족보 비교

유튜브에서 비디오를 시청하면서 여러분이 부여한 좋음과 싫음에 대한 기록을 이용하여 만족도가 높은 개인화된 비디오를 추천하도록 학습하고자 한다. 다음 중 Mitchell의 머신러닝의 정의에서 과업

T에 해당하는 행동은 무엇인가?

- 여러분이 비디오를 좋음 또는 싫음으로 라벨링하는 것을 관찰함
- 비디오를 좋음 또는 싫음으로 분류함
- 비디오를 추천 또는 비추천으로 표시함
- 좋음 또는 싫음으로 정확하게 분류한 비디오의 수

답변

답안: 3) 비디오를 추천 또는 비추천으로 표시함

이유

Mitchell의 머신러닝 정의에서 과업

T는 머신러닝 시스템이 수행해야 하는 행동을 의미합니다. 여기서는 유튜브가 사용자가 좋음과 싫음을 기록한 데이터를 바탕으로 사용자가 좋아할 만한 비디오를 추천하는 것이 목표이므로,

T에 해당하는 행동은 "비디오를 추천 또는 비추천으로 표시함"입니다.

경사 하강에 관한 다음 설명 중 올바른 것을 모두 고르시오

1. 학습률은 경사 하강으로 최소값에 도달하는 속도를 조절한다
2. 경사 하강의 가장 좋은 시나리오는 국소 최소값에 도달하는 것이다
3. 비용 함수의 전역 최소값에 해당하는 파라미터가 가장 좋은 파라미터이다
4. 파라미터 값을 업데이트할 때, 변화량을 적용하기 전에, 계산된 값을 저장하기 위한 임시 변수가 필요하다

답변

답안: 1) 학습률은 경사 하강으로 최소값에 도달하는 속도를 조절한다, 3) 비용 함수의 전역 최소값에 해당하는 파라미터가 가장 좋은 파라미터이다

이유

경사 하강법에서 학습률은 경사 하강의 속도를 조절하는 역할을 합니다. 학습률이 너무 크면 발산하고, 너무 작으면 수렴 속도가 느려집니다. 비용 함수의 전역 최소값에 해당하는 파라미터가 모델의 성능을 최적화하는 가장 좋은 파라미터임을 나타내고 있습니다. 국소 최소값에 도달하는 것은 최선의 시나리오가 아닙니다. 경사 하강법에서 필요한 임시 변수가 아니라, 각 파라미터 업데이트 과정에서 직접적으로 변화량을 적용합니다.

특정 값 스케일링에 관한 다음 설명의 빈칸에 들어갈 가장 적절한 단어를 고르시오

"특징 값의 갯수가 (___)이고, 특징들이 (___) 범위의 값을 가지고 있다."

1. 하나, 비슷한
2. 여러 개, 다른
3. 여러 개, 비슷한
4. 하나, 다른

답변

답안: 3) 여러 개, 비슷한

이유

특징 값 스케일링은 여러 개의 특징 값을 비슷한 범위로 조정하는 것을 의미합니다. 스케일링의 목적은 모델 학습 과정에서 모든 특징들이 비슷한 영향을 미치도록 하기 위해서입니다. 따라서 빈칸에 들어갈 단어는 "여러 개"와 "비슷한"이 맞습니다.

로지스틱 회귀에 관한 다음 설명 중 올바른 것을 모두 고르시오

1. 로지스틱 회귀의 분류 경계선은 직선이나 비선형 곡선 모두 가능하다
2. 로지스틱 회귀의 분류 경계선은 데이터를 잘 못 분류하지 않는다
3. 로지스틱 회귀의 분류 경계선은 언제나 직선이다
4. 로지스틱 회귀의 분류 경계선은 언제나 비선형 곡선이다

답변

답안: 1) 로지스틱 회귀의 분류 경계선은 직선이나 비선형 곡선 모두 가능하다

이유

로지스틱 회귀는 기본적으로 직선 경계선을 생성하지만, 다중 클래스 로지스틱 회귀 또는 다항 로지스틱 회귀에서는 비선형 경계선도 생성할 수 있습니다. 따라서 로지스틱 회귀의 분류 경계선은 직선이나 비선형 곡선 모두 가능합니다.

