앙상블(ensemble)은 주어진 데이터에서 여러 개의 분석 모형을 만들고 각 학습 결과를 결합하여 예측 모형을 구축하는 기법으로, 단일 예측 모형보다 더 나은 일반화 성능을 갖는 것을 목표로 한다. 앙상블 기법의 종류에는 보팅(voting), 배깅(bagging), 부스팅(boosting) 등이 있다.

- **보팅**: 서로 다른 알고리즘을 사용한 여러 분석 모형의 결과를 두고 투표를 통해 최종 예측 결과를 결정한다.
- **배깅(Bootstrap aggregating)**: 간단하고 강력한 앙상블 기법이다. 먼저, 분석 모형의 수만큼 부트스트랩 데이터를 만들고 같은 알고리즘으로 병렬적으로 학습한 각각의 학습 결과를 결합해 최종 예측 모형을 만든다.
- **랜덤 포레스트**: 의사결정 트리를 개별 모형으로 사용하는 모형 결합 방법으로, 독립변수의 차원을 랜덤하게 감소시킨 다음 그중에서 독립변수를 선택하는 방법이다.
- **부스팅**: 여러 개의 연결된 약한 분석 모형(weak learner)을 순차적으로 학습하며 맞추지 못한 부분에 가중치를 부여함으로써 하나의 강한 분석 모형(strong learner)으로 만드는 앙상블 기법이다.