

AVERAGE TOP-K LOSS

LEARNING WITH ~

REFERENCE

- ▶ Fan, Yanbo, et al.
"Learning with average top-k loss."
Advances in neural information processing systems. 2017.

ABSTRACT

ABSTRACT

- ▶ Average top-k (AT_k) 손실
 - ▶ 지도 학습을 위한
 - ▶ 새로운 통합 손실
- ▶ AT_k 는
 - ▶ 학습 데이터셋에 대해
 - ▶ k 개의 가장 큰 개별적 손실들을 평균낸 것

ABSTRACT

- ▶ AT_k 가
 - ▶ 유명한 두 통합 손실 방법들
 - ▶ Average loss와 maximum loss의
 - ▶ 일반화임을 보임
- ▶ 두 방법의 장점을 더하고 단점을 완화한 것
 - ▶ 다른 데이터 분포에 더 잘 적응

INTRODUCTION

INTRODUCTION

- ▶ 지도 학습이란
 - ▶ 레이블링된 훈련 예제 $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ 를 이용해
 - ▶ data/features x 로부터
 - ▶ 타겟 y 를 예측

INTRODUCTION

- ▶ 전형적으로 함수 f 를 찾음으로써 해결할 수 있음
 - ▶ **통합 손실(Aggregation loss)**를 최소화하는 함수 f
 - ▶ 훈련 샘플들에 대한 각 손실로부터 구성됨

INTRODUCTION

- ▶ $\min_f \{ \mathcal{L}(L_z(f)) + \Omega(f) \}$
