**2025년 1학기 머신러닝1 과목 프로젝트 서술형 문제 답안지**

**Project 1 - 와인 분류 최적화 프로젝트**

**2단계-2번. 정규화 전/후 KNN 성능 차이가 크게 발생한 이유를 거리 기반 알고리즘의 특성과 연관 지어 설명하세요.**

**답안:**  
KNN에서 정규화 전후 성능 차이가 크게 발생한 이유는 다음과 같습니다:

1. **스케일 차이로 인한 거리 계산 왜곡**: 와인 데이터셋의 각 특성은 서로 다른 범위를 가집니다. 예를 들어, 'proline' 특성은 수백에서 수천의 값을 가지는 반면, 'hue' 특성은 0-2 범위의 값을 가집니다.
2. **거리 기반 알고리즘의 한계**: KNN은 유클리드 거리를 기반으로 가장 가까운 이웃을 찾습니다. 정규화하지 않으면 큰 스케일의 특성(proline, color\_intensity 등)이 거리 계산을 지배하게 되어, 상대적으로 작은 스케일의 특성들(hue, nonflavanoid\_phenols 등)의 영향이 무시됩니다.
3. **실험 결과**: 정규화 전 테스트 정확도 72.22%에서 정규화 후 94.44%로 약 22%포인트 향상되었습니다. 이는 StandardScaler를 통해 모든 특성을 평균 0, 표준편차 1로 표준화하여 각 특성이 동등한 가중치로 거리 계산에 기여할 수 있게 되었기 때문입니다.

**3단계-2번. Best parameter를 기준으로 훈련 정확도와 테스트 정확도의 차이를 bias-variance tradeoff 관점에서 정량적으로 분석하세요.**

**답안:**  
최적 파라미터 {'max\_depth': 3, 'min\_impurity\_decrease': 0.0, 'min\_samples\_split': 2}를 가진 결정트리 모델의 성능 분석:

* **훈련 정확도**: 99.30%
* **테스트 정확도**: 94.44%
* **정확도 차이**: 4.85%

**Bias-Variance Tradeoff 분석**:  
주어진 조건에 따른 분석:

* 훈련과 테스트 차이(4.85%) < 0.2 (20%) → 분산이 높지 않음
* 훈련 정확도(99.30%) ≥ 0.9 (90%) → 편향이 높지 않음
* 훈련과 테스트 차이(4.85%) < 0.05 (5%) → 조건을 만족하지 않음(훈련 정확도가 0.9보다 크기 때문)

**결론**: 좋은 일반화 성능을 가지고 있습니다. max\_depth=3으로 제한하여 모델 복잡도를 적절히 조절하고, 과적합을 방지하면서도 충분한 학습 능력을 유지하여 편향과 분산 사이의 균형을 잘 맞췄습니다.

**5단계-1번. 성능 측면에서 가장 우수한 모델은 무엇인가요? 그 이유는 무엇인가요?**

**답안:**  
성능 측면에서 가장 우수한 모델은 **Random Forest**입니다.

**이유**:

1. **테스트 정확도**: 100%로 모든 모델 중 최고 성능을 보였습니다.
2. **과적합 없음**: 훈련 정확도와 테스트 정확도 차이가 0%로 완벽한 일반화를 보입니다.
3. **교차 검증 성능**: 평균 97.86%로 높은 정확도를 유지하였으며, 각 fold 간 편차도 상대적으로 작아 예측 성능이 안정적이었습니다.
4. **앙상블 효과**: 다수의 결정트리를 결합함으로써 개별 트리의 과적합 문제를 해결하고 예측 안정성이 향상되는 효과를 가져왔습니다.

**5단계-2번. 각 모델의 과대적합 여부를 훈련 데이터와 테스트 데이터 정확도 차이로 판단해보세요.**

**답안:**  
훈련-테스트 정확도 차이 분석:

1. **KNN (정규화 후)**: 차이 4.15% → 경미한 과적합
2. **Decision Tree**: 차이 4.85% → 경미한 과적합
3. **Random Forest**: 차이 0% → 과적합 없음 (이상적)
4. **Gradient Boosting**: 차이 5.56% → 다른 모델에 비해 과적합 경향이 상대적으로 가장 큽니다.

