교환가능성과 컨포멀 예측

: 대칭성이 제공하는 통계적 보장

들어가며

동전을 100번 던져서 앞면과 뒷면이 나온 순서를 기록했다고 해보자. 이 기록의 앞 50번과 뒤 50번을 통째로 맞바꾼다고 해서 우리가 "이 동전은 공정한가?"라고 내릴 판단이 달라질까? 아마 그렇지 않을 것이다. 순서 자체가 중요한 정보가 아니라면 결과가 어떤 차례로 적혀 있는지는 그저 정리 방식의 문제에 가깝다. 우리는 보통 '언제 나왔는지'보다는 '무엇이 얼마나 나왔는지'를 보고 판단한다. 그래서 관측된 순서를 바꿔도 전체적인 성질은 그대로일 것이라고 자연스럽게 생각한다. 하지만 이 익숙한 직관 뒤에는 생각보다 강한 가정이 숨어 있다. 바로 "이 데이터는 순서를 섞어도 본질이 변하지 않는다"는 가정이다.

통계학에서는 이러한 성질을 **교환가능성(exchangeability)**이라 정의한다. 교환가능성은 단순한 기술적 조건 같지만 실제로는 데이터의 해석과 일반화 가능성을 결정짓는 핵심 기제다. 특히 최근 머신러닝의 **컨포멀 예측(conformal prediction)**은 이 교환가능성을 유일한 이론적 가정으로 삼아 모델 구조에 구애받지 않는 강건한 불확실성 추정을 가능케 한다 [5]. 이 글에서는 교환가능성의 정의와 이 대칭성이 갖는 중요성, 그리고 이러한 가정이 어떻게 컨포멀 예측의 유한표본(finite-sample) 보장으로 연결되는지를 단계적으로 고찰하고자 한다.

교환가능성: 인덱스는 정보를 담지 않는다

통계학에서 관측치 $X_1, X_2, \ldots, X_n$가 *교환가능(exchangeable)*하다고 말할 때, 이는 이 값들의 순서를 어떻게 바꾸더라도 전체적인 확률적 성질이 달라지지 않는다는 뜻이다. 수학적으로는 관측치들의 결합분포가 인덱스의 임의의 순열(permutation)에 대해 불변(invariant)임을 의미하며, 임의의 순열 π에 대해 다음이 성립한다.

$$(X_1, \ldots, X_n) \overset{d}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)})$$

여기서 $\overset{d}{=}$는 "같은 분포를 따른다"는 뜻이다.

이 정의가 말하고자 하는 핵심은 단순하다. 관측치에 붙은 번호, 즉 인덱스는 확률적으로 아무런 정보를 담고 있지 않다는 것이다. X_1이 첫 번째로 관측되었고 X_{17}이 열일곱 번째로 관측되었다는 사실 자체는, 데이터가 어떻게 생성되었는지에 대해 추가적인 단서를 제공하지 않는다. 인덱스는 그저 관측값들을 구분하기 위해 붙인 기술적인 라벨일 뿐이며 중요한 것은 값들이 "언제" 나왔는지가 아니라 "무엇이" 나왔는지다. 확률적 관점에서 의미를 갖는 것은 관측값들의 집합 그 자체이지 그 나열 순서가 아니다.

이 점은 일상적인 비유로도 쉽게 이해할 수 있다. 잘 섞인 카드 더미를 떠올려 보자. 특정 카드가 다섯 번째에 있느냐, 스무 번째에 있느냐는 사실은 그 카드의 성질이나 카드 더미의 구성에

대해 아무런 정보를 주지 않는다. 중요한 것은 어떤 카드들이 들어 있는지이지, 각 카드가 어디에 놓였는지는 우연의 결과일 뿐이다. 교환가능성은 바로 이런 직관을 데이터의 언어로 공식화한 개념이다.