 - ▶ $\mathcal{L}(L_z(f))$: 통합 손실
 - ▶ $L_z(f) = \{l_i(f)\}_{i=1}^n$: 훈련 샘플 각각의 개별 손실의 누적
 - ▶ $l_i(f) = l(f(\mathbf{x}_i), y_i)$
 - ▶ $\Omega(f)$: Regularizer

INTRODUCTION

- ▶ 통합 손실 \mathcal{L} 로 주로 쓰이는 것은 평균
 - ▶ 모든 데이터의 손실의 평균
 - ▶ 편향되지 않음

INTRODUCTION

- ▶ Maximum 손실
 - ▶ 전체 데이터 중 손실이 가장 큰 하나
- ▶ Top-k 손실
 - ▶ Maximum 손실의 일반화
 - ▶ Maximum 손실 대비 아웃라이어의 영향을 경감
 - ▶ 그러나 convex하지 않음
 - ▶ 경사 하강법을 통한 최소화가 보장되지 않음

AVERAGE TOP-K LOSS

AVERAGE TOP-K LOSS

- ▶ 저자들이 제안하는 average top-k (AT_k) 손실
 - ▶ 최대 k 개의 개별 손실의 평균
 - ▶ $\mathcal{L}_{avt-k}(L_z(f)) = \frac{1}{k} \sum_{i=1}^k l_{[i]}(f)$
- ▶ AT_k 손실의 최소화를 MAT_k 학습이라 칭할 것

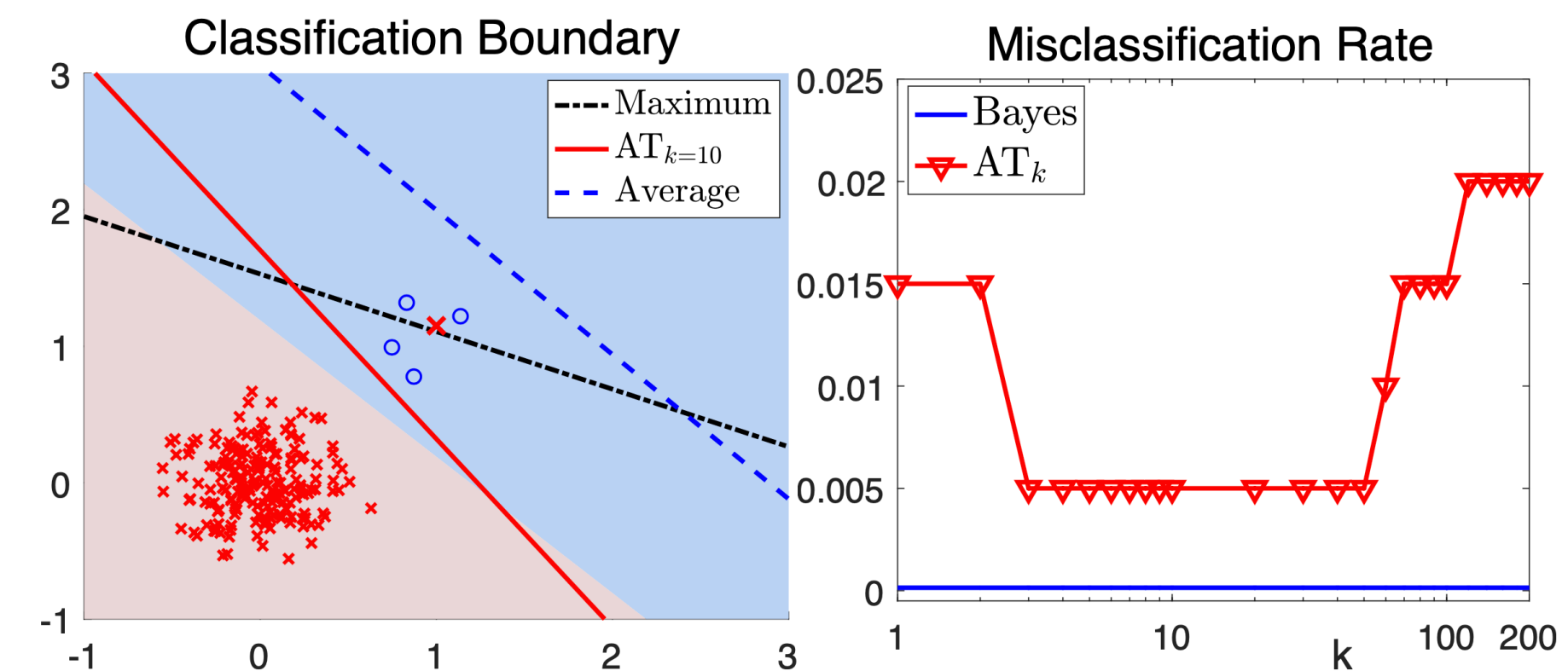
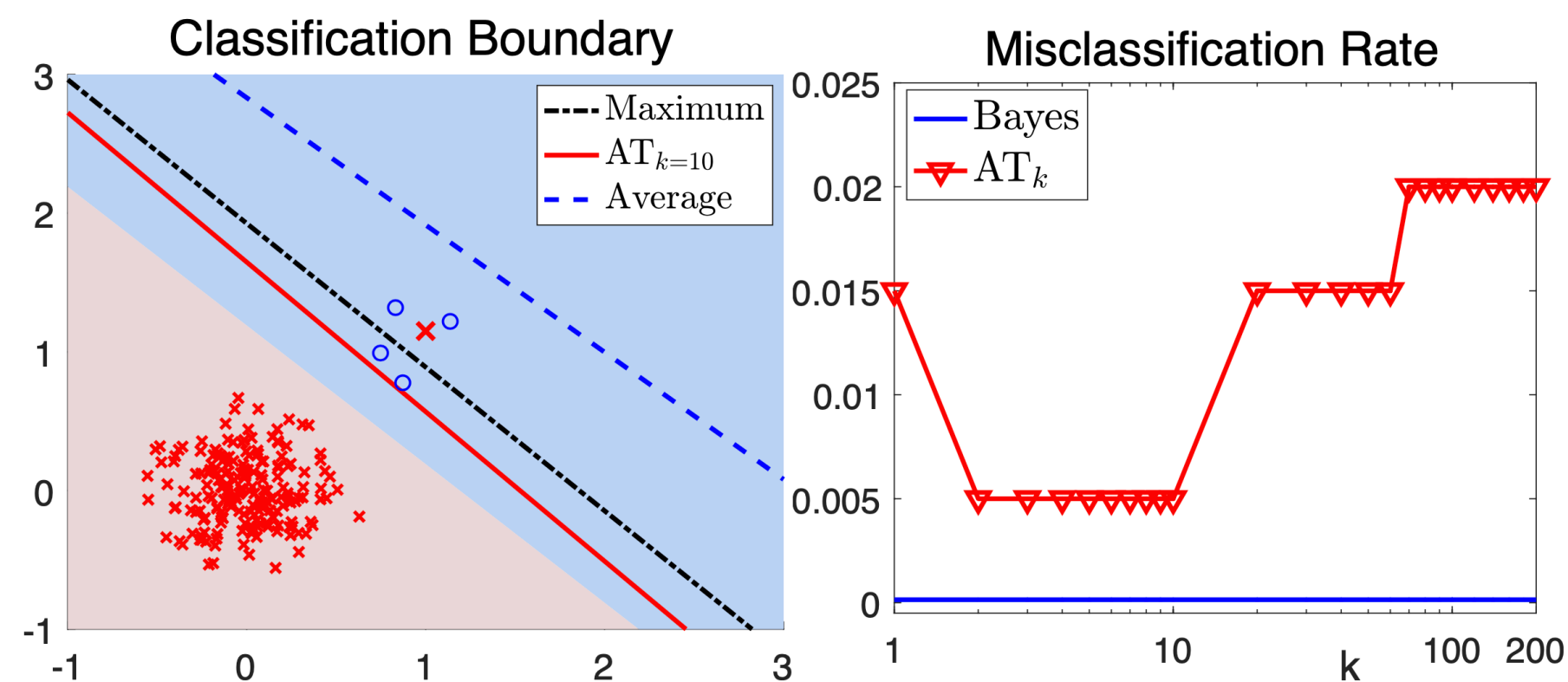
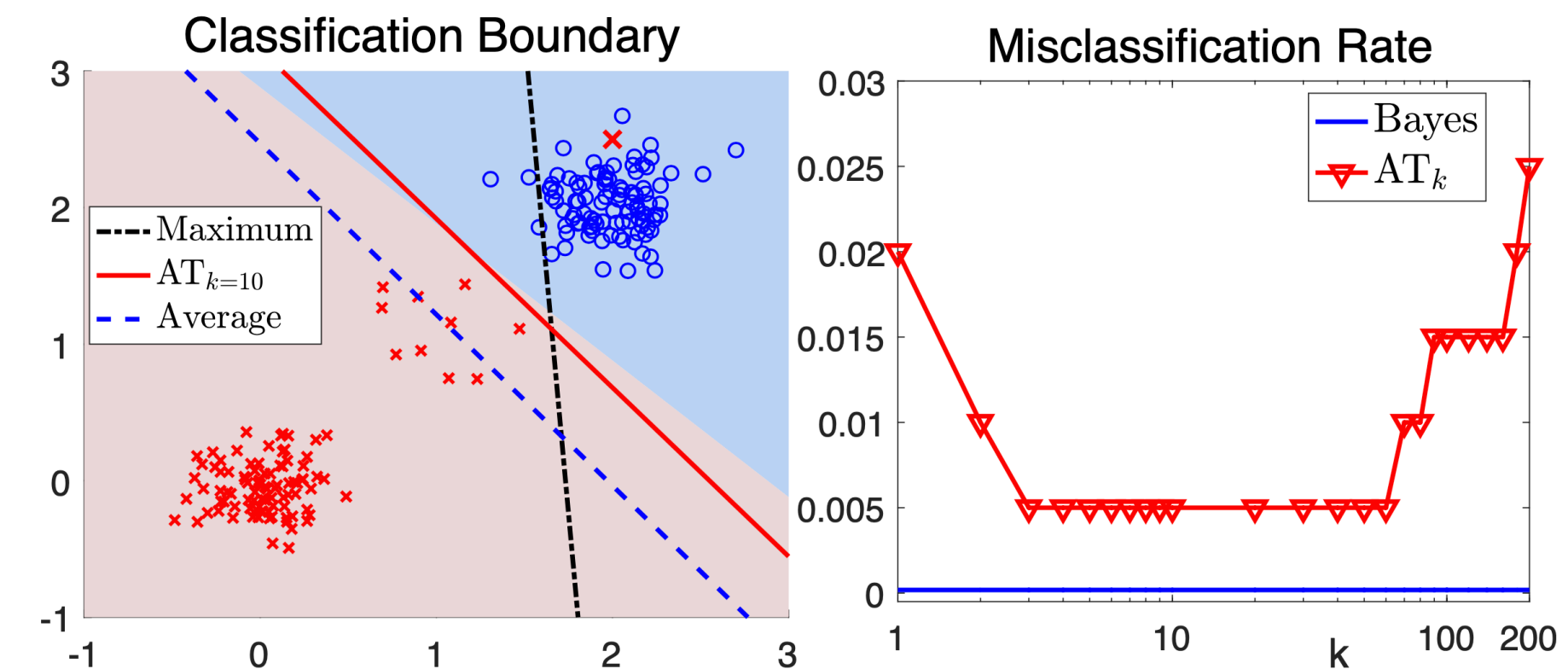
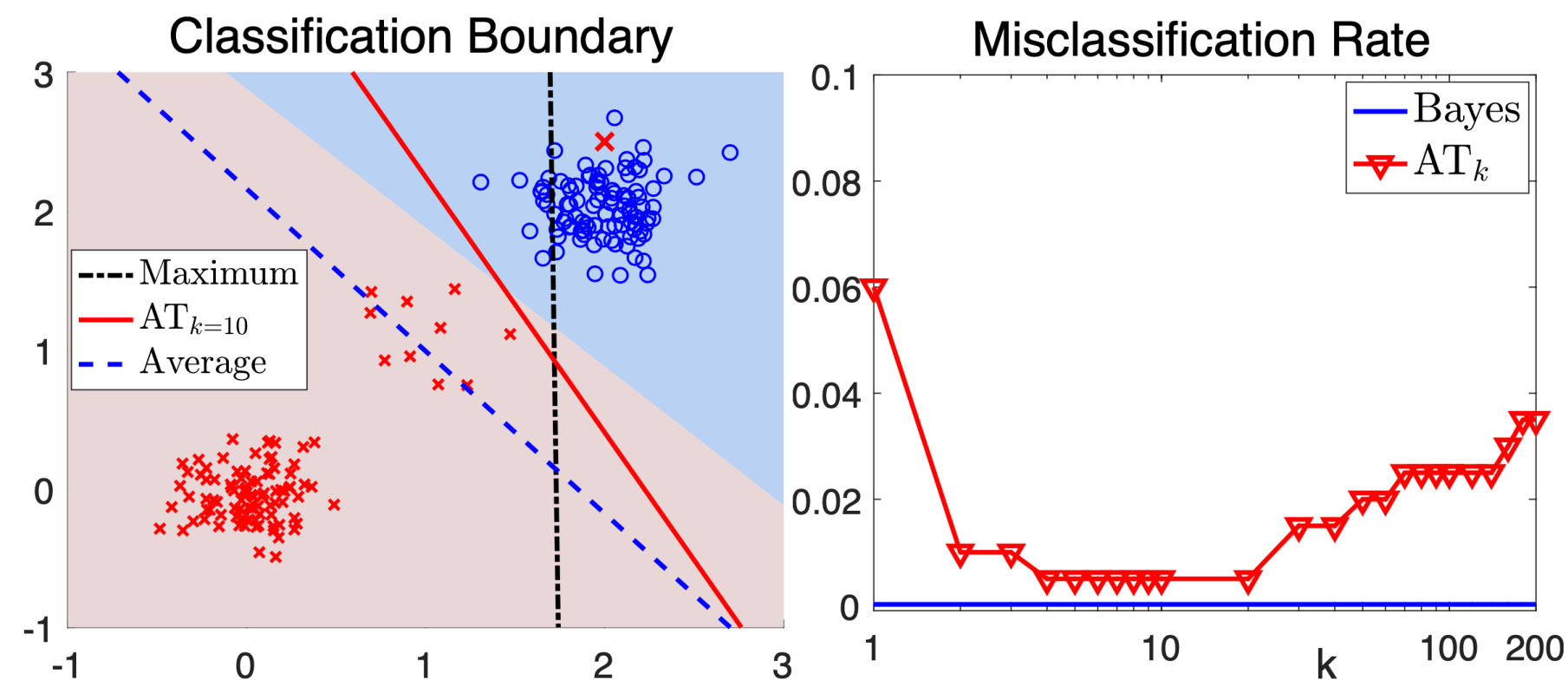
AVERAGE TOP-K LOSS