**결론**: Random Forest가 가장 균형 잡힌 성능을 보이며, Gradient Boosting이 상대적으로 과적합 위험이 가장 높습니다.

**5단계-3번. 세 모델((1) 결정 트리, (2) 랜덤 포레스트, (3) 그래디언트 부스팅)의 학습 방식, 예측 안정성, 과적합 위험성의 측면에서 구조적 차이를 설명하세요.**

**답안:**

**1. 결정 트리**:

* 학습 방식: 단일 트리로 특성을 순차적으로 분할하여 결정 경계 생성
* 예측 안정성: 낮음 (작은 데이터 변화에 민감)
* 과적합 위험성: 높음 (depth 제한 없으면 훈련 데이터를 완벽하게 반영함)

**2. 랜덤 포레스트**:

* 학습 방식: 배깅(Bootstrap Aggregating) + 특성 무작위 선택으로 다수 트리 학습 후 투표
* 예측 안정성: 높음 (앙상블을 통한 분산 감소)
* 과적합 위험성: 낮음 (개별 트리 과적합이 평균화로 상쇄)

**3. 그래디언트 부스팅**:

* 학습 방식: 이전 학습기의 잔차를 다음 학습기가 학습하고 학습기를 계속 추가해 가면서 잔차를 줄여가는 방식이전 모델의 오차를 보정
* 예측 안정성: 중간 (learning rate와 n\_estimators에 민감)
* 과적합 위험성: 중간-높음 (잘못된 하이퍼파라미터로 점진적 과적합 가능)

**5단계-4번. 최종적으로 어떤 모델을 선택하겠습니까? 선택 이유는 무엇인가요?**

**답안:**  
최종 선택 모델: **Random Forest**

**선택 이유**:

1. **데이터 분포**: 와인 데이터는 13개 특성으로 구성된 중간 규모 데이터셋으로, Random Forest의 특성인 무작위 선택이 효과적으로 작동합니다.
2. **정규화 여부**: Random Forest는 트리 기반 모델로 스케일에 불변하여 정규화 없이도 우수한 성능을 보입니다.
3. **모델 복잡도**: 적절한 복잡도로 편향과 분산의 균형이 잘 맞춰졌습니다. 개별 트리의 높은 분산을 앙상블로 효과적으로 감소시켰습니다.
4. **앙상블 효과**:
   1. 배깅을 통해 분산이 감소하였습니다.
   2. 특성 무작위 선택으로 다양성이 확보되었습니다.
   3. 개별 모델 오류를 상호 보완시키는 효과를 보였습니다.
5. **실증적 성능**: 테스트 정확도 100%, 과적합이 관찰되지 않았으며, 교차 검증에서 안정적인 성능으로 실용성과 신뢰성을 동시에 확보하였습니다.

따라서 Random Forest 모델은 와인 데이터셋에서 실험한 세 종류의 모델 중 가장 적절한 모델입니다.

**Project 2 - 당뇨병 진행도 예측 최적화 프로젝트**

**4단계-3번. 두 모델의 과대적합 가능성 여부를 두 모델의 구조 차이 (학습 방식, 예측 방식, 하이퍼파라미터 민감도)를 통해 비교 서술하세요.**

**답안:**

**Random Forest Regressor**:

* **학습 방식**: 독립적인 다수 트리를 병렬 학습
* **예측 방식**: 모든 트리 예측값의 평균
* **하이퍼파라미터 민감도**: 상대적으로 낮음
* **과적합 가능성**: 비교적 큼 (훈련-테스트 R² 차이: 0.28 차이)

**Gradient Boosting Regressor**:

* **학습 방식**: 순차적으로 이전 모델의 오차를 보정하는 트리 추가
* **예측 방식**: 모든 트리의 가중합
* **하이퍼파라미터 민감도**: 높음 (learning\_rate, n\_estimators에 매우 민감)
* **과적합 가능성**: 비교적 낮음 (훈련-테스트 R² 차이: 0.17)

**결론**: Random Forest는 모델 복잡도 제약 없이 트리를 많이 학습하다 보니 과적합이 발생했으며, 이는 작은 데이터셋에서 각 트리가 과도하게 복잡해졌기 때문입니다. Gradient Boosting은 상대적으로 나은 일반화 성능을 보였습니다.

