Good Practices in Data Science

Elena Tuzhilina

Description

What is the goal (SMART) / deliverable?

Currently there are quite many recommendations for the "right way" to do DS. The practical goal of this project is not only to compile the list of such recommendations but also to explain and provide strong evidence (through simulation or real data example) why it is important to follow these recommendations. It is aimed to help building a bridge between statistics and other fields of applications to people with different level of expertise.

- *6 week goal*: make final decisions on the format, start collecting ideas and references, try to implement one idea;
- *Longer term goal*: set up a biweekly updated platform, find other people in the DS community that would be interested in contributing to the subject (build a network);

Time needed to complete the project?

- 2 weeks to write a new post/chapter
- 4 months to build a base for the platform

Resources needed?

- *people*: 1-4 people (about 10 hours per one concept)
- *money*: for blog for website server and platform services (?)

How does this help you? How does this help DSI? How does this help Stanford? How does this help the world?

This project will allow me to structure the statistical background I currently have. It could help people who recently started their journey in DS to avoid data misinterpretation while doing statistical analysis. It could be of a great interest for people who already have some experience in DS helping to develop deeper understanding of some statistical phenomena.

Ideas for brainstorm

Alternative names

- Good practices in Data Science
- Data Science pitfalls
- What If? In Data Science.
- How to treat your data well
- Tricky Data Science

Post/Chapter structure

1. Difficulty level
2. Preamble:
    - definition
    - intuition
3. Explaining consequences:
    - simulation and/or real data
    - well-known cases (papers criticized in media)
4. Theoretical justification (formal proof of some concepts)
5. Links to further reading

Format

- blog (what platform to use?)
- tutorial (Jupyter notebook, Google Colab)
- ebook (with code)
- course (Google classroom)

Good practices list

Some of these recommendations are data specific.

Procedure

1. **Problem:** Need to do test validation.
   Consequence: Could overfit train data.
   Solution: Can use held-out set or cross validation.

2. **Problem:** Need to take into account correlated observations in the data while doing train and test splits (e.g. if closed relatives are included in the data).
   Consequence: Results could be not replicable on new data set.
   Solution: Can use block cross validation.

3. **Problem:** Need to remove the effect of strong predictors (e.g. gender) when you study correlation between two variables.
   Consequence: Misinterpretation of the result (e.g. maybe gender explains 90% of your correlation).
   Solution: Can regress the effect of a strong predictor from both variables of interest.

Data transformation

4. **Problem:** Need scaling before doing regularization (e.g. ridge/lasso) as long as PCA and KNN.
   Consequence: The method will pay more attention to variables on larger scale.
   Solution: Do scaling before applying method.

5. **Problem:** Need centering before applying PCA.
   Consequence: The principal direction could be way off.
   Solution: Center the data before applying PCA.

Regression

6. **Problem:** Need to check for the outliers.
   Consequence: The fit could be way off.
   Solution: Plot the data before fitting, do some checks after (e.g. leverage statistics, Cook's distance).

7. **Problem:** Do not throw away all variable with large p-value.
   Consequence: Large p-value can be explained by correlation between predictors, so you can remove informative ones.
   Solution: Use some formal procedure (e.g. step-wise).

8. **Problem:** Do not pick between two sub-models based on R^2 only.
   Consequence: Could overfit the data.
   Solution: Need to account for number of parameters (e.g. use Cp or Adjusted R^2).

9. **Problem:** Misinterpretation of regression coefficients: the change in response if predictor is increased by one unit.
   Consequence: Incorrect conclusion drawn from data.
   Solution: Increase in response if predictor is increased by one unit and other predictors are fixed (so can be negative is there is some correlation in the data).

Testing

10. **Problem:** Misinterpretation of large p-value: we can accept H_1.
    Consequence: Incorrect conclusion drawn from data.
    Solution: Can't reject H_0.

11. **Problem:** Need to do correction while doing multiple testing.
    Consequence: Large FWER.
    Solution: Use Bonferonni, Tukey's, Holm's correction.

Data

12. **Problem:** Need regularization if number of predictors is compatible to the sample size.
    Consequence: Could overfit the data.
    Solution: Reduce number of features via PCA or apply regularization.
13. **Problem:** Need take into account class size if data is imbalanced.
    Consequence: Method could pay more attention to over-represented class.
    Solution: Use weighted metric or F1 score.

Conceptual

1. Correlation $\neq$ causation
2. Bias in data
3. Data-dredging

What If?

1. I run two sample t-test instead of paired one?
2. I do not check normality assumption before running t-test?
3. I do not check regression assumptions?

References

1. Wikipedia: Misuse of statistics.
2. Science Forum: Ten common statistical mistakes to watch out for when writing or reviewing a manuscript.
3. Statistical Mistakes and How to Avoid Them.
4. Common mistakes in statistics.
5. Reviewer's quick guide to common statistical errors in scientific papers.