- ▶ Average top-k (AT_k) 손실
- ▶ Average loss와 maximum loss의 일반화
 - ▶ $k = 1$ 이면 maximum loss
 - ▶ $k = n$ 이면 average loss
- ▶ AT_k 는 1이나 n 이 아닌 적당한 값을 찾아 설정
 - ▶ (어찌보면 당연히) 더 좋은 성능

AVERAGE TOP-K LOSS

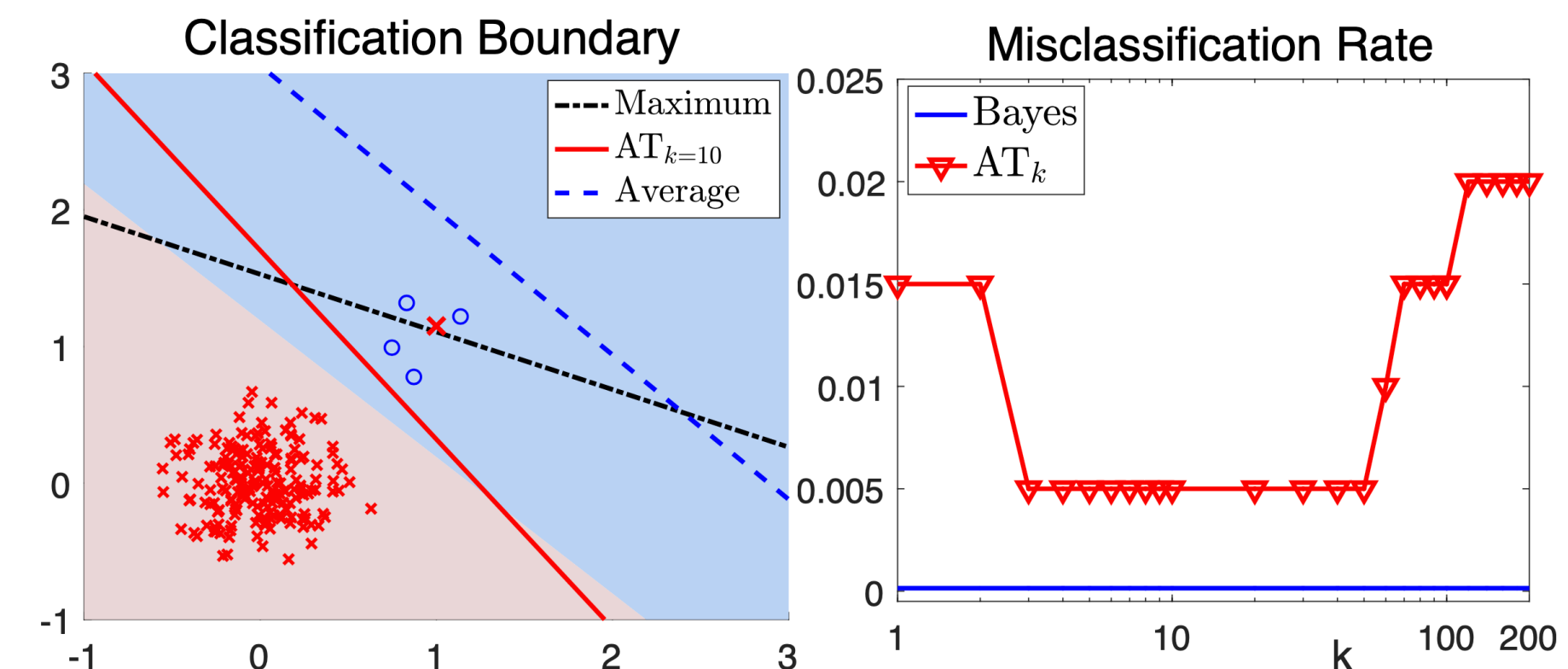
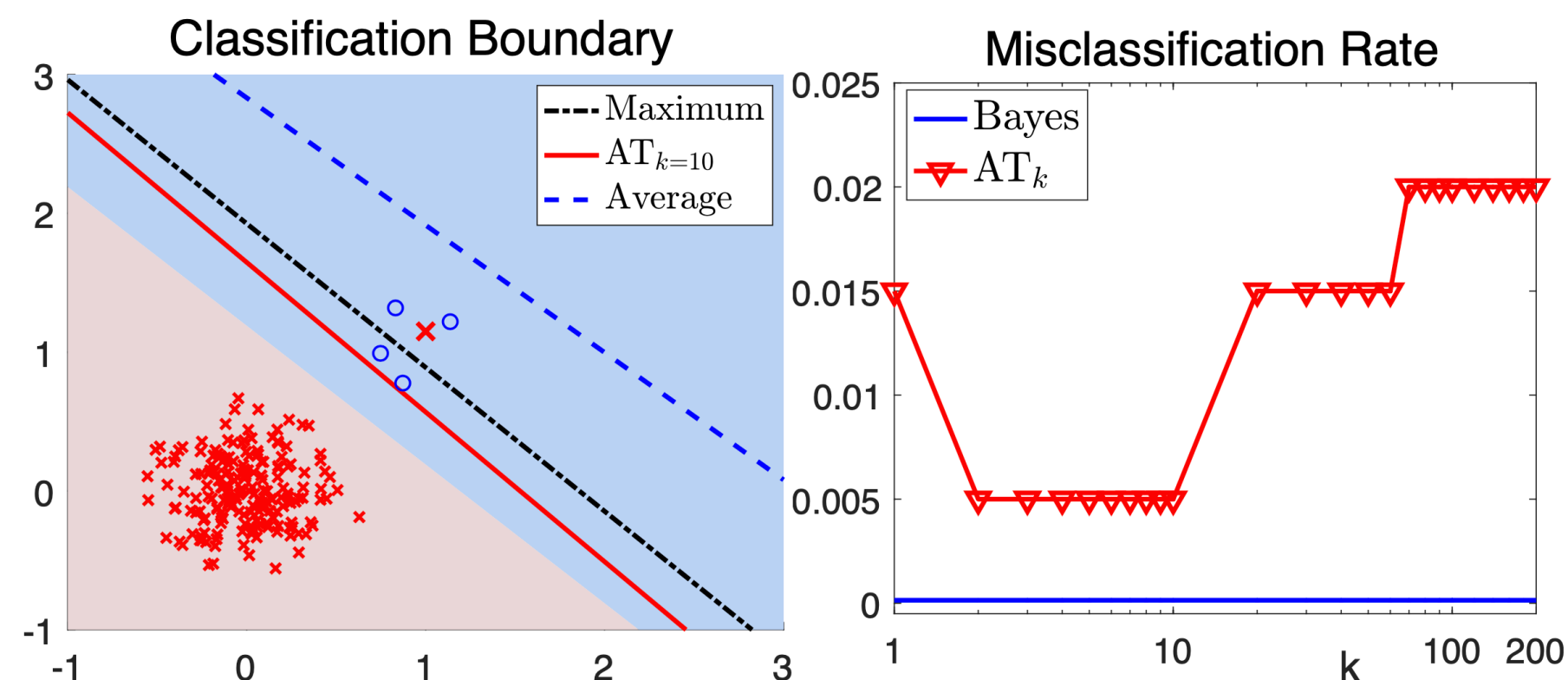
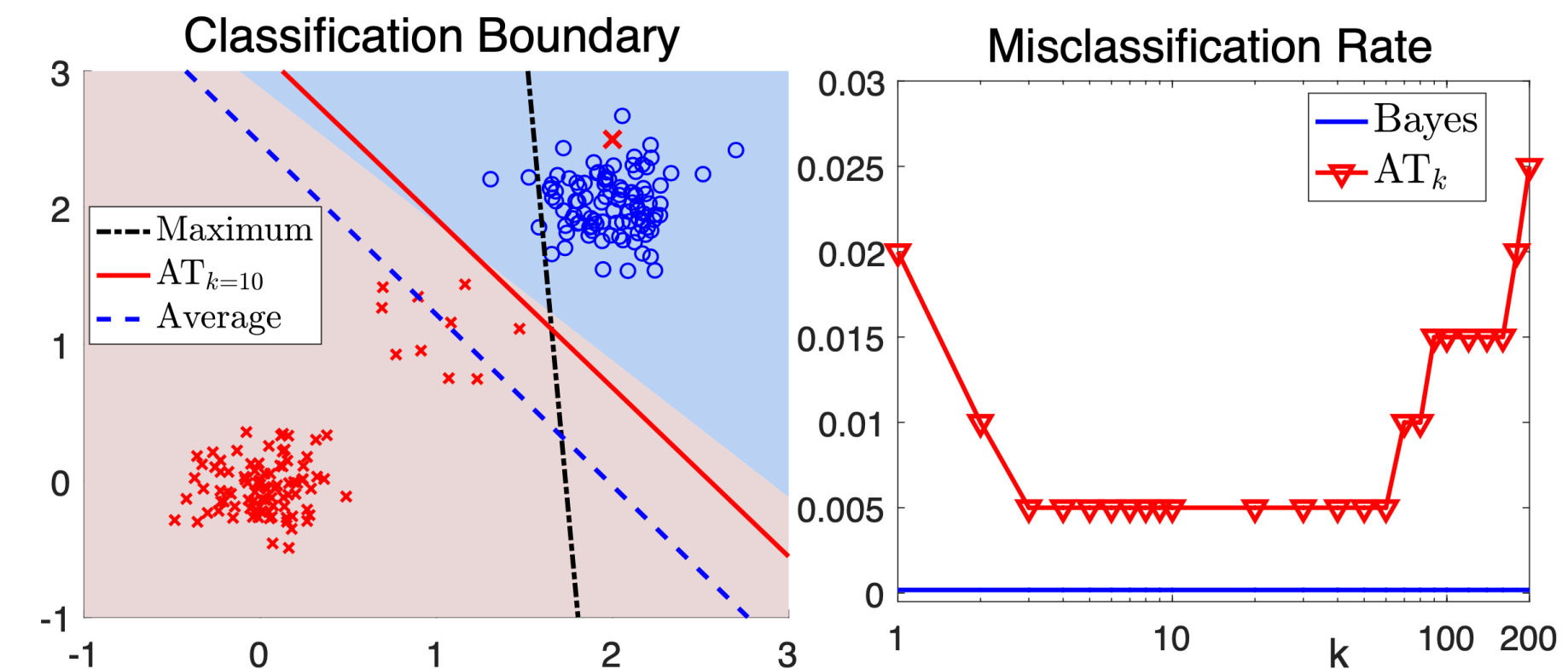
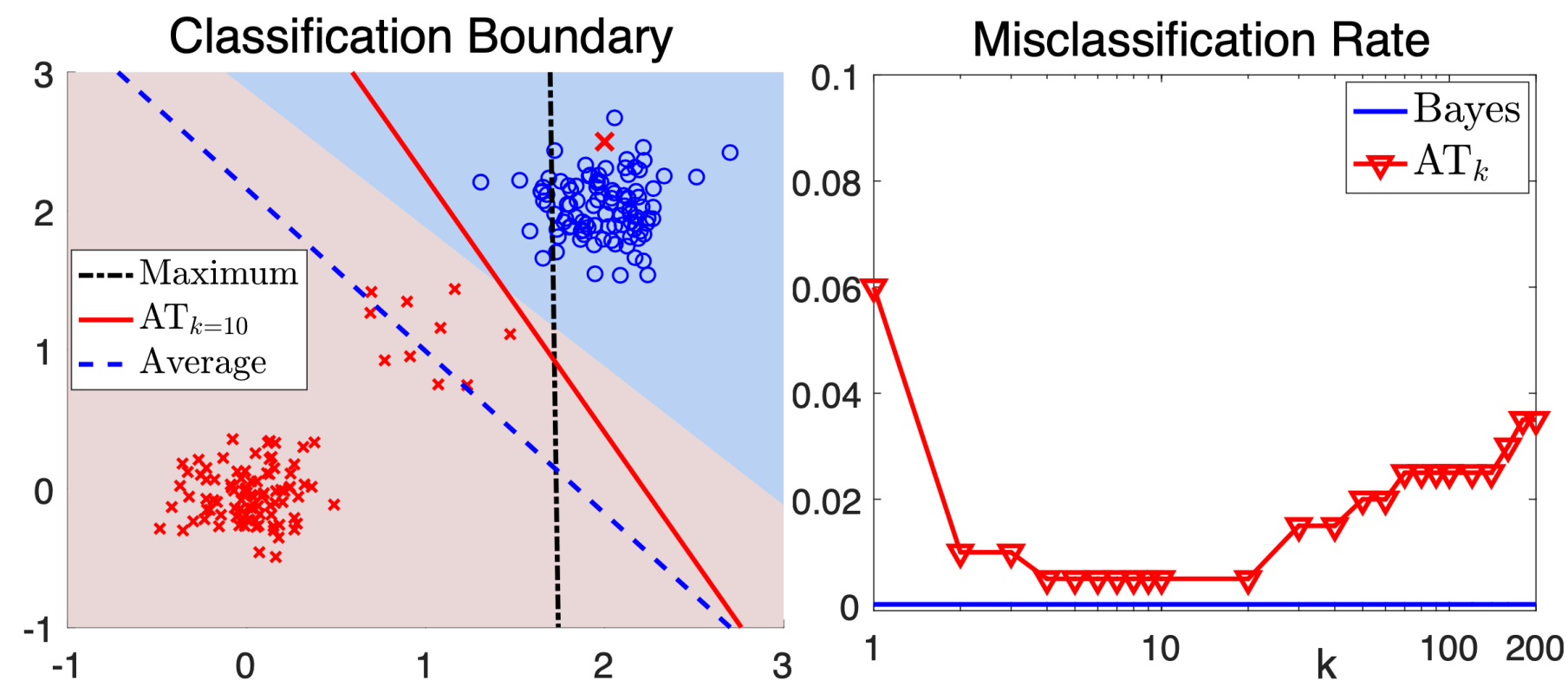
▶ 아웃라이어가 있는 바이너리 분류 문제