이러한 순열 대칭성(permutation symmetry)이 성립하면, 관측치들 사이에 미리 정해진 위계나 역할은 존재하지 않는다. 모든 데이터는 동등한 자격으로 추론에 참여하며 이는 통계적 판단과 일반화를 가능하게 하는 중요한 이론적 토대가 된다.

i.i.d. 가정의 완화

통계학의 가장 익숙한 출발점은 관측치들이 i.i.d.(independent and identically distributed)라는 조건이다. 이는 모든 데이터가 서로 독립이며 동일한 분포에서 추출되었음을 전제하며 수많은 고전 이론의 출발점이 된다. i.i.d. 데이터는 자연스럽게 교환가능성을 만족한다. 서로 독립이고 분포도 같으니 관측 순서를 어떻게 바꾸든 전체적인 확률 구조가 달라질 이유가 없다. 하지만 중요한 점은 그 반대가 항상 성립하지는 않는다는 것이다. 교환가능성은 i.i.d.보다 훨씬 느슨한 조건이며 데이터들 사이의 독립성까지 요구하지는 않는다.

이를 이해하기 위해 잠재변수(latent variable)를 포함한 간단한 계층적 모형을 생각해보자.

$$\theta \sim P(\theta), \qquad X_i \mid \theta \overset{i.i.d.}{\sim} P(X \mid \theta)$$

이 모형에서는 먼저 보이지 않는 변수 θ가 하나 정해지고, 그 다음에 각 관측치 X_i가 이 θ를 기준으로 생성된다. θ가 주어졌을 때는 X_i들이 서로 독립이지만, θ를 알 수 없는 상태에서 보면 이야기가 달라진다. 하나의 관측값이 유난히 크다면, 이는 θ가 특정한 범위에 있을 가능성을 높이고, 그 결과 다른 관측값에 대한 기대에도 영향을 미치게 된다. 즉, 관측치들 사이에 의존성이 생긴다. 그럼에도 불구하고 이 데이터는 여전히 교환가능하다. 모든 관측치는 같은 잠재변수 θ를 통해 동일한 방식으로 생성되었기 때문에 순서를 어떻게 바꾸든 결합분포는 변하지 않는다. 다시 말해 이 데이터는 i.i.d.는 아니지만 교환가능성은 만족한다.

이 관점에서 보면 교환가능성은 i.i.d. 가정에서 가장 강한 제약인 독립성을 내려놓고, 대신 핵심적인 대칭성만을 남긴 개념이라 할 수 있다. 이처럼 최소한의 구조만을 요구하는 가정은 뒤에서 살펴볼 컨포멀 예측의 이론적 토대가 된다. 데이터 사이에 복잡한 의존성이 있거나 모델의 형태가 명확하지 않더라도, 여전히 의미 있는 통계적 보장을 가능하게 하는 이유가 바로 여기에 있다.

대칭성의 응용: Conformal Prediction

최근의 머신러닝 모델은 복잡한 함수 근사를 통해 뛰어난 예측 정확도를 달성하고 있다. 그러나 실제 의사결정 과정에서는 단순한 점 예측(point prediction)만으로는 불충분하며 예측에 수반되는 불확실성을 정량화하는 것이 중요하다. 특히 새로운 관측치에 대한 예측을 어느 수준까지 신뢰할 수 있는지 명확한 기준이 필요한데, 이를 위해 예측 구간(prediction interval) $C(X)$를 고려하며 이상적인 예측 구간은 다음과 같은 유한표본 커버리지 조건을 만족한다.

$$\mathbb{P}\big(Y_{n+1} \in C(X_{n+1})\big) \geq 1 - \alpha.$$

이 부등식은 이미 관측된 데이터 $(X_1, Y_1), \ldots, (X_n, Y_n)$를 바탕으로 어떤 절차를 통해 예측 구간 함수 $C(\cdot)$를 구성한 뒤, 앞으로 새롭게 들어올 입력값 X_{n+1}에 대해 구간 $C(X_{n+1})$을 제시했을 때 그 구간이 실제 반응값 Y_{n+1}을 포함할 확률이 최소 $1-\alpha$임을 의미한다. 다시 말해, $\alpha = 0.1$이라면 같은 과정을 여러 번 반복했을 때 우리가 제시한 예측 구간이 적어도 90% 이상의 경우에 참값을 포함해야 한다는 뜻이다. 전통적인 통계 방법에서는 선형성, 정규성, 또는 오차 구조에 대한 가정을 통해 이러한 구간을 도출한다. 그러나 고차원 · 비선형 모델이 일반적인 현대 머신러닝 환경에서는 이러한 가정을 검증하거나 정당화하기가 쉽지 않다.

