ECOM90025 Advanced Data Analysis

Tutorial 12 - a quick revision & brief summary[1]

Zheng Fan

The University of Melbourne

[1] Selected contents from the lecture materials. The slides do not include everything we learned.

Introduction

Zheng Fan

- Ph.D. student in Economics
- Email: fan.z@unimelb.edu.au
- Tutorial code and slides: github.com/zhengf1/ADA2022

Seek help?

- Ed discussion board
- Consultations: refer to Canvas for details

Bootstrap

To understand estimation uncertainty: such as accessing the standard error of $\bar{X}$

1. non-parametric bootstrap: keep resampling from our actual data.
2. parametric bootstrap: it generates the sample from the distribution we expect.

Sample variance is a biased estimate of the true population variance $\rightarrow$ need de-biase.

Bootstrap of regressions: randomly draw n rows with replacement as data to run regression

False Discovery (FD)

$$\text{FD Proportion} = \frac{\#\text{ false positive}}{\#\text{ tests called significant}}$$

If you have 1000 noise variables and test with 5% significance level, you would expect about 50 false discoveries.

$\rightarrow$ Using the Benjamini-Hochberg logic

k-Fold Cross-Validation

To evaluate machine learning models through out-of-sample fit.

A popular choice ($k = 10$) was found to provide a good trade-off of computational cost and bias in an estimate of model performance.

LASSO: least absolute shrinkage and selection operator

$$min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

The choice of λ

- AIC; BIC: less favorable in this contents
- AICc: embedded in “gamlr" function. To calculate, there are functions available.
- **cross validation**: “cv.gamlr" (or you may code yourself)

Multinomial Logistic Regression

emmm. I have nothing to mention here.

Treatment Effect (TE)

AB trial is also known as a completely randomized design

- A is the control group and B is the treatment group
- Average Treatment Effect (ATE)

$$ATE = E(y|d = 1) - E(y|d = 0) = \bar{y}_1 - \bar{y}_0$$

Difference-in-Differences:

$$Y = \alpha + \gamma \cdot time + \lambda \cdot intervention + \delta\,(time \cdot intervention) + \varepsilon$$

The treatment is $(time \cdot intervention)$, so the TE is δ

The main identifying assumption of DiD:

- common/parallel "trends" in outcomes in treated and control groups.

PCA, PCR, PLS

In PCA, the first component T_1 is constructed by taking

$$T_1 = \gamma_1^T X$$

where $||\gamma_1|| = 1$, and maximize $var(T_1)$

- can be easily obtained via "T = prcomp(X, rank = p, scale=TRUE)".

In PLS, the first component T_1 is constructed by taking

$$T_1 = \phi_1^T X$$

where $||\phi_1|| = 1$, and maximize $cov(Y, T_1)$

- can be easily done by repeatedly estimating marginal regression (algorithm 20 & 21).

Topic models

Latent Dirichlet Allocation (LDA):

1. A topic k is a probability vector θ_k
2. Each θ_k is a vector. With probability θ_{kj}, word j could be generated if it is from topic k

See a graphical representation:

CART

Nonparametric Modelling: to grow a tree.

1. x_i is a vector. Split on an element x_{ij}.
2. splitting at jth element produces two children left:$\{x_k, y_k : x_{kj} \leq x_{ij}\}$, and right:$\{x_k, y_k : x_{kj} > x_{ij}\}$
3. Loss function: $\sum_{k \in \text{left}} (y_k - \overline{y}_{\text{left}})^2 + \sum_{k \in \text{right}} (y_k - \overline{y}_{\text{right}})^2$

We may use cross-validation to prune the tree: the number of leaves.

Trees have high variance: similar samples can produce very different trees.

- Bagging is a way to reduce the variance (Bootstrapping).
- Drawback with bagging: the B trees are not independent.
- Random forests (RF): reduces the correlation between the B of trees, by growing the bootstrap trees in a specific way.

The end

Thanks for your attention! 

Tutorial code and slides: github.com/zhengf1/ADA2022

Good luck with your final project!