▶ 200개의 샘플



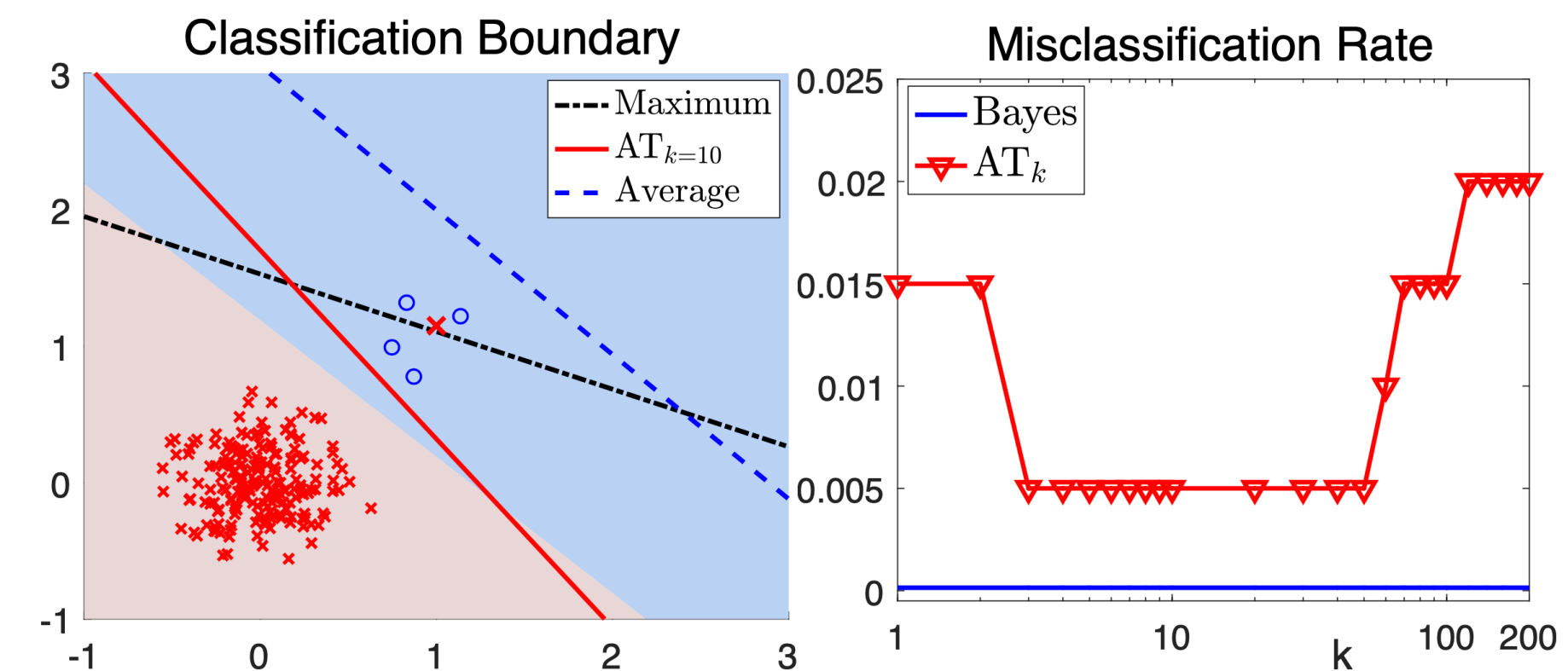
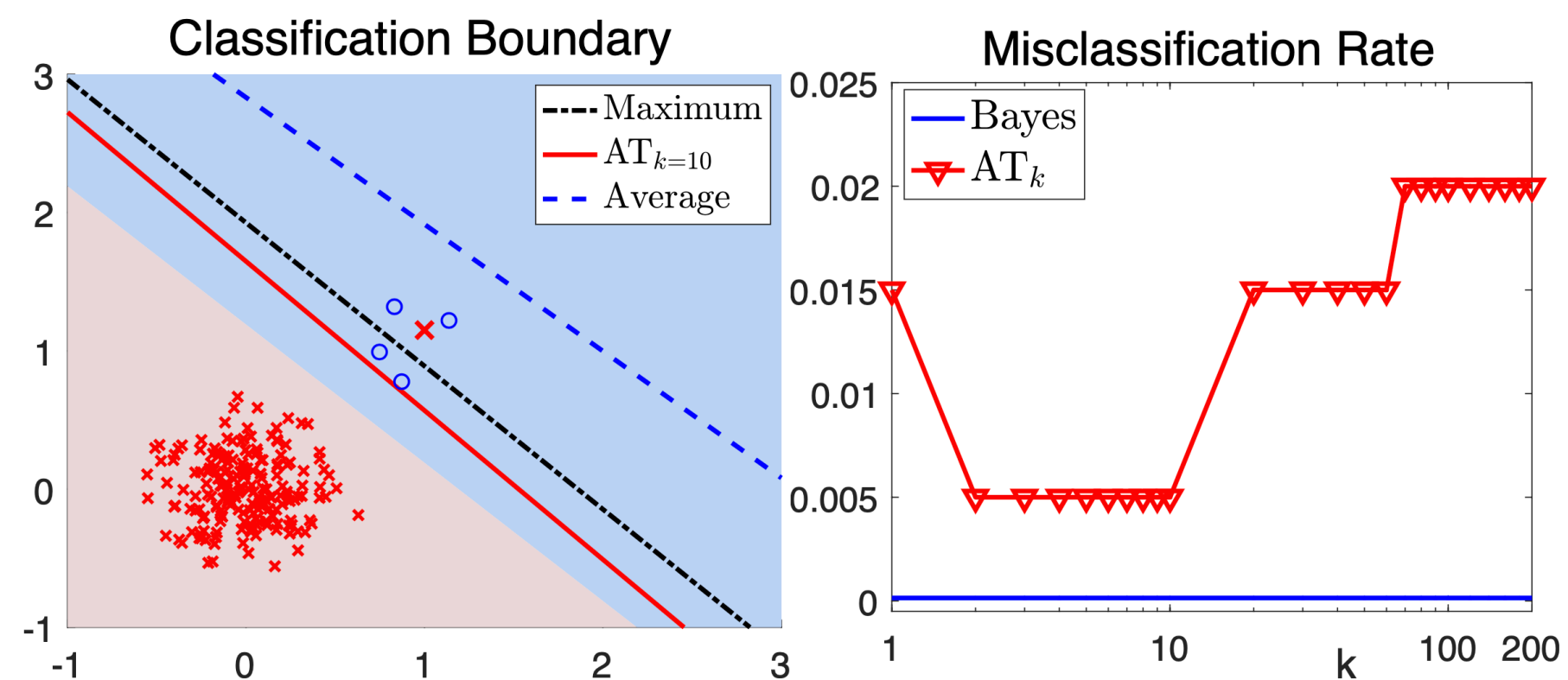
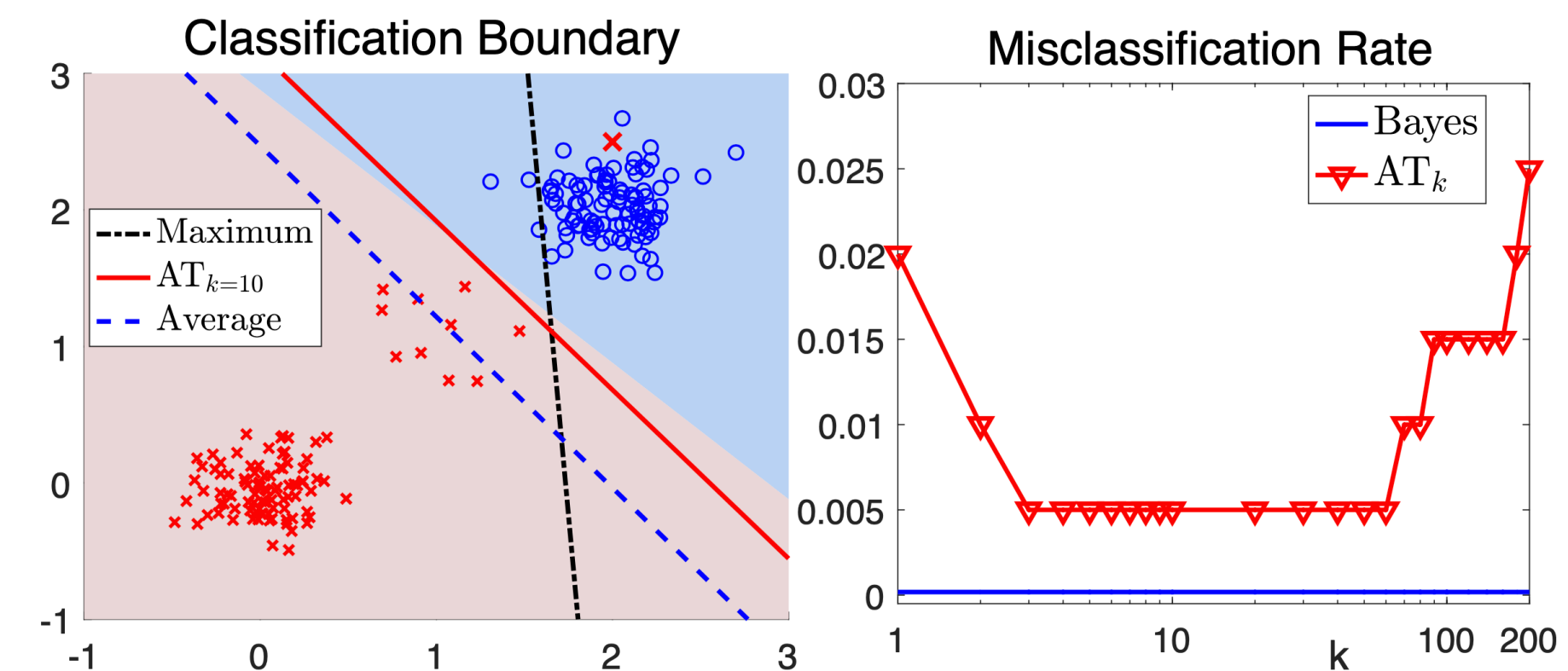
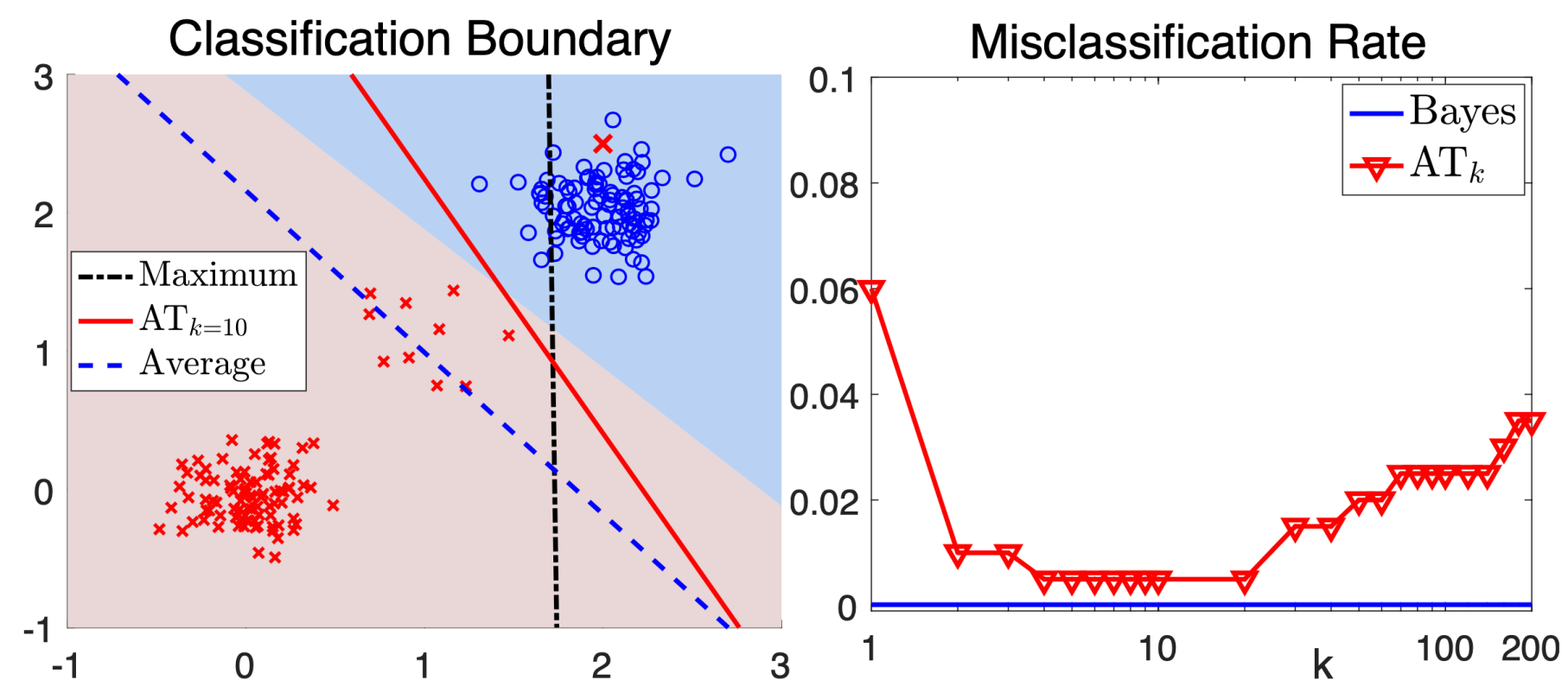
AVERAGE TOP-K LOSS