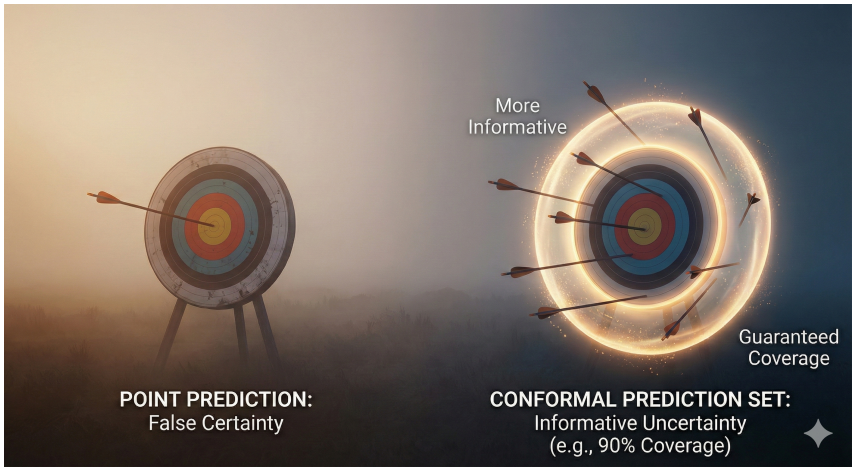

Figure 1: 점 예측과 컨포멀 예측 집합의 대비. 왼쪽의 점 예측은 단일 값만을 제시함으로써 과도한 확실성을 암시한다. 반면 오른쪽의 컨포멀 예측 집합은 불확실성을 명시적으로 표현하면서도 사전에 정해진 수준의 커버리지(예: 90%)를 보장한다. 출처: Google Gemini

컨포멀 예측은 이 난제를 교환가능성이라는 최소한의 가정만으로 해결한다 [5]. 본 방법론의 핵심은 예측 구간의 타당성을 모델의 구조나 분포 가정에 기대는 대신, 데이터가 갖는 순열 대칭성(permutation symmetry)으로부터 직접 도출한다는 점이다. 이를 위해 컨포멀 예측은 각 관측치가 "얼마나 데이터에 어울리지 않는지"를 수치화한 비적합도 점수(nonconformity score)를 도입한다. 예측 함수 $\hat{f}$와 점수 함수 $S(\cdot)$가 주어졌을 때,

$$S_i = S(X_i, Y_i; \hat{f}) \qquad (i = 1, \ldots, n+1)$$

로 점수를 정의하자. 중요한 사실은 $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$가 교환가능하다면, 이들로부터 계산된 점수들의 결합분포 역시 순열에 대해 불변이라는 점이다. 다시 말해 새로운 점수 S_{n+1}은 기존 점수들과 확률적으로 완전히 대등한 위치에 놓이며, 어떤 관측치의 점수가 "특별히" 더 크거나 작을 이유가 없다.

이 대칭성의 직접적인 귀결로, S_{n+1}이 $\{S_1, \ldots, S_{n+1}\}$ 가운데 차지하는 *순위(rank)*는 모든 값이 동일한 확률을 갖는다. 즉, 새로운 점수의 순위는 균등분포(uniform distribution)를 따르게 된다. 컨포멀 예측은 바로 이 "순위의 균등성"을 이용하여 점수의 상위 α 비율을 배제하도록 임계값을 정하고, 그 결과로 분포 가정이나 모델의 정확도와 무관한(distribution-free) 유한표본 커버리지를 달성한다.

작동 원리와 통계적 보장: Split Conformal Prediction

스플릿 컨포멀 예측(split conformal prediction)은 컨포멀 예측의 여러 변형 가운데 가장 이해하기 쉽고 실제로도 가장 널리 사용되는 방법이다 [3]. 이 절에서는 스플릿 컨포멀 예측을 중심으로, 컨포멀 예측이 어떻게 작동하며 어떤 통계적 보장을 제공하는지 살펴본다.

스플릿 컨포멀 예측의 절차는 전체 데이터를 학습용과 교정용($\mathcal{I}_{\text{cal}}$)으로 분할하는 것으로 시작한다. 학습 데이터는 오직 예측 함수 $\hat{f}$를 추정하는 데 사용되며, 이때 선형 회귀나 신경망 등 어떤 알고리즘을 사용해도 무방하다.