**5단계-1번. Ridge, Lasso, Random Forest, Gradient Boosting 중 가장 성능이 우수한 모델을 선택하고, 이유를 R², MSE, 해석 가능성 측면에서 설명하세요.**

**답안:**

실험 결과를 바탕으로 각 모델(Lasso, Random Forest, Gradient Boosting)의 성능을 표로 정리한 결과는 다음과 같습니다.

****

**R² (결정계수) 측면 분석**

테스트 R² 성능 순위에서 Lasso (기본, α=1.0)과 Lasso (Grid Search)가 0.4669로 실험을 수행한 모델 중 가장 높은 예측 성능을 보였습니다. Lasso (Random Search)가 0.4641, Gradient Boosting이 0.4597, Random Forest가 0.4544 순으로 뒤를 이었습니다.

과적합 정도 분석에서 훈련 R²와 테스트 R² 간의 차이를 통해 일반화 성능을 평가한 결과, Lasso 모델들은 0.0556~0.0595의 매우 낮은 과적합을 보인 반면, Random Forest는 0.2794의 상대적으로 심각한 과적합, Gradient Boosting은 0.1733의 중간 수준 과적합을 나타냈습니다.

**MSE (평균제곱오차) 측면 분석**

테스트 MSE 성능에서도 Lasso 모델들이 우수한 결과를 보였습니다. Lasso (기본, α=1.0)과 Grid Search 버전이 2824.57로 최저 MSE를 기록했으며, Random Search 버전이 2839.29, Gradient Boosting이 2862.84, Random Forest가 2890.81 순으로 나타났습니다.

MSE 관점에서 특히 주목할 점은 앙상블 모델들이 훈련 데이터에서는 매우 낮은 MSE를 보였으나(Random Forest 1617.33, Gradient Boosting 2230.03), 테스트 데이터에서는 오히려 Lasso보다 높은 MSE를 기록하여 과적합 현상을 명확히 보여주었습니다.

**Ridge 회귀의 일반적 특성 분석**

Ridge 회귀는 L2 정규화를 사용하는 선형 회귀 모델로, 회귀 계수의 제곱합에 페널티를 부여하여 과적합을 방지합니다. Lasso와 달리 Ridge는 계수를 완전히 0으로 만들지 않고 축소시키는 특징이 있어, 모든 특성이 반영되는 데이터셋에서 안정적인 성능을 보입니다.

일반적으로 당뇨병과 같은 의료 데이터에서 Ridge 회귀는 다중공선성 문제가 있을 때 특히 효과적이며, Lasso와 유사한 R² 성능(0.46~0.47 범위)을 보일 것으로 예상되나 특성 선택 능력은 제한적이라 볼 수 있습니다. Ridge는 모든 특성을 유지하면서 계수를 축소시키 때문에 Lasso보다 해석이 어려운 특성을 가집니다.

**해석 가능성 측면 비교**

Lasso 회귀는 L1 정규화로 인한 자동 특성 선택 기능을 제공하여 불필요한 특성의 계수를 정확히 0으로 만들어 모델의 해석 가능성을 크게 향상시킵니다. 당뇨병 진행도 예측에서 10개의 생리학적 특성 중 실제로 중요한 핵심 요인들만을 선택적으로 활용할 수 있어 의료진의 임상적 의사결정에 직접적으로 도움을 줍니다.

Ridge 회귀는 모든 특성을 유지하면서 계수를 축소시키므로 Lasso보다 해석이 어렵고, 앙상블 모델들(Random Forest, Gradient Boosting)은 복잡한 트리 구조와 앙상블 메커니즘으로 인해 해석 가능성이 현저히 낮아 의료 분야에서 활용하기 어렵습니다.