- ▶ 위: 데이터의 분포가 multimodal (뚜렷하게 구분되는)
- ▶ 아래: 레이블의 분포가 극심하게 불균형



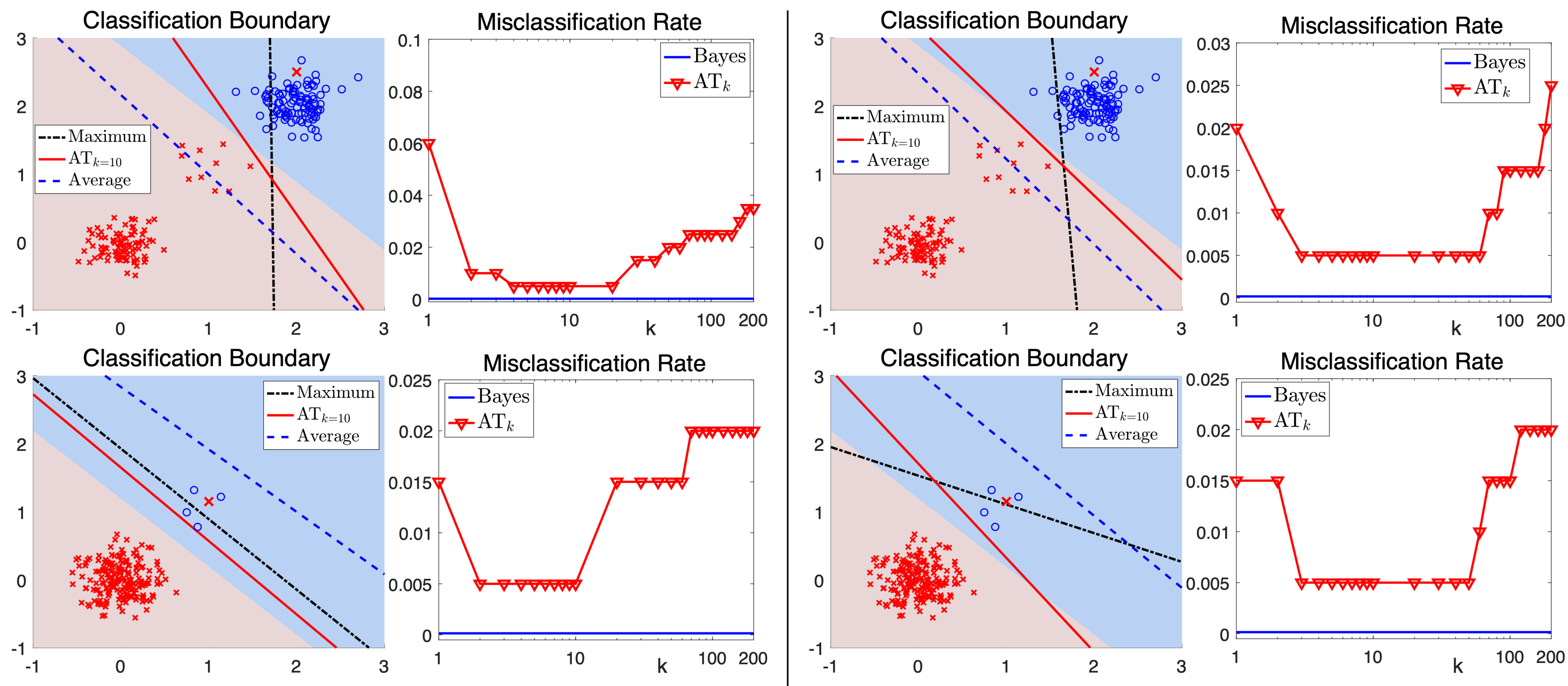
AVERAGE TOP-K LOSS

- ▶ 왼쪽: Hinge loss
- ▶ 오른쪽: Logistic loss



AVERAGE TOP-K LOSS

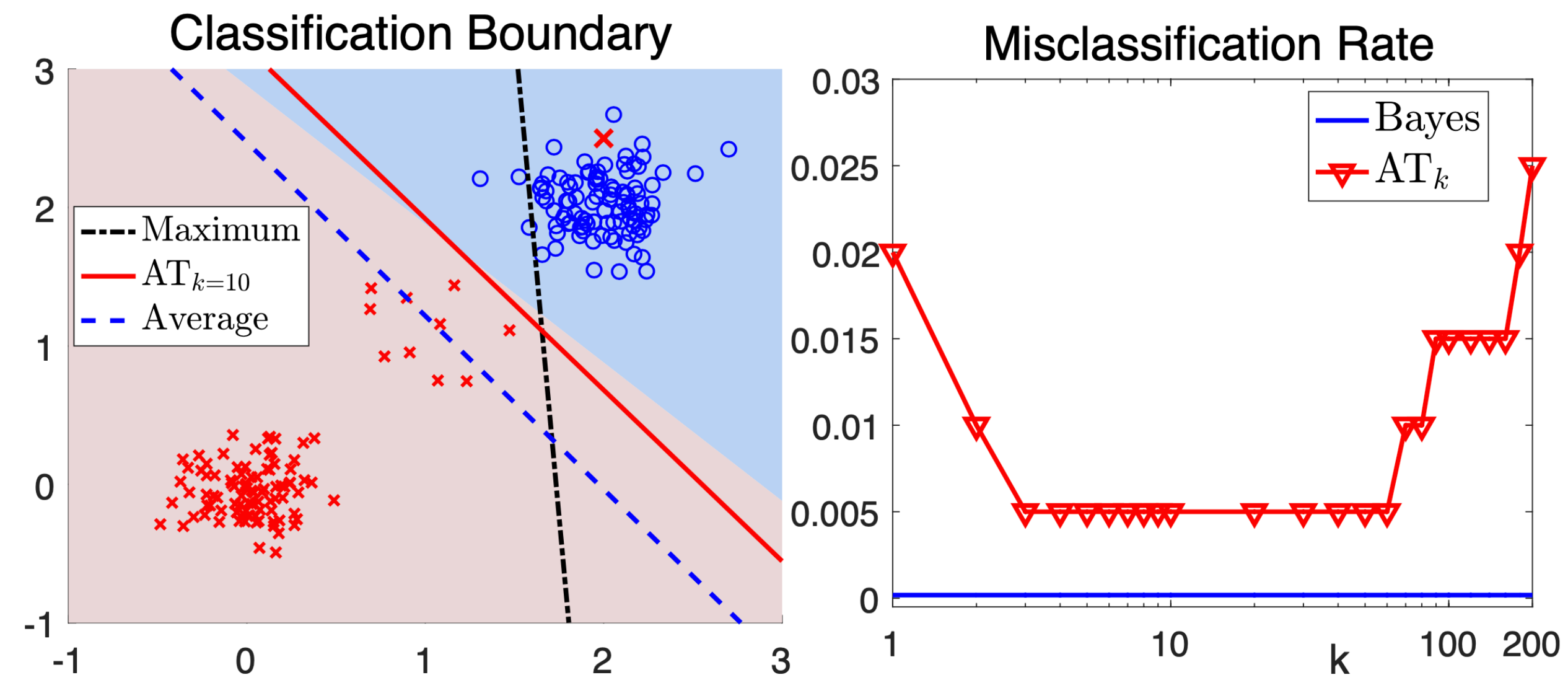
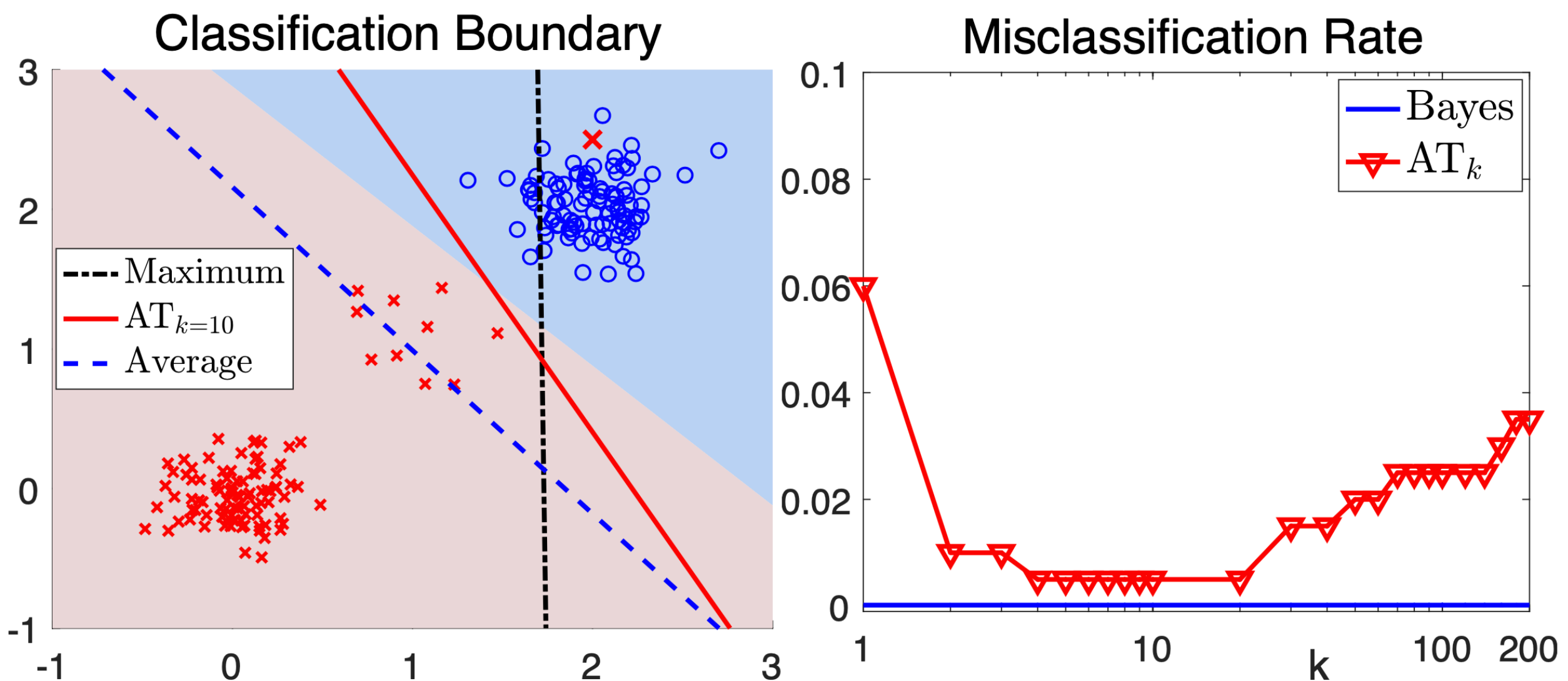