Multiclass 분류기에 관한 다음 설명의 빈칸에 들어갈 가장 적절한 단어를 고르시오

"부류의 개수가 (N)인 분류 문제는 (___)개의 이진 분류기를 이용하여 Multiclass 분류기를 구현할 수 있다."

1. (N + 2)
2. (N + 1)
3. (N)
4. (2N)

답변

답안: 3) (N)

이유

Multiclass 분류 문제는 부류의 개수가 (N)일 때 (N)개의 이진 분류기를 이용하여 구현할 수 있습니다. 이는 "One-vs-Rest" 방식으로 각 클래스에 대해 이진 분류기를 학습시켜 Multiclass 분류를 수행하는 방법입니다.

다음 설명 중 과적합의 특징에 해당하는 것을 모두 고르시오

1. 학습 데이터에 대한 학습 성능이 우수하다
2. 테스트 데이터에 대한 성능이 우수하다
3. 학습 데이터에 대한 학습 성능이 나쁘다
4. 테스트 데이터에 대한 성능이 나쁘다

답변

답안: 1) 학습 데이터에 대한 학습 성능이 우수하다, 4) 테스트 데이터에 대한 성능이 나쁘다

이유

과적합(overfitting)은 모델이 학습 데이터에 지나치게 맞추어져서, 학습 데이터에서는 성능이 우수하지만 테스트 데이터에서는 일반화되지 않아 성능이 떨어지는 현상을 의미합니다. 따라서 학습 데이터에 대한 성능이 우수하고, 테스트 데이터에 대한 성능이 나쁩니다.

정규화에 관한 다음 설명 중 올바른 것을 모두 고르시오

1. 정규화는 정규 방정식의 계산과정에서 역행렬이 존재하지 않도록 한다
2. 정규화는 예측 함수 (hypothesis function)에 적용된다
3. 정규화는 비용 함수 (cost function)에 적용된다
4. 정규화는 파라미터 업데이트에 적용된다

답변

답안: 3) 정규화는 비용 함수 (cost function)에 적용된다, 4) 정규화는 파라미터 업데이트에 적용된다

이유

정규화는 비용 함수에 페널티 항을 추가하여 모델이 과적합되지 않도록 하는 방법입니다. 이는 비용 함수에 적용되어 모델의 복잡도를 줄이며, 정규화 항이 포함된 비용 함수를 최소화하는 과정에서 파라미터 업데이트에도 영향을 미칩니다. 역행렬의 존재 여부와는 관계가 없습니다.

다음 문장의 빈칸에 들어갈 가장 적절한 단어를 고르시오

"다층 신경망의 구조는 외부로부터 입력을 받아들여 활성 값을 계산하는 (___)들이 여러 개의 (___)으로 구성되어 있다."

1. 가중치, 뉴런
2. 가중치, 층
3. 뉴런, 가중치
4. 뉴런, 층

답변

답안: 4) 뉴런, 층

이유

다층 신경망의 구조는 여러 층(layer)으로 구성되어 있으며, 각 층은 여러 뉴런(neuron)으로 이루어져 있습니다. 따라서 빈칸에 들어갈 단어는 "뉴런"과 "층"이 맞습니다.

다음 신경망에 관한 다음 설명 중 올바른 것을 모두 고르시오

1. 신경망은 스스로 학습하는 파라미터를 가진 선형 회귀 모델로 볼 수 있다
2. 중간층의 뉴런의 개수는 입력층의 뉴런의 개수와 동일하다
3. 다층 신경망은 입력층을 포함하여 레이어의 수가 2개 이상인 경우를 의미한다
4. 손으로 쓴 숫자를 분류하는 문제에서 출력 층은 10개의 뉴런으로 구성된다

답변

답안: 3) 다층 신경망은 입력층을 포함하여 레이어의 수가 2개 이상인 경우를 의미한다, 4) 손으로 쓴 숫자를 분류하는 문제에서 출력 층은 10개의 뉴런으로 구성된다

이유

3번 설명은 다층 신경망의 정의에 부합합니다. 다층 신경망은 최소한 하나 이상의 은닉층을 포함하여 레이어가 2개 이상인 구조입니다. 4번 설명은 손으로 쓴 숫자를 분류하는 문제(예: MNIST 데이터셋)에서 0부터 9까지의 숫자를 분류하기 위해 출력 층이 10개의 뉴런으로 구성됩니다. 1번과 2번 설명은 부정확합니다. 신경망은 반드시 선형 회귀 모델로 볼 수 없으며, 중간층의 뉴런 수는 입력층과 동일할 필요가 없습니다.

