M1399.000500:
통계적 기계학습 (Statistical Machine Learning)
2025년 1학기 (1st Semester)

김지수 (Jisu KIM)

2025-03-05

기계학습(Machine Learning)에는 여러 분야들이 있습니다.

- 기계학습에는 여러 분야들이 있지만, 대표적으로 다음과 같습니다.
- 지도학습(Supervised Learning)
  - 입력변수를 이용하여 출력변수를 예측
  - 회귀(regression), 분류(classification) 등
- 비지도학습(Unsupervised Learning)
  - 입력변수 간의 복잡한 관계를 규명(출력변수가 없음)
  - 군집분석(clustering), 주성분분석(principal component analysis), 요인분석(factor analysis) 등
- 강화학습(reinforcement learning)
  - 행동에 따라 변화하는 환경에서 의사결정 방법을 학습
  - multi-armed bandit problem, markov decision process 등
- 이 수업에선 대부분 지도학습(Supervised Learning)을 다루고, 비지도학습(Unsupervised Learning)을 살짝 다룹니다.

통계적 기계학습(Statistical Machine Learning)에서는 기계학습의 통계적 이론을 배웁니다.

- 자료(data)는 항상 랜덤한 잡음(noise)을 가지고 있습니다.
- 랜덤한 잡음(noise)이 기계학습 알고리즘에 어떠한 영향을 끼치는지, 알고리즘이 잘 작동하는지 알기 위해서는 통계적 분석이 필요합니다.
- 기계학습에서 통계 이론은 다음과 같은 분석을 해줍니다:
  - 과적합(overfitting)의 이해
  - 차원의 저주(curse of dimensionality) 이해
  - 알고리즘의 통계적 이론: 일치성(consistency), 최적합(optimality), 불확실성의 정량화(uncertainty quantification) 등

지도학습(Supervised Learning)의 기본 모형

- 입력(Input) / 설명 변수(Covariate) : $x \in \mathbb{R}^d$
- 출력(Output) / 반응 변수(Response): $y \in \mathcal{Y}$
- 모형(Model) : $y \approx f(x)$, $f \in \mathcal{F}$
- 손실 함수(loss function): $l(y, a)$
- 목적: $f^0 = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(Y,X)} [l(Y, f(X))]$ 을 찾는 것을 목적으로 합니다.
- 예측: $\hat{f}$가 추정한 함수이고, 새로운 자료 x가 들어오면, y를 $\hat{f}(x)$로 추정합니다.

선형 분류기(Linear Classifier)

- class 가 두 가지인 경우를 생각합니다. (즉, $\mathcal{Y} = \{0, 1\}$ 또는 $\mathcal{Y} = \{-1, 1\}$)
- 선형 분류기(Linear Classifier)는 결정 경계(decision boundary)가 선형이 되는 모형을 사용합니다. 즉, 결정 경계가 다음과 같은 형태입니다:
$$\{x : \beta_0 + x^\top \beta = 0\}.$$
- 이는, $\mathcal{Y} = \{-1, 1\}$일 때, 선형함수 $f(x) = \beta_0 + x^\top \beta$에 대해, 해당하는 분류기(classifier) $G : \mathbb{R}^p \to \mathcal{Y}$를 다음과 같이 만드는 것과 같습니다:
$$G(x) = \mathrm{sign} f(x).$$

선형 분류기(Linear Classifier) 중 선형판별분석(Linear Discriminant Analysis)와 로지스틱 분류(Logistic Classificaion)를 다룹니다.