▶ 아웃라이어는 **X**로 표시됨



AVERAGE TOP-K LOSS

▶ Multimodal 데이터

- ▶ 평균 손실: 아웃라이어에 대해 잘 구분, 적은 분포는 잘 구분하지 못함
- ▶ 최대 손실: 적은 분포도 잘 구분, 아웃라이어를 잘 다루지 못해서 성능이 좋지 못함

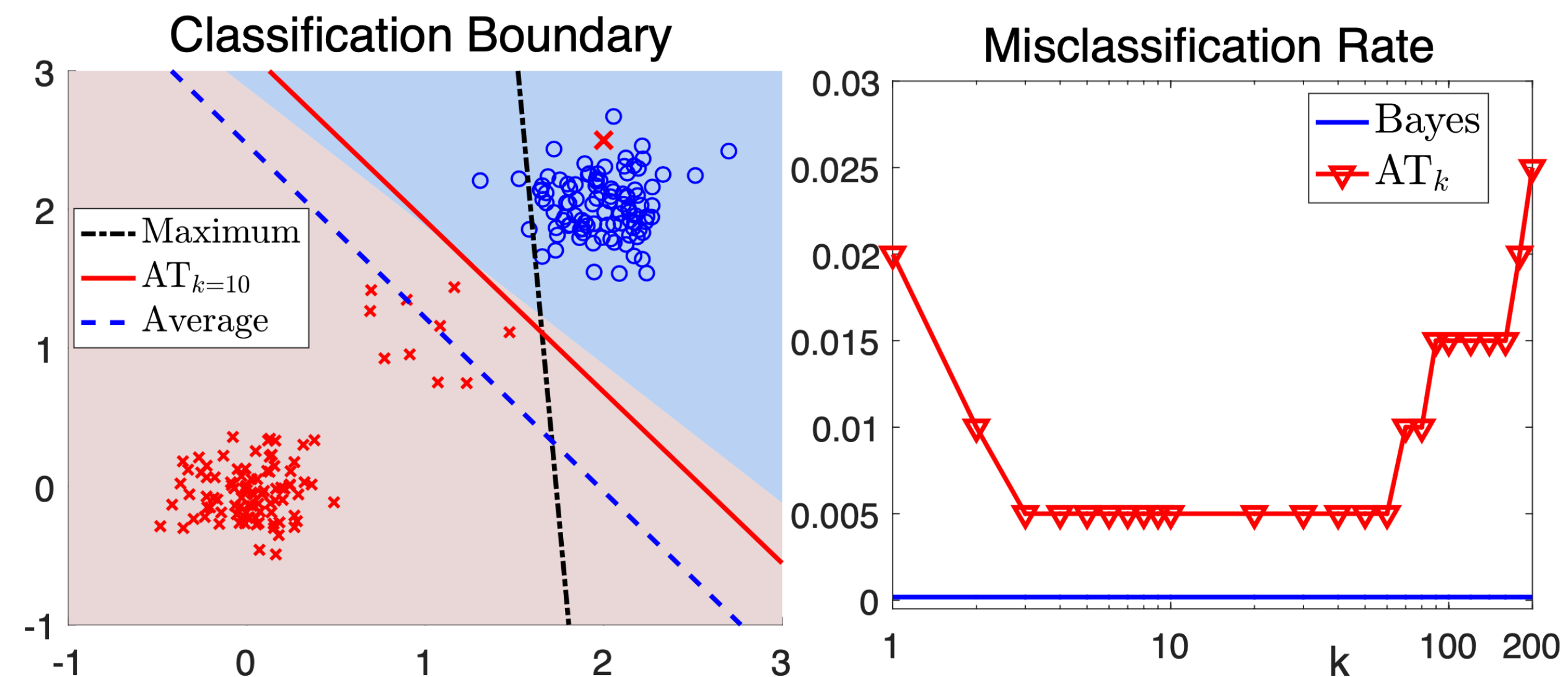
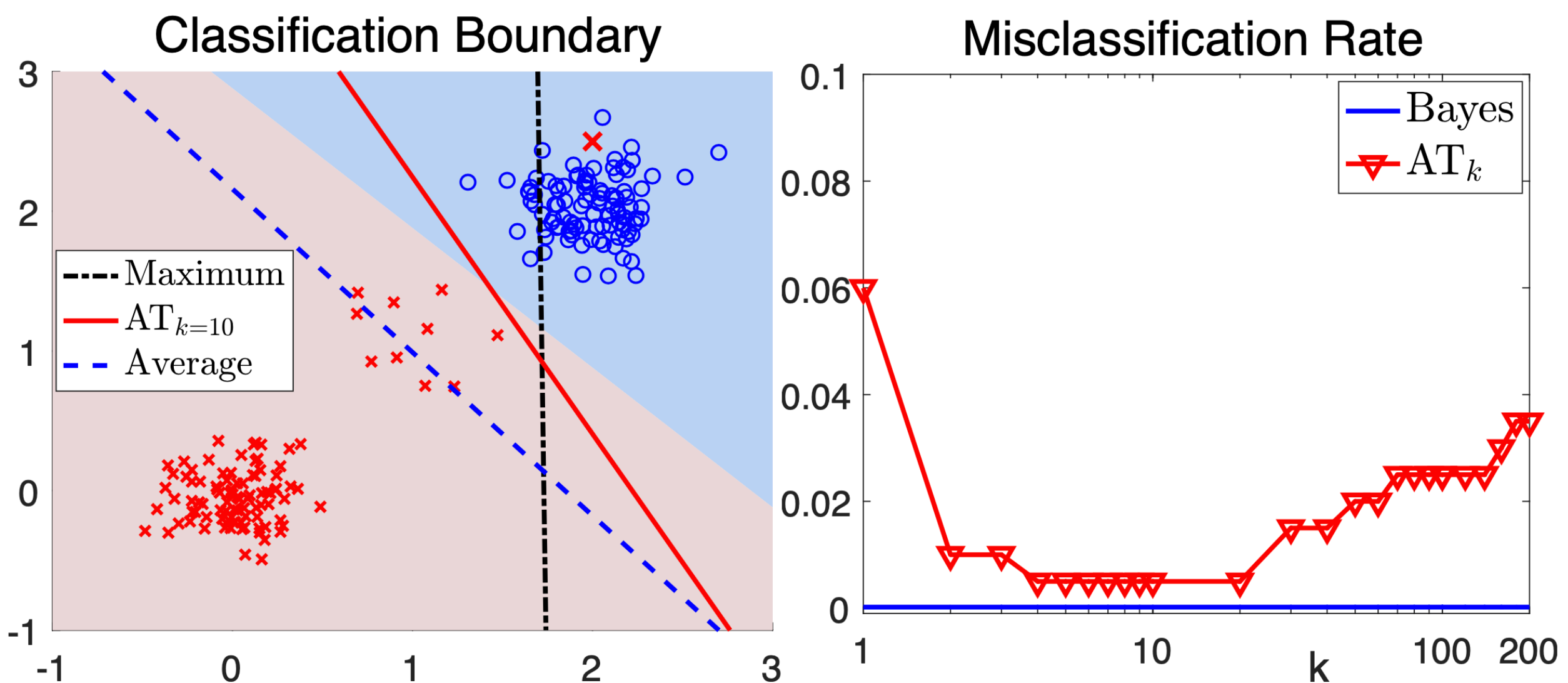


