The gap between market needs and education in AI: A viewpoint from Trusting Social

Mai Hai Thanh
Lead Data Scientist, Trusting Social

“Advancing Data Science and Technology to Deliver Financial Access for All”

2019-11

Before we begin

- Hard-learned lessons that are often not available in textbooks
- A mixture of data science, machine learning, and soft things
- More or less true for different cases
- This is a sharing session with university lecturers

A typical academic research flow

- Define problem
- Review related work
- Propose a new method/algorithm
- Evaluate performance
- Write and publish papers

A typical industrial research flow

- Define problem
- Review related work
- Propose a new method/algorithm
- Evaluate performance
- Deploy on production system

Not much difference at the first glance.

A typical industrial research flow

- Define problem: make sure you know
  - What your problem is
  - What the expected outputs are
  - **Why you should do it**

    Most students don't ask **Why**

    Don't spend 6 months working hard and figuring out that you are working on something nobody needs

A typical industrial research flow

- Review related work
  - What have been done by your team-mates?
  - What are the strengths and weaknesses of previous methods?

  Don't blindly start from scratch

  Don't live in an isolated island

A typical industrial research flow

- Propose a new method/algorithm
  - Asking teammates and managers to read 10k lines of code is not appropriate in most cases
  - Make sure you can describe clearly what you have done

    If the author cannot explain the new method clearly, who can?

A typical industrial research flow

- Evaluate performance
  - Make sure you know what metrics are best to measure your success
    - If you cannot measure your success, how can you know that you are doing well?

  - Accuracy, recall, precision, F1, FAR, TAR, AUC... choose one or two metrics that suit best to your need
    - Why is that the most suitable metric?
      Need to clearly understand the problem. Don't just blindly follow your advisor/supervisor/boss or even a random online tutorial.

ML without data? No way!

- Make sure you know
  - **What** kind of input data you have
    (text, image, transactional data?)
    Where & **How** you can get them
    (which sources? who?)
    When you can get them
    (every month/week/day/hour/minute/second?)

  - We will have to adjust the model development process depending on the What/Where/How/When aspects of the data

ML without data? No way!

- Input data is important, both **labels** and **raw** data to compute features
  - Garbage in -> garbage out
  - Labeled data is very expensive in many cases
    Examples: loan performance, faked ID
  - Students should also study about data collection

ML without data? No way!

- 20%+ of your time is to clean input data
  - Higher percentage if your data is really big
  - Don't be surprised, don't be sad
  - Be prepared with appropriate skills

The first model

Interviewer: what do you do?
Interviewee: I tried to solve problem X, using a supermarket's transactional data to predict customer demand for some special products.
Interviewer: how?
Interviewee: I created a deep neural network with 10 convolutional layers, 10 pooling layers, and 10 fully connected layers. The accuracy is 85%, which makes my supervisor happy.

The first model

A junior AI engineer: I will work hard 3 months and build a sophisticated model that can achieve 90% accuracy. I think 90% is a good target.

Boss: What! Our team could get 91% accuracy with just 10 lines of code.

The first model

- The baseline: your first ML model should be very simple (but not too simple)
    - Can be done quickly, still give a decent accuracy
    - More complex ML models must be compared to the baseline

    More complex models often cost more time & money to develop, deploy, and maintain - more dangerous for most companies.

Deep learning and traditional methods

- Neural networks are beautiful
  - Work very well with images and natural language text
  - But usually do not work well with transactional/tabular data - around **80%+** of the data most companies have

- **Random forest, gradient boosted decision trees, SVM, logistic regression** algorithms are as important as deep neural networks

Deep learning and traditional methods

- Don't be shy to choose the one that works best for you, even though it might not sound sexy
- We should balance students' training time for both deep learning and traditional methods

Just a tiny fraction of AI engineers & data scientists will work for Google's self-driving car team. Most of us will not.

Insights, insights, insights

- If you are working with transactional/tabular data, at least 50% of your time should be used to find **insights** about the problem and do feature engineering
    - Not just deep learning has limitations
    - Need to understand the problem to have better features and models
    - Better insights can lead to better NN models too
    - ML models should not be 100% black boxes (why? model bias, operational debt, legal requirement...)

    Students should seriously practice finding insights

Hyperparameter tuning

- Yes, it is important
- But often takes less than 5% of your “brain" time
- Grid search is useful. More advanced methods (such as Bayesian optimization) are even more useful
- Don't spend too much time for it

- What about AutoML?
  - Anyone can build a model. Tools are available everywhere. Data scientists and AI engineers should retire?
  - Remember the previous slides!

Bias vs. Variance

- Basic but very important
- Never overlook

- Check your code, check again and again
- 80-20 train-test split is not enough. Should always have a completely independent test set.
    - For time series data, should have an out-of-time test set

Big data

- Big data processing mindset, skills, and experience are extremely important
  - Toy project's data is often small. Real-life project's data is often very big
  - Find gold in the forest quickly before starving. Speed of code, speed of experimental iterations, speed of improving the models
  - SQL skills, Spark skills, and fundamental database knowledge are very useful

Most Vietnamese students are not good at big data processing.
We should adjust our AI curriculum.

Big data

- But it is not easy to find big data sets to practice, right?
    - No, every data set will be big enough if you put enough constraints on the available resources (set limits for RAM, CPU, processing time,...)
    - The Misfit example

- I cannot practice Spark because I don't have a big data cluster?
    - Set up a cluster of virtual machines on a PC/laptop
    - Spark can be set up easily on a PC/laptop, too

Team work

- Weakness of most Vietnamese students
  - Some students think they are geniuses (but that is not the truth) and don't try to work well with others
  - Even if someone is a genius, he/she cannot do everything
  - Universities should have many team projects, with clear reports about individuals' contributions. Advisors should always remind students to respect their peers.

Purpose

- Ask your students! Why do you want to study & work on data science / ML / AI?

Purpose

- Ask your students! Why do you want to study & work on data science / ML / AI?
  - Good income, but usually you will not become a millionaire
  - However, you will be able to do something meaningful and big

- Wrong expectations lead to low productivity and high turnover rate
  - A waste of time and money for both employers and employees
  - If a student just wants to be rich, he should immediately quit

Is the gap easy to be closed? With your help, YES

Thanks!

Questions?