모델을 한 번 학습하고 나면 이제 교정 데이터에 대해 "이 모델이 얼마나 빗나갔는지"를 측정한다. 가장 흔한 방법은 잔차(residual)의 절댓값을 사용하는 것이다.

$$S_i = |Y_i - \hat{f}(X_i)| \qquad (i \in \mathcal{I}_{\text{cal}})$$

이 점수는 각 관측치가 모델의 예측에서 얼마나 멀리 떨어져 있는지를 나타낸다. 값이 클수록 해당 관측치는 모델 입장에서 덜 "자연스럽다"고 볼 수 있다. 여기서 전체 데이터의 개수를 n이라 하고 그 중 $m = |\mathcal{I}_{\text{cal}}|$개를 교정용으로 사용하고 있다고 생각해 보자. 즉 지금 시점에서는 이미 m개의 교정 점수 $\{S_1, \ldots, S_m\}$가 계산되어 있는 상황이다 (편의상 교정 데이터의 인덱스를 $1, \ldots, m$으로 다시 번호를 매겨 두었다). 이제 학습과 교정에 사용되지 않았던 *완전히 새로운* 관측치 (X_{n+1}, Y_{n+1})가 하나 들어왔다고 하자. 이 데이터에 대해서도 같은 방식으로 점수를 계산해

$$S_{m+1} = |Y_{n+1} - \hat{f}(X_{n+1})|$$

로 정의한다. 여기서 S_{m+1}의 아래첨자 $m+1$은 이미 계산해 둔 교정 점수 m개에 새 점수를 하나 더 *임시로 추가해* 함께 비교하기 위해 붙인 번호라고 이해하면 된다. 즉 실제로 교정 데이터를 늘려서 다시 학습하는 것이 아니라, $(S_1, \ldots, S_m)$ 뒤에 S_{m+1}을 하나 얹어 $(S_1, \ldots, S_m, S_{m+1})$의 상대적인 크기 순서를 살펴보는 것이다. 이 "점수를 하나 더 추가해 순위를 비교한다"는 관점이 컨포멀 예측의 중요 아이디어다.

이제 핵심적인 관찰을 하나 짚고 넘어가자. 학습과 교정을 분리했기 때문에 $\hat{f}$는 교정 데이터와는 독립적으로 고정된 함수처럼 취급할 수 있다. 따라서 교환가능성 가정 하에서는

$$(S_1, \ldots, S_m, S_{m+1})$$

이 점수들 역시 서로 대등한 교환가능한 대상이 된다.

이 대칭성에서 곧바로 중요한 결론이 나온다. 새로운 점수 S_{m+1}이 이 점수들 가운데에서 차지하는 순위는 특정 위치에 치우치지 않고 균등하게 분포한다는 것이다. 그 결과

$$\mathbb{P}\big(S_{m+1} \le S_{(\lceil (m+1)(1-\alpha) \rceil)}\big) \ge 1 - \alpha$$

가 성립한다. 여기서 $S_{(k)}$는 교정 점수 $S_1, \ldots, S_m$를 크기순으로 정렬했을 때의 k번째 값이다. 즉, 교정 데이터에서 관측된 점수들의 경험적 분포를 기준으로

$$q = S_{(\lceil (m+1)(1-\alpha) \rceil)}$$

를 임계값으로 잡으면 새로운 점수가 이를 넘지 않을 확률이 최소 $1-\alpha$임이 보장된다.

마지막으로 이 임계값을 예측 구간으로 옮기면 된다.

$$C(X_{n+1}) = \left[\hat{f}(X_{n+1}) - q,\ \hat{f}(X_{n+1}) + q\right].$$

이렇게 구성된 구간은 분포의 형태나 모델의 정확도에 대한 추가 가정 없이도

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$$

라는 유한표본 커버리지를 엄밀하게 만족한다 (요약본: Figure 3). 다시 말해, 데이터가 많아질 때까지 기다릴 필요도 없고, 모델이 "잘 맞는다"는 믿음을 전제할 필요도 없다. 오직 교환가능성이라는 대칭성만으로 얻어진 결과다. 보다 자세한 증명과 이론적 확장은 [1]를 참고하면 된다.