- 선형 분류기(Linear Classifier)에는 크게 세 가지 접근방식이 있습니다:
- 결합확률분포(joint probability distribution)을 모형화합니다 : 선형판별분석(Linear Discriminant Analysis), 나이브 베이즈 분류기 (Naive Bayes Classifier) 등
- 조건부밀도함수 $P(\mathrm{class}|X)$를 모형화합니다: 다수의 회귀분석 (regression) 기반 분류기(classifier), 특히 로지스틱 회귀(Logistic Regression)를 이용한 로지스틱 분류(Logistic Classification)
- 다른 그룹 간의 차이를 최대화합니다: 퍼셉트론(perceptron), 서포트 벡터 머신(support vector machine) 등
- 이 수업에선 선형판별분석(Linear Discriminant Analysis)과 로지스틱 분류(Logistic Classification)를 다룹니다.

선형 회귀(Linear Regression)에서 자료의 차원(dimension)이 너무 클 경우 문제가 생깁니다.

- 선형 회귀(Linear Regresison) 문제 $Y = X\beta + \epsilon$, $X \in \mathbb{R}^{n\times d}$에서, 자료의 차원(dimension) d가 표본의 수 n에 비해서 상대적으로 클 때, 최소제곱법(least square method)에 기반한 간단한 회귀 모형을 적합하는 데에 여러 문제가 생깁니다.
- 다중공선성(multicollinearity): 설명 변수들이 서로 상관이 커져서 β의 추정을 어렵게 합니다. 예를 들어, $d > n$이면 최소제곱법이 $\hat{\beta}$의 값을 유일하게 결정하지 못합니다.
- 과적합(overfitting): 적합하는 모형(fitted model)이 참 모형(true model)에 비해 너무 복잡해지고, 차원의 저주(curse of dimensionality)와 같은 현상을 유발합니다.

선형 회귀(Linear Regression)에서 자료의 차원(dimension) 문제를 해결하는 방법으로 축소추정량(Shrinkage Estimator)을 사용합니다.

- 선형 회귀(Linear Regresison) 문제 $Y = X\beta + \epsilon$, $X \in \mathbb{R}^{n \times d}$에서, 자료의 차원(dimension) p가 표본의 수 n에 비해서 상대적으로 클 때, 여러 해결 방법이 있습니다.
- 변수 선택(Variable Selection): 최적부분집합선택(Best Subset Selection), 전진선택(Forward Selection), 후진선택(Backward Selection) 등
- 축소추정량(Shrinkage Estimator): 능선회귀(Ridge Regerssion), 라쏘(Lasso) 등
- 차원축소(Dimension Reduction): 주성분회귀(Principal Component Regression), 부분최소법(Partial Least Square) 등
- 이 수업에선 축소추정량(Shrinkage Estimator), 특히 능선회귀(Ridge Regerssion)와 라쏘(Lasso)를 다룹니다.

비선형 회귀(non-linear regression)를 푸는 데에 기저 전개 (basis expansion) 또는 핵평활(kernel smoothing)을 활용합니다.

- 선형 회귀(linear regresison) $Y = X\beta + \epsilon$, $X \in \mathbb{R}^{n \times d}$에서는 평균반응 $\mathbb{E}[Y|X] = X\beta$이 X에 대해 선형인데, 이런 선형 관계가 깨질 수 있습니다.
- 비선형 회귀(non-linear regresison) $Y = f(X) + \epsilon$ 를 푸는 데에는 두 가지 방법이 있습니다.
- 기저 전개(basis expansion): 설명변수 X에 비선형 변환(non-linear transformation) 을 취한 후 거기에서 선형 회귀(linear regression)를 풉니다.
  - 회귀함수가 $f(x) = \sum_{j=1}^{q} \beta_j h_j(x)$ 와 같은 형태가 됩니다. 이 때, 각 $h_j : \mathbb{R}^d \to \mathbb{R}$ 을 기저함수(basis function)라 합니다.
  - 다항회귀(polynomial regression), 회귀스플라인(regression spline), 평활스플라인(smoothing spline) 등이 있습니다.
- 핵평활(kernel smoothing): 핵함수(kernel function)을 사용하여 국소적(local)으로 선형 회귀(linear regression)를 풉니다.
  - Nadaraya-Watson 추정량, 국소선형회귀(local linear regression) 등이 있습니다.