오차 역전파 알고리즘에 관한 다음 설명 중 올바른 것을 모두 고르시오

1. 가중치 값은 오차 역전파 과정에서 업데이트 된다
2. 어느 뉴런의 델타 값은 그 이전 층에 있는 뉴런의 델타 값의 영향을 받는다
3. 빠른 수렴을 위해서는 가중치 값들이 0으로 초기화되어야 한다
4. 비용 함수는 중간층 뉴런에 대해 계산된다

답변

답안: 1) 가중치 값은 오차 역전파 과정에서 업데이트 된다, 2) 어느 뉴런의 델타 값은 그 이전 층에 있는 뉴런의 델타 값의 영향을 받는다

이유

1번 설명은 오차 역전파 알고리즘에서 가중치가 업데이트되는 과정을 설명하며, 이는 사실입니다. 2번 설명도 각 뉴런의 델타 값이 이전 층의 뉴런들로부터 전파된 오차에 영향을 받는다는 점에서 맞습니다. 3번 설명은 잘못된 것으로, 가중치는 일반적으로 작은 랜덤 값으로 초기화됩니다. 4번 설명은 비용 함수가 출력층에서 계산된다는 점에서 잘못되었습니다.

다음 오차 역전파 학습 알고리즘에 관한 설명 중 올바른 것을 모두 고르시오

1. 가중치는 균일하게 분포된 랜덤 값으로 초기화하는 것이 좋다
2. 0 값으로 가중치를 초기화하면 계산량이 줄어든다
3. 학습 알고리즘을 반복할 때마다 가중치를 초기화하면 수렴이 용이하다
4. 동일한 값으로 가중치를 초기화하면 학습 알고리즘의 수렴속도가 빠르다

답변

답안: 1) 가중치는 균일하게 분포된 랜덤 값으로 초기화하는 것이 좋다

이유

1번 설명은 맞습니다. 가중치는 일반적으로 작은 랜덤 값으로 초기화됩니다. 이는 모든 뉴런이 동일한 출력 값을 가지지 않도록 하기 위해서입니다. 2번 설명은 잘못되었습니다. 0 값으로 초기화하면 뉴런이 동일한 값을 가지게 되어 학습이 이루어지지 않습니다. 3번 설명도 잘못되었습니다. 학습 알고리즘을 반복할 때마다 가중치를 초기화하면 학습이 이루어지지 않습니다. 4번 설명도 잘못되었습니다. 동일한 값으로 초기화하면 뉴런이 동일한 값을 가지게 되어 학습이 이루어지지 않습니다.

다음 문장의 빈칸에 들어갈 가장 적절한 단어를 고르시오

"모델 선택에서 모든 후보 모델들은 (___) 데이터 만을 이용하여 학습시키고, (___) 데이터를 이용하여 테스트하고 성능이 가장 우수한 모델을 선택한 다음, (___) 데이터를 이용하여 일반화 오차를 계산한다."

1. 검증, 테스트, 학습
2. 검증, 학습, 테스트
3. 검증, 학습, 테스트
4. 학습, 검증, 테스트

답변

답안: 4) 학습, 검증, 테스트

이유

모델 학습과 검증 및 테스트 과정은 일반적으로 다음과 같이 이루어집니다:

1. 학습 데이터: 모델을 학습시키는 데 사용됩니다.
2. 검증 데이터: 학습된 모델의 성능을 평가하고, 하이퍼파라미터 튜닝에 사용됩니다.
3. 테스트 데이터: 최종적으로 선택된 모델의 일반화 오차를 평가하기 위해 사용됩니다.