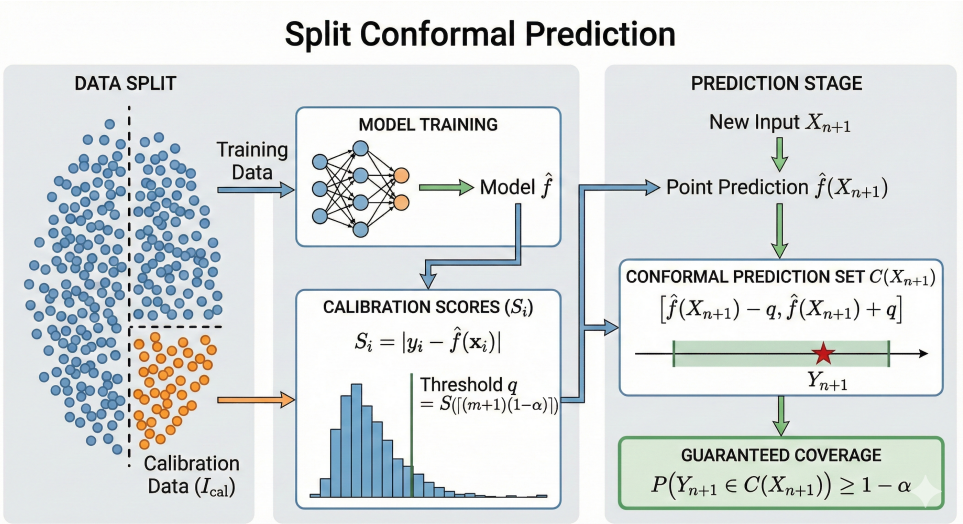

Figure 2: 스플릿 컨포멀 예측의 작동 원리. 출처: Google Gemini

한계와 확장

지금까지의 논의는 하나의 중요한 전제 위에 서 있다. 바로 데이터가 교환가능하다는 가정이다. 그러나 현실의 데이터가 항상 이 가정을 만족하는 것은 아니다. 예를 들어 시계열 데이터처럼 시간이 흐르며 관측되는 경우나, 모집단의 성질이 점차 변하는 분포 이동(distribution shift) 상황에서는 관측 순서 자체가 의미 있는 정보를 담는다. 이때는 순서를 바꾸어도 괜찮다는 대칭성이 깨지며 그 결과 기존 컨포멀 예측이 제공하던 커버리지 보장 역시 더 이상 그대로 적용되지 않는다.

이러한 한계를 보완하기 위해 최근에는 여러 확장 기법들이 제안되고 있다. 대표적인 예가 가중 컨포멀 예측(weighted conformal prediction)이다. 이 방법은 모든 관측치를 동일하게 취급하는 대신, 현재 예측하려는 분포에 더 가깝다고 판단되는 관측치에 더 큰 가중치를 부여해 비적합도 점수를 계산한다. 이 밖에도 분포의 변화를 명시적으로 추정한 뒤 이를 보정하는 접근법들이 활발히 연구되고 있다 [4].

한편, 설령 교환가능성이 성립하더라도 컨포멀 예측이 제공하는 보장에는 또 하나의 근본적인

한계가 있다. 컨포멀 예측의 커버리지는 본질적으로 *주변적(marginal)*이다. 다시 말해

$$\mathbb{P}\big(Y_{n+1} \in C(X_{n+1})\big) \geq 1 - \alpha$$

라는 보장은 모든 입력 X_{n+1}에 대해 평균적으로 성립하는 확률일 뿐 특정 입력값 $X_{n+1} = x$에 대해 개별적으로 보장되는 것은 아니다(Figure 3 참조).

이 때문에 전체적으로는 약속된 커버리지를 만족하더라도 입력 공간의 어떤 영역에서는 예측 구간이 필요 이상으로 넓어질 수 있고 다른 영역에서는 지나치게 좁아질 수도 있다. 이러한 문제의식 속에서 조건부 커버리지(conditional coverage)를 목표로 하거나 입력 공간을 국소적으로 나누어 각 영역별로 컨포멀 예측을 수행하는 방법들이 제안되고 있다. 다만 유한 표본에서 분포 가정 없이 완벽한 조건부 커버리지를 달성하는 것은 불가능함이 증명된 만큼 [2] 이러한 접근법들은 필연적으로 추가적인 구조적 가정이나 근사를 필요로 한다.