앙상블(Ensemble)은 여러 학습기(learner)를 모아 하나의 성능 좋은 학습기를 만듭니다.

- 앙상블(Ensemble)은 여러 학습기(learner) / 분류기(classifier)를 모아 하나의 성능 좋은 학습기 / 분류기를 만듭니다.
- 배깅(Bagging), 부스팅(Boosting), 랜덤포레스트(Random Forest) 등이 있습니다.

모형 선택(Model Selection)과 모형 평가(Model Assessment)

- 모형 선택(Model Selection): 다른 모형(model)들의 성능을 비교해 최적 모형을 선택합니다.
- 모형 평가(Model Assessment): 최적 모형을 선택한 후 그 모형의 예측 오차(prediction error) 내지는 일반화 오차(generalized error)를 추정합니다.
- 모형 선택을 위해선 모형 평가를 정확하게 할 필요는 없고, 서로 다른 모형 간에 줄세우기만 할 수 있으면 됩니다. (가장 좋은 모형, 두 번째로 좋은 모형, etc)

모형 선택(Model Selection)의 여러 방법들을 배웁니다.

- 모형 선택(Model Selection)의 여러 방법들을 배웁니다:
- AIC(Akaike Information Criterion), BIC(Bayesian Information Criterion), 교차타당성(Cross Validation), 부트스트랩(Bootstrap) 등

고차원에서 함수를 추정하는 여러 기법들을 알아봅니다.

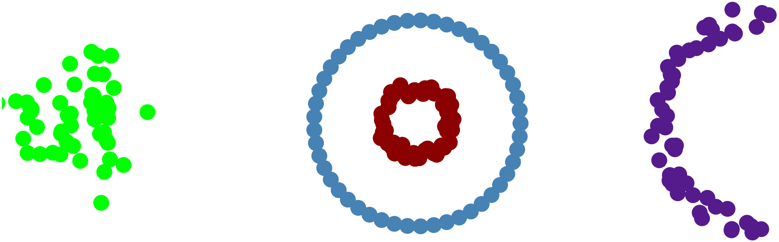
- 저차원에서는 함수를 추정하는 여러 방법들이 있습니다.
- 고차원에서는 차원의 저주(curse of dimensionality)로 인해 잘 작동하지 않습니다.
- 고차원에서 함수를 추정하는 여러 기법들을 알아봅니다: 일반화가법모형(Generalized Additive Models), 사영추적회귀 (Projection Pursuit Regression), 다변량 적응 회귀스플라인 (Multivariate Adaptive Regression Spline, MARS) 등

Reproducing Kernel Hilbert Space는 기저 전개에서 무한한 개수의 기저를 활용하도록 해 줍니다.

- 앞서 비선형 회귀(non-linear regresison)를 푸는 데에 기저 전개(basis expansion)나 핵평활(kernel smoothing)을 활용하였습니다.
- 기저 전개에서는 무한한 개수의 기저를 명시적으로 계산할 수 없기 때문에 기저의 개수가 유한개로 제한됩니다.
- 이를 해결하는 기법 중에 하나가 RKHS(Reproducing Kernel Hilbert Space) 입니다.
- 회귀함수가 $f(x) = \sum_{j=1}^{\infty} \beta_j h_j(x)$ 와 같이 무한한 기저 함수의 합으로 나타내더라도, RKHS 에서는 $f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$와 같은 형태로 나타내집니다. 이 때 $k(\cdot, \cdot)$가 kernel 함수입니다.
- 다른 일반적인 경우에도 선형 문제를 비선형 문제로 확장할 수 있는 중요한 방법론입니다.

군집분석은 자료들을 같은 군집에 속하는 자료들끼리 좀 더 유사하도록 몇 개의 군집으로 묶는 과정입니다.