AVERAGE TOP-K LOSS

▶ Multimodal 데이터

▶ AT_k 손실: 최적 분류기에 가깝게 위치, 가장 좋은 성능

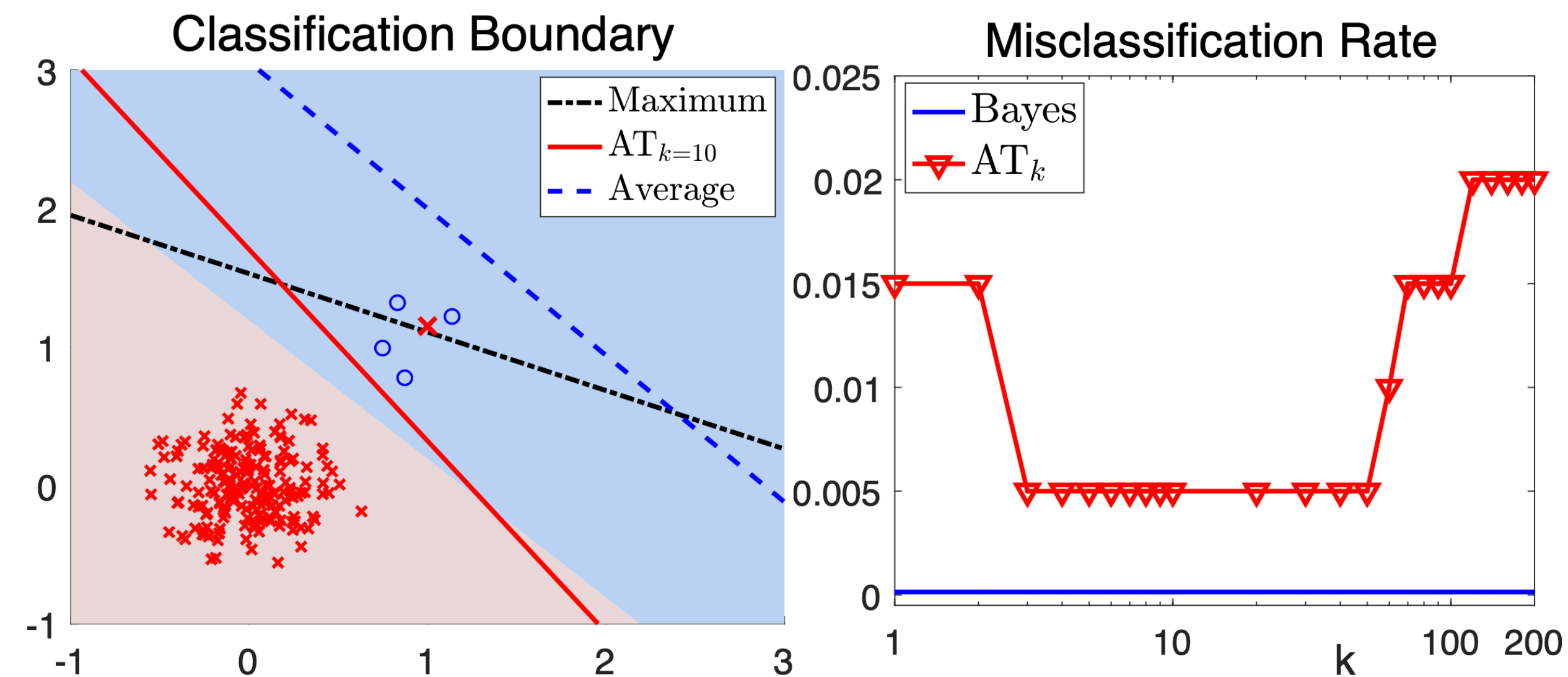
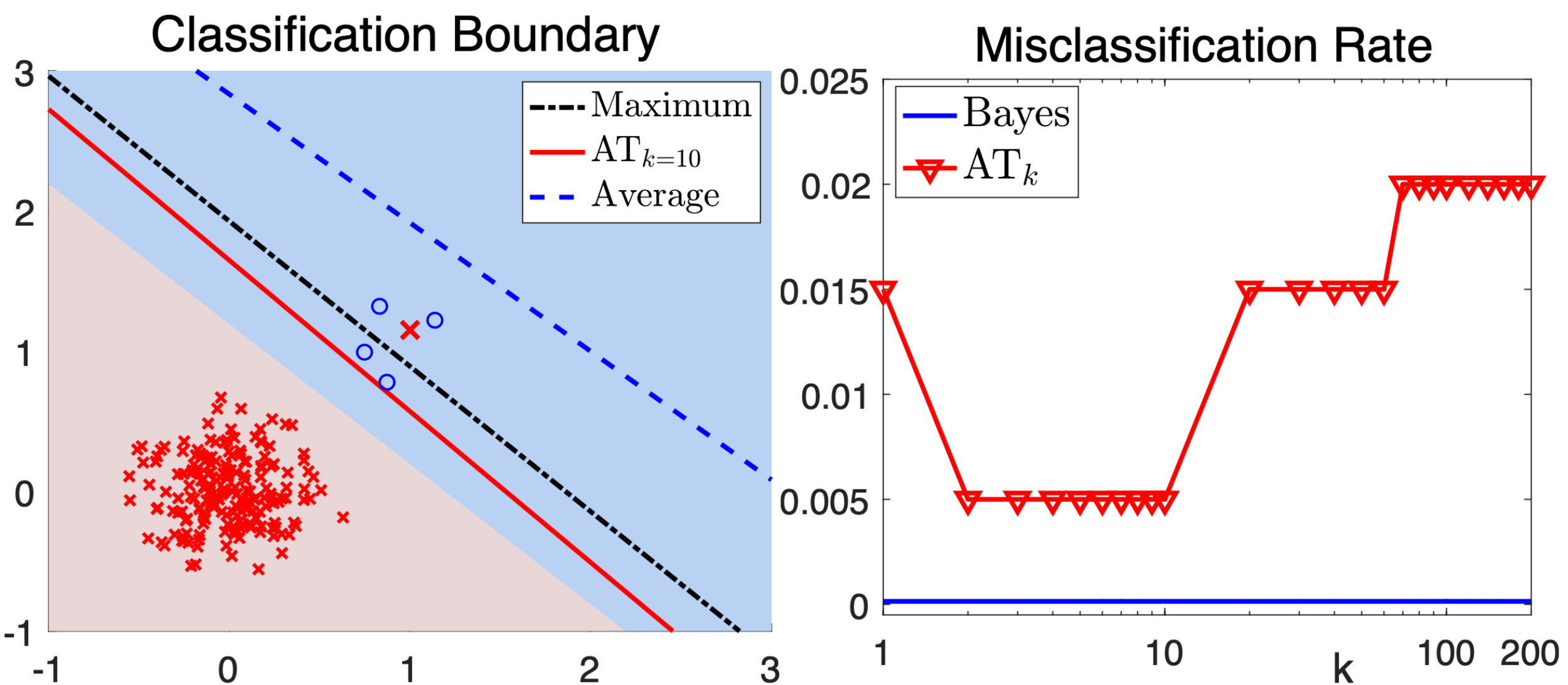
▶ 적절한 k 의 값은 grid search로 찾음