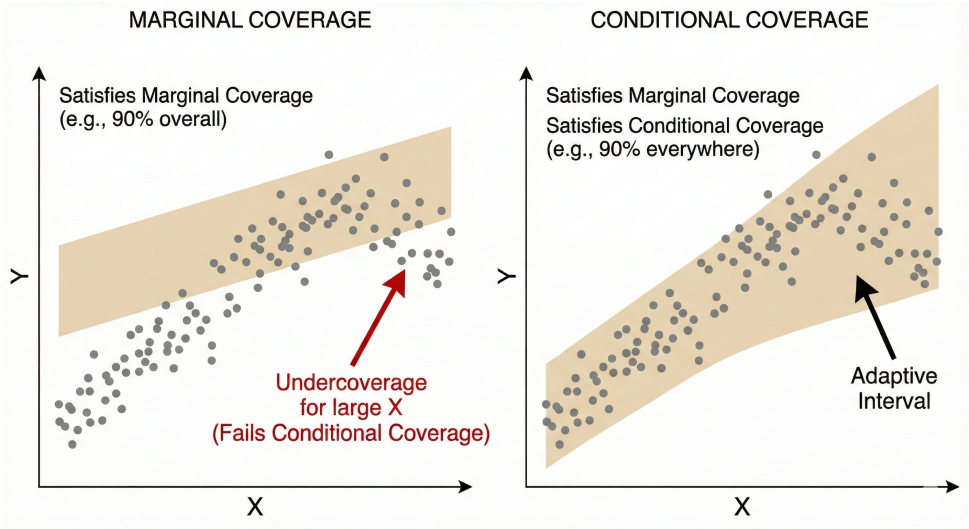

Figure 3: Marginal coverage와 conditional coverage의 차이를 나타낸 그림. 왼쪽의 예측 구간은 주변적 커버리지는 만족하지만 조건부 커버리지는 만족하지 않으며 특히 X_{n+1}의 값이 클 경우 과소 커버리지(undercoverage)가 발생한다. 반면 오른쪽 예측 구간은 두 가지 커버리지 조건을 모두 만족한다. 출처: Google Gemini

맺으며

교환가능성은 "관측 순서에는 정보가 없다"는 매우 단순한 생각에서 출발한다. 하지만 이 소박한 대칭성은 복잡한 분포 가정이나 정교한 모델 분석 없이도 불확실성을 정량화할 수 있게 해주는 강력한 이론적 기반이 된다. 컨포멀 예측은 그 가능성을 가장 분명하게 보여주는 사례다. 우리는 모델의 내부가 어떻게 작동하는지 데이터가 정확히 어떤 분포에서 생성되었는지를 완벽히 알지 못하더라도 데이터가 통계적으로 대등하다는 사실 하나만으로 신뢰할 수 있는 예측 구간을 만들 수 있다.

현대의 데이터 분석 환경에서 모델은 점점 더 크고 복잡해지며 종종 블랙박스에 가깝다. 이런 상황에서 모든 통계적 가정을 하나하나 검증하는 일은 현실적으로 어렵다. 그래서 가정을 쌓아 올리는 대신 데이터가 본질적으로 만족하는 최소한의 구조에 주목하는 접근이 중요해진다. 그런 의미에서 "이 데이터는 교환가능한가?"라는 질문은 단순한 기술적 점검을 넘어 우리가 어떤 수준의 신뢰와

일반화를 기대할 수 있는지를 가늠하는 가장 근본적인 출발점이라고 할 수 있다.

References

[1] A. N. Angelopoulos, R. F. Barber, and S. Bates, Theoretical foundations of conformal prediction, *arXiv preprint arXiv:2411.11824*, 2024.

[2] J. Lei and L. Wasserman, Distribution-free prediction bands for non-parametric regression, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 1, pp. 71–96, 2014.

[3] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, Distribution-free predictive inference for regression, *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1094–1111, 2018.

[4] R. J. Tibshirani, R. Foygel Barber, E. J. Candès, and A. Ramdas, Conformal prediction under covariate shift, *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[5] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*, Springer, 2005.