- 자료에 목표가 되는 군집이 표기되어 있지 않기 때문에, 군집분석은 비지도학습으로 분류됩니다.
- 다양한 방법론이 있습니다: k-means, mixture models, density-based clustering, hierarchical clustering, spectral clustering 등
- 여러 군집분석 방법들의 관련 이론을 배웁니다.



Concentration inequality는 확률변수의 행태를 확률적으로 통제합니다.

- 약한 대수의 법칙(law of large number)은, $X_1, \ldots, X_n$이 iid 자료이고 모평균 $\mathbb{E}[X_1] = \mu < \infty$일 때, 표본평균이 모평균으로 확률수렴함을 뜻합니다:

$$\lim_{n \to \infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0 \text{ for all } \epsilon > 0.$$

- 표본평균을 더 일반적인 함수로 확장시켰을 때 확률변수의 행태를 다음과 같이 확률적으로 부등호로 통제하는 것 (및 이에 필요한 기술들)을 Concentration inequality라 합니다:

$$P\left(\sup_{f \in \mathcal{F}} |f(X_1, \ldots, X_n) - \mathbb{E}\left[f(X_1, \ldots, X_n)\right]| > \epsilon\right) < \delta.$$

Concentration inequality는 확률변수의 행태를 확률적으로 통제합니다.

- 표본평균을 더 일반적인 함수로 확장시켰을 때 확률변수의 행태를 다음과 같이 확률적으로 부등호로 통제하는 것 (및 이에 필요한 기술들)을 Concentration inequality라 합니다:

$$P\left(\sup_{f\in\mathcal{F}} \left|f(X_1,\ldots,X_n) - \mathbb{E}\left[f(X_1,\ldots,X_n)\right]\right| > \epsilon\right) < \delta.$$

- 다음과 같은 concentration inequality를 배웁니다: Hoeffding's inequality, McDiarmid's inequality, Bernstein's inequality, Rademacher Complexity, VC dimension, uniform bound 등

추정량(estimator)의 최대위험(maximum risk)은 추정량의 최악의 경우에 오차의 기대값(expected error)입니다.

- 추정량(estimator) $\hat{\theta}_n$ 의 최대위험(maximum risk)은 최악의 경우에 추정량 $\hat{\theta}_n$ 이 만들어낼 수 있는 오차의 기대값(expected error)입니다.
-
$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\ell \left(\hat{\theta}_n(X),\ \theta(P) \right) \right]$$
    - $X = (X_1, \cdots, X_n)$는 고정된 분포 P에서 추출하고, P는 확률분포의 집합 $\mathcal{P}$에 속합니다.
    - 추정량 $\hat{\theta}_n$ 은 자료 X의 임의의 함수입니다.
    - 손실함수 $\ell(\cdot, \cdot)$는 추정량 $\hat{\theta}_n$의 오차를 잽니다.

미니맥스 위험(minimax risk)은 모수(parameter) 추정의 통계적 어려움을 묘사합니다.

- 미니맥스 위험(minimax risk) R_n은 최악의 경우에도 잘 작동하는 추정량(estimator)의 위험(risk)입니다. 이를 표본크기(sample size)의 함수로 봅니다.
- $$R_n = \inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\ell \left(\hat{\theta}_n(X),\ \theta(P) \right) \right]$$
  - $X = (X_1, \cdots, X_n)$는 고정된 분포 P에서 추출하고, P는 확률분포의 집합 $\mathcal{P}$에 속합니다.
  - 추정량 $\hat{\theta}_n$ 은 자료 X의 임의의 함수입니다.
  - 손실함수 $\ell(\cdot, \cdot)$는 추정량 $\hat{\theta}_n$의 오차를 잽니다.

심층학습은 여러 층으로 쌓인 구조로 되어 있는 모형을 학습합니다.

- 심층학습은 폭발적인 발전에 비해 통계적 이론 분석은 매우 더딘 편입니다.
- 심층학습에 관련된 통계 학습 이론을 배웁니다.