AVERAGE TOP-K LOSS

▶ 불균형 데이터

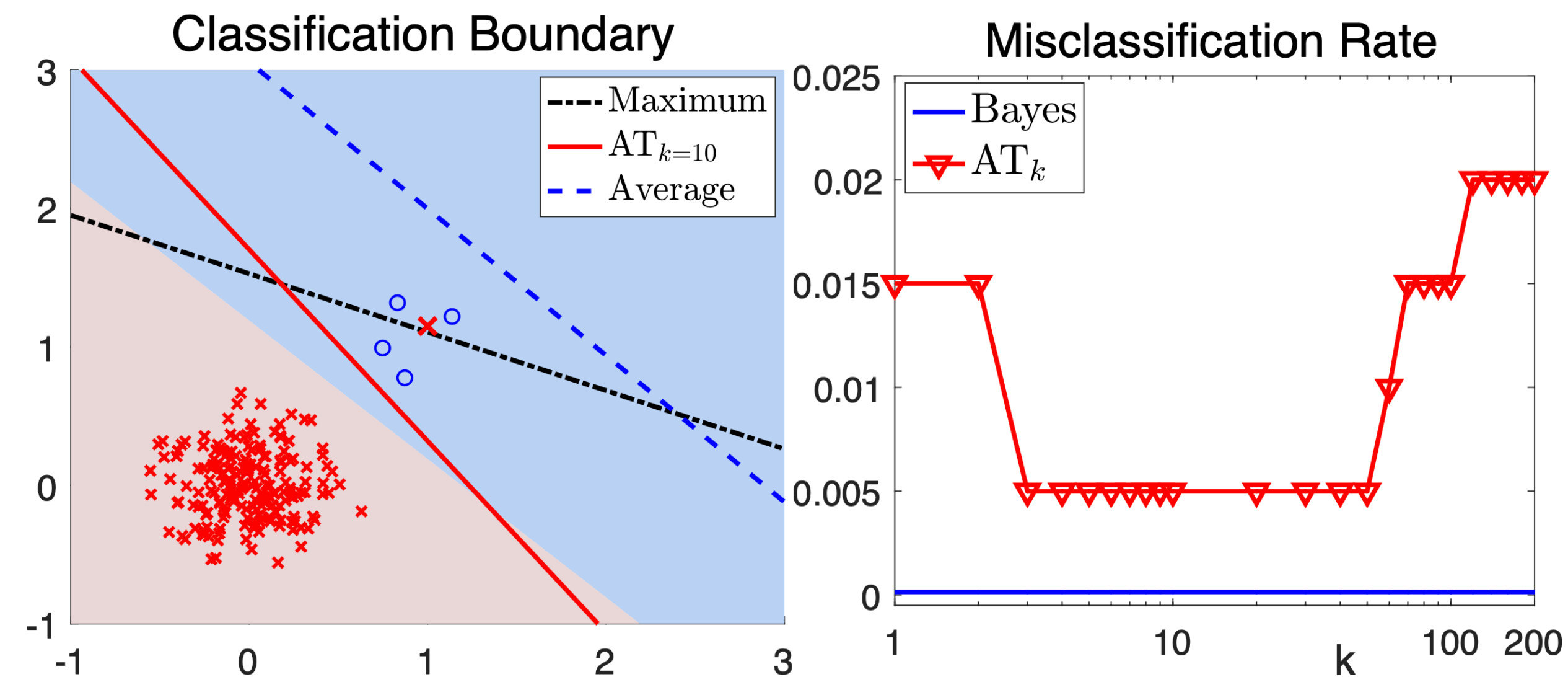
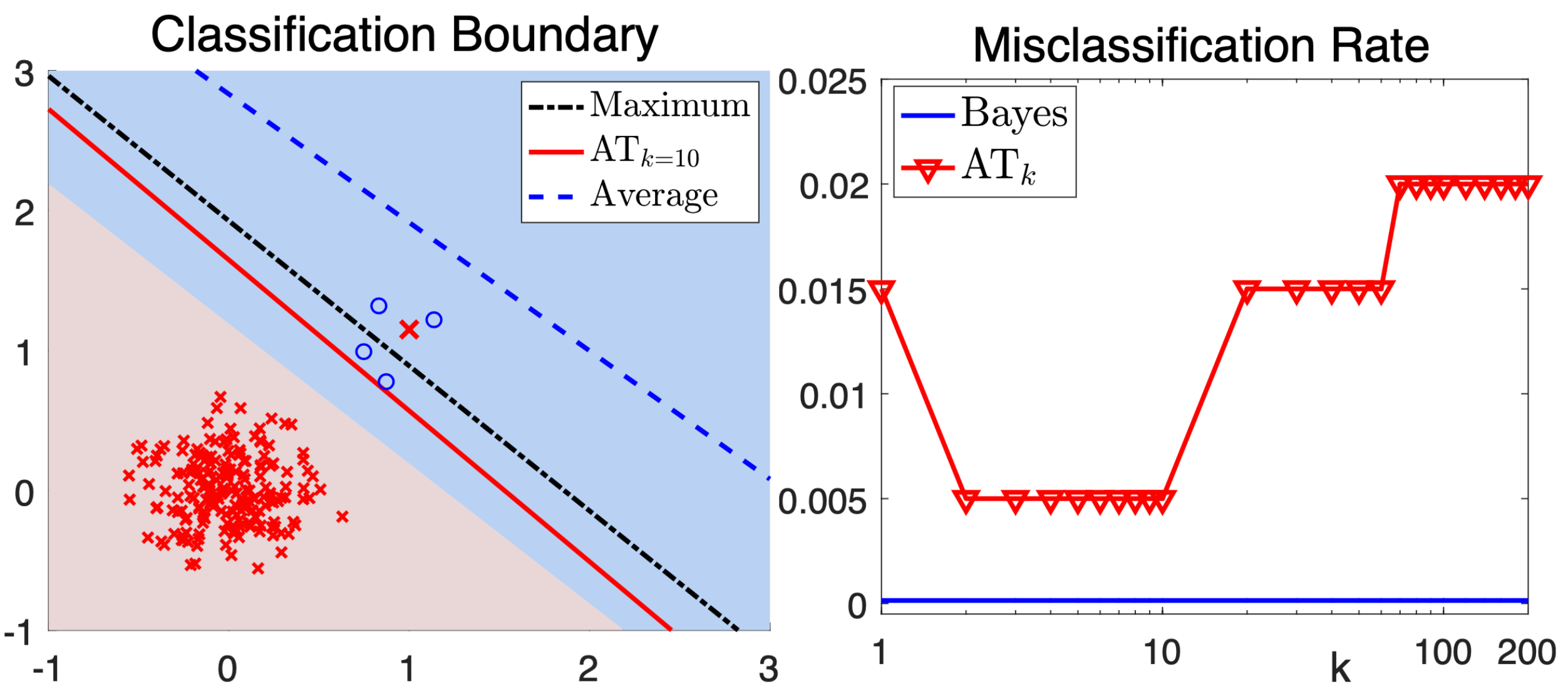
- ▶ 평균 손실: 레이블의 불균형 때문에 잘 구분하지 못함
- ▶ 최대 손실: 아웃라이어를 잘 다루지 못해서 성능이 좋지 못함



AVERAGE TOP-K LOSS

▶ 불균형 데이터

▶ AT_k 손실: 최적 분류기에 가깝게 위치, 가장 좋은 성능



AVERAGE TOP-K LOSS

- ▶ Average top-k (AT_k) 손실
 - ▶ 평균 및 최대 손실의 장단점의 균형을 맞춘 손실
 - ▶ 불균형하고 multimodal한 데이터 분포에도 잘 적응
 - ▶ 일반화

AVERAGE TOP-K LOSS

- ▶ Convex함
 - ▶ 증명 생략 (원 논문 참고)
 - ▶ 경사 하강법으로 효율적으로 수렴
 - ▶ 최적의 해를 구할 수 있음

EXPERIMENTS

EXPERIMENTS

- ▶ Binary 분류 문제와 regression 문제에서 각각
 - ▶ AT_k 손실과 평균 및 최대 손실을 비교
- ▶ MAT_k 학습에 경사 하강법을 사용해 최적화

EXPERIMENTS

- ▶ Binary 분류 문제
- ▶ 8개 데이터셋에 대해 평균 잘못 분류한 정도 (%)

	Logistic Loss			Hinge Loss		
	Maximum	Average	AT_{k^*}	Maximum	Average	AT_{k^*}
Monk	22.41(2.95)	20.46(2.02)	16.76(2.29)	22.04(3.08)	18.61(3.16)	<u>17.04(2.77)</u>
Australian	19.88(6.64)	14.27(3.22)	11.70(2.82)	19.82(6.56)	14.74(3.10)	<u>12.51(4.03)</u>
Madelon	47.85(2.51)	40.68(1.43)	39.65(1.72)	48.55(1.97)	40.58(1.86)	<u>40.18(1.64)</u>
Splice	23.57(1.93)	17.25(0.93)	16.12(0.97)	23.40(2.10)	<u>16.25(1.12)</u>	<u>16.23(0.97)</u>
Spambase	21.30(3.05)	8.36(0.97)	8.36(0.97)	21.03(3.26)	7.40(0.72)	7.40(0.72)
German	28.24(1.69)	25.36(1.27)	23.28(1.16)	27.88(1.61)	<u>24.16(0.89)</u>	<u>23.80(1.05)</u>
Titanic	26.50(3.35)	22.77(0.82)	22.44(0.84)	25.45(2.52)	<u>22.82(0.74)</u>	22.02(0.77)
Phoneme	28.67(0.58)	25.50(0.88)	24.17(0.89)	28.81(0.62)	22.88(1.01)	22.88(1.01)

EXPERIMENTS

- ▶ Regression 문제
- ▶ 4개 데이터셋에 대한 Root Mean Square Error (RMSE) 평균

	Square Loss			Absolute Loss		
	Maximum	Average	AT_{k^*}	Maximum	Average	AT_{k^*}
Sinc	0.2790(0.0449)	0.1147(0.0060)	0.1139 (0.0057)	0.1916(0.0771)	0.1188(0.0067)	0.1161(0.0060)
Housing	0.1531(0.0226)	0.1065(0.0132)	0.1050 (0.0132)	0.1498(0.0125)	0.1097(0.0180)	0.1082(0.0189)
Abalone	0.1544(0.1012)	<u>0.0800</u> (0.0026)	0.0797 (0.0026)	0.1243(0.0283)	0.0814(0.0029)	0.0811(0.0027)
Cpusmall	0.2895(0.0722)	0.1001(0.0035)	0.0998 (0.0037)	0.2041(0.0933)	0.1170(0.0061)	0.1164(0.0062)

DISCUSSION

DISCUSSION

- ▶ 일반화된 방법인 AT_k 를 제시
 - ▶ 장점을 결합하고 단점을 완화된 해결책
- ▶ 단순한 경사 하강법 외에도
 - ▶ 더 강력한 최적화 방법을 사용하는 것을 연구 중
- ▶ 비지도 학습 등에 적용도 중요한 주제가 될 수 있을 것

AVERAGE TOP-K LOSS

LEARNING WITH ~