Outro

데이터 분석 들어가기

- 데이터 분석 프로세스
  문제 정의 - 데이터 수집 - 데이터 전처리 - 데이터 분석 - 리포팅 피드백

- 정규화와 데이터 스케일링
  정규화(normalization): min-max normalization, Z-score normalization
  스케일링: Log / power / exponential scaling

통계적 분석

- 확률의 성질
- 변수, 확률변수, 확률 분포, 확률 함수의 정의
- 통계적 분석은 모수 추정의 과정

- 기댓값과 분산의 성질
- 독립의 정의(공분산), 결합확률분포
- 이산 확률분포: Bernoulli, Binomial, Poisson
- 연속 확률분포: Uniform, Normal(Gaussian), Standard normal(Z-score)

- Unbiased estimate(표본 분산)
- 중심극한정리

통계적 분석

- Pandas Dataframe을 이용한 기술통계값

- 통계적 추정의 정의
- 점 추정(MLE)
- 구간 추정(t분포: 모평균 구간 추정, 카이 제곱 분포: 모분산 구간 추정)

- 가설검정의 정의 및 오류(type 1 error, 유의수준)
- 모평균 차에 대한 가설 검정(Z-score, t-score), 모분산 비에 대한 가설검정(F분포)
- ANOVA(SSB, SSE, SST, F분포)

데이터 시각화

- Matplotlib label, tick, legend, marker, color등 기본 문법
- subplots, axes 활용
- plt.scatter, plt.bar, plt.hist, plt.boxplot, plt.violinplot

- Seaborn relplot: 2개 이상의 변수 간의 관계
  scatter plot, line plot
- Seaborn displot: 1개 이상의 변수 값의 분포
  hist plot, kde, heatmap, contour plot
- Seaborn catplot: 범주형 데이터의 분포
  strip plot/swarm plot, box plot/violin plot, bar plot/point plot

회귀 분석/데이터 모델링

- Regression / Classification task
- 데이터 모델링에서 MLE, MAP의 의미

- Linear regression의 cost function, Ordinary Least Squares, Gradient descent
- Overfitting, regularization(Ridge, Lasso)
- Logistic regression, SVM

- Random forest, decision tree
- Naive bayes, bayes theorem

- Regression evaluation: MSE, MAE, R-square
- Classification evaluation: Precision/recall(False positive란?), F1 score
- Feature analysis: .coef, .feature_importances_, pearsonr, spearmanr

마치며...

- Pattern recognition and machine learning(Bishop)
- 결국 머신러닝, 딥러닝의 모든 기초는 통계와 수학
- 수식을 이해하고 수식으로 표현하려고 하는 습관
- 데이터 시각화는 다른 사람들의 예시를 최대한 많이 참고(논문, 보고서)