다음 문장의 빈칸에 들어갈 가장 적절한 단어를 고르시오

"(___)을 그려보면 과적합이나 부족 적합 문제의 여부를 쉽게 식별할 수 있다."

1. 예측 함수
2. 학습 데이터
3. 학습 곡선
4. 분류 경계선

답변

답안: 3) 학습 곡선

이유

학습 곡선을 그려보면 모델의 학습 과정에서 과적합(overfitting)이나 부족 적합(underfitting) 문제를 쉽게 식별할 수 있습니다. 학습 곡선은 학습 데이터와 검증 데이터에 대한 모델의 성능을 학습 과정 동안 그래프로 나타

낸 것입니다.

다음 중 불균형 데이터 분류 문제를 나타내는 설명을 모두 고르시오

1. 테스트 데이터보다 학습 데이터가 훨씬 더 많은 문제이다
2. 한 부류의 데이터의 수가 아주 적은 반면에 다른 부류의 데이터의 수가 아주 많은 문제이다
3. 데이터 부류의 개수가 10개 이상인 문제이다

답변

답안: 2) 한 부류의 데이터의 수가 아주 적은 반면에 다른 부류의 데이터의 수가 아주 많은 문제이다

이유

불균형 데이터 문제는 클래스 간의 데이터 분포가 크게 차이 나는 문제를 말합니다. 즉, 한 클래스의 데이터 수가 다른 클래스의 데이터 수에 비해 매우 적을 때 발생합니다. 1번과 3번 설명은 불균형 데이터 문제와 관련이 없습니다.

다음 중 올바른 설명을 모두 고르시오

1. 정확도는 높을수록 좋고, 재현율 값은 작을수록 좋다
2. 평균값은 정확도(precision)와 재현율(recall)의 균형을 이루기 위한 좋은 평가 척도이다
3. F1 점수는 정확도(precision)와 재현율(recall)의 균형을 이루기 위한 좋은 평가 척도이다

답변

답안: 3) F1 점수는 정확도(precision)와 재현율(recall)의 균형을 이루기 위한 좋은 평가 척도이다

이유

1번 설명은 잘못되었습니다. 정확도(precision)와 재현율(recall) 모두 높을수록 좋은 것이며, 재현율 값이 작을수록 좋다는 것은 틀린 설명입니다. 2번 설명에서 '평균값'은 모호하며, 정확도와 재현율의 균형을 맞추기 위해 사용되는 주요 척도는 F1 점수입니다. F1 점수는 정확도와 재현율의 조화 평균으로, 두 값 사이의 균형을 이루기 위한 좋은 평가 척도입니다.

다음 중 올바른 설명을 모두 고르시오

1. SVM은 언제나 로지스틱 회귀보다 우수하다
2. SVM은 학습 데이터의 수가 특정 값의 수보다 약간 더 많을 때 로지스틱 회귀보다 우수하다
3. SVM은 학습 데이터의 수가 특정 값의 수보다 훨씬 많을 때 로지스틱 회귀보다 우수하다
4. SVM은 로지스틱 회귀보다 우수하지만 계산량이 많다

답변

답안: 4) SVM은 로지스틱 회귀보다 우수하지만 계산량이 많다

이유

1번 설명은 사실이 아닙니다. SVM이 항상 로지스틱 회귀보다 우수한 것은 아니며, 데이터의 특성에 따라 다릅니다. 2번과 3번 설명도 구체적인 데이터 양에 따른 우수성을 일반화할 수 없습니다. 4번 설명은 SVM이 로지스

틱 회귀보다 계산량이 많아 학습이 더 복잡하다는 점에서 맞는 설명입니다. SVM은 고차원 데이터에서 성능이 좋지만, 계산 비용이 많이 듭니다.

다음 문장의 빈칸에 들어갈 가장 적절한 단어를 고르시오

"일반적으로 가장 많이 사용하는 클러스터링 알고리즘은 K-means이며, 클러스터 중심을 초기화한 다음 (___)과 (___)의 두 단계를 반복적으로 수행한다."