**5단계-2번. 모델의 일반화 성능, 해석 가능성, 계산 효율성 측면에서의 장단점을 서술하세요.**

**답안:**

**일반화 성능 측면**

Lasso Regression의 **뛰어난 일반화 성능**은 실험 결과에서 명확히 입증되었습니다. 훈련 R² 0.5225와 테스트 R² 0.4669 간의 차이가 0.0556에 불과하여 새로운 데이터에 대한 예측 신뢰성이 매우 높습니다. 이는 L1 정규화가 모델 복잡도를 효과적으로 제어하여 훈련 데이터의 노이즈에 과도하게 적응하지 않고 일반적인 패턴을 학습했음을 의미합니다.

**장점:**

* 정규화 효과로 인한 안정적인 예측 성능
* 새로운 환자 데이터에 대한 높은 신뢰도
* 실제 임상 환경에서의 실용적 활용 가능성
* 과적합 최소화로 인한 견고한 모델 성능

**단점:**

* α 하이퍼파라미터에 대한 민감도가 존재합니다(다만, α=1.0이 안정적인 최적값으로 확인되어, 본 실험에 사용된 당뇨병 환자의 임상 데이터에서는 제한적).
* 선형 관계만 학습 가능하여 복잡한 비선형적 데이터 및 패턴에서는 성능이 저하될 수 있습니다.

**해석 가능성 측면**

Lasso Regression의 좋은 해석 가능성은 의료 분야에서 가장 중요한 장점입니다. L1 정규화로 인해 중요하지 않은 특성의 계수가 정확히 0이 되어 자동적인 특성 선택이 이루어지므로, 의료진이 어떤 생리학적 지표들이 당뇨병 진행에 가장 큰 영향을 미치는지 명확히 파악할 수 있습니다.

**장점:**

* 선형 모델의 단순성으로 인해 상대적으로 투명한 의사결정 과정
* 각 특성의 기여도를 정량적으로 해석 가능
* 의료 AI 시스템에서 요구되는 설명 가능성 충족
* 자동 특성 선택을 통한 핵심 위험 인자를 상대적으로 쉽게 식별 가능
* 임상 가이드라인 개발 및 환자 상담에 직접 활용 용이

**단점:**

* 비선형 관계를 포착하지 못하는 선형 모델의 한계
* 각 특성 간 상호작용 효과 및 특성 간 관계를 직접적으로 모델링하지 못함
* α 값에 따라 선택되는 특성이 달라질 수 있는 잠재적 불안정성 (다만, 실험을 수행한 당뇨 데이터에서 최적 α=1.0에서는 안정적)

**계산 효율성 측면**

Lasso Regression의 우수한 계산 효율성은 실제 의료 현장에서의 활용도를 높입니다. 선형 모델의 단순한 구조로 인해 훈련 시간이 매우 짧고, 예측 과정도 즉시 수행 가능하여 실시간 환자 모니터링 시스템에 적합합니다.

**장점:**

* 빠른 훈련 속도로 대용량 의료 데이터 처리 가능
* 실시간 예측을 통한 즉각적인 임상 지원
* 제한된 컴퓨팅 자원 환경에서도 안정적 동작
* 메모리 사용량이 적어 모바일 기기나 임베디드 시스템과 같이 컴퓨팅 성능이 제한적인 환경에서도 활용 가능
* 모델 업데이트 시 빠른 재학습 가능

**단점:**

* Coordinate descent 알고리즘의 수렴 특성상 매우 작은 α 값에서 수렴 시간이 길어질 수 있음 (다만, 최적 α=1.0에서는 문제없음)
* 특성 스케일링 전처리 과정이 필수적으로 요구됨
* 대규모 특성을 가진 데이터에서는 정규화 계산이 상대적으로 비효율적일 수 있음

**종합 평가**

Lasso Regression (기본, α=1.0)은 당뇨병 진행도 예측 과제에서 예측 성능, 일반화 능력, 해석 가능성, 계산 효율성 모든 측면에서 균형 잡힌 우수한 성능을 보여주었습니다. 특히 의료 분야의 특수성을 고려할 때, 복잡한 앙상블 모델보다 해석 가능하고 안정적인 선형 모델이 더 적합하다는 점을 실증적으로 확인했습니다.

실제 임상 환경에서 요구되는 투명성, 신뢰성, 효율성을 모두 만족하는 최적의 모델로 평가됩니다.