1. Cluster movement, Centroid assignment
2. Cluster initialization, Centroid assignment
3. Cluster assignment, Centroid initialization
4. Cluster assignment, Centroid movement

답변

답안: 4) Cluster assignment, Centroid movement

이유

K-means 알고리즘은 클러스터 중심(centroid)을 초기화한 다음, 각 데이터 포인트를 가장 가까운 중심으로 할당(cluster assignment)하고, 그 후 클러스터 중심을 데이터 포인트들의 평균으로 이동(centroid movement)시키는 두 단계를 반복합니다.

K-means 알고리즘에 관한 다음 설명 중 올바른 것을 모두 고르시오

1. 여러 개의 국소 최적 값에 빠질 수 있다
2. 클러스터의 수가 일정하면 동일한 결과를 얻을 수 있다
3. 랜덤 초기화를 여러 번 반복하면 국소 최적값에 빠지는 것을 피할 수 있다
4. 랜덤 초기화를 하였을 때 국소 최적 값에 빠질 수 있다

답변

답안: 1) 여러 개의 국소 최적 값에 빠질 수 있다, 3) 랜덤 초기화를 여러 번 반복하면 국소 최적값에 빠지는 것을 피할 수 있다, 4) 랜덤 초기화를 하였을 때 국소 최적 값에 빠질 수 있다

이유

1번 설명은 맞습니다. K-means 알고리즘은 여러 개의 국소 최적 값에 빠질 수 있습니다. 3번 설명도 맞습니다. 랜덤 초기화를 여러 번 반복하면 다양한 초기 중심을 시도해 볼 수 있으므로 국소 최적값에 빠지는 것을 피할 가능성이 높아집니다. 4번 설명도 맞습니다. 랜덤 초기화를 하더라도 여전히 국소 최적값에 빠질 수 있습니다. 2번 설명은 틀렸습니다. 클러스터의 수가 일정해도 초기화에 따라 다른 결과를 얻을 수 있습니다.

다음 중 PCA를 사용하는 이유가 아닌 것은?

1. 학습 알고리즘의 실행 시간을 빠르게 함
2. 특징 값의 개수를 줄여 과적합을 예방함
3. 데이터를 저장하는 데 필요한 메모리와 디스크 공간을 줄임

답변

답안: 3) 데이터를 저장하는 데 필요한 메모리와 디스크 공간을 줄임

이유

PCA(주성분 분석)는 고차원 데이터를 저차원으로 축소하여 학습 알고리즘의 실행 시간을 줄이고, 과적합을 예방하는 데 도움이 됩니다. 그러나 PCA는 주로 차원 축소 및 데이터 시각화를 목적으로 하며, 데이터를 저장하는 데 필요한 메모리와 디스크 공간을 줄이는 것이 주요 목적은 아닙니다.

이상 데이터 검출에 관한 다음 설명 중 올바른 것을 모두 고르시오

1. 어떤 데이터의 확률이 임계값보다 작으면 그 데이터는 이상 데이터로 간주된다
2. Joint 확률 분포 함수는 언제나 개별 특정 값의 확률분포 함수의 곱에 의하여 구할 수 있다
3. 이상 데이터 샘플은 학습, 교차 검증, 테스트 데이터 셋으로 균등하게 배분해야 한다
4. 어떤 데이터의 확률이 임계값보다 크면 그 데이터는 이상 데이터로 간주될 수 있다

답변

답안: 1) 어떤 데이터의 확률이 임계값보다 작으면 그 데이터는 이상 데이터로 간주된다

이유

1번 설명은 맞습니다. 이상 데이터 검출에서는 데이터의 확률이 임계값보다 작으면 이상 데이터로 간주됩니다. 2번 설명은 틀렸습니다. Joint 확률 분포 함수는 개별 확률분포 함수의 곱으로 항상 구할 수 없습니다. 3번 설명도 틀렸습니다. 이상 데이터는 학습, 검증, 테스트 데이터 셋으로 균등하게 배분하지 않습니다. 4번 설명은 틀렸습니다. 데이터의 확률이 임계값보다 크면 일반적으로 이상 데이터로 간주되지 않습니다.