Getting Started and Troubleshooting

How to Get Started – First Steps

Tips for getting started:

- Complete onboarding (e.g., training)

- Take inventory (tools, code, data, environment)

- Understand the big picture. What is the goal?

- Look for intermediate milestones:
   > Forming training set / valuation set
   > Decide on metrics
   > Coding data pipelines

- Thinking about functionality

  > What needs to happen? (build lists, embed objects, build classifier, …)

  > Are there modules for the functionality?
    Can check for built-ins, papers w code, Hugging Face, ...
    What do you need to code / build?

- Break out large tasks and add to roadmap

- Look for techniques (NLP, time series, ...)
  The type of data will help guide

  Helpful approaches:
  > Literature review
    Has this problem / similar problem been solved? How?

Troubleshooting

Let's discuss previous issues and others you're facing

Ideas are proposed. Please share what's working / not working.

Unclear Problem Definition

Each project should have a problem statement, but clarity may vary.

Q: What if the sponsor doesn't clearly define the problem?

You might propose ideas at intersection of what's achievable & interesting

Lack of Data

Each project should have data, but coverage and quality may vary

Q: What if the data won't support the objective?

You might propose ideas on what is achievable
- Solve for subset of population?
- Pivot the question?
- If labeled dataset is small, you might train on labeled data, predict on unlabeled, and review cases with high/low probability.
- Can more data be collected? Can supplemental data be created?

Each sponsor should meet w students as needed.

Q: What if the meetings aren't productive enough?

You might try:
- Forming an agenda / agreeing on it beforehand
- Sharing results beforehand to discuss during meeting
- Decide Next Steps in the meeting
- Call out decision points in meeting
- Ask for feedback

Upskilling

Ideally, the project brings together skills you've learned in the program.

Q: What if I don't have the skills to solve the problem?

> In some cases, you will need to learn new things
> Your sponsor / mentor might be able to suggest resources

Upskilling

Ideally, the project brings together skills you've learned in the program.

Q: What if I don't have the skills to solve the problem?

> In some cases, you will need to learn new things
> Your sponsor / mentor might be able to suggest resources

> Keep a list of open questions, skills needed
   Break these down into manageable tasks (e.g., learn n-gram models)
> You might distribute the work across teammates and share learnings

Sparse Feedback

Your mentor and sponsor should provide feedback on your work

Q: What if feedback isn't arriving often enough?

You can try:

> Checking if there is a better medium for feedback (email, meetings, .

> Asking if there is a better way to share results
   - Are questions clear enough?
   - Are things summarized?
    (e.g., assemble separate questions into single slide)

Sparse Feedback

Your mentor and sponsor should provide feedback on your work

Q: What if feedback isn't arriving often enough?

You can try:

> Understanding if there are busier/quieter periods

> Working on aspects not requiring the feedback

Model Doesn't Generalize

You've built a model on training data
The performance on new data is worse

Q: Why does my model perform much worse on new data?

This is usually a symptom of overfitting. Can consider:
- k-fold cross validation (for ML models)
- Dropout layers (for deep neural networks)
- More suitable data (larger dataset, more recent data)
- Looking at the distribution of training data, new data.
  Is there drift in distribution? If so, training set isn't sufficient.

Unstable Model

You've built a model
When you change hyperparameters slightly, the results change drastically

Q: Why is my model so unstable?

While we tune models for better performance,
a drastic change in output is not desirable

> The model may be too complex
Can try simpler model and examine sensitivity

> Did you cross validate the model? This is a good practice when feasible

Final Thoughts

You have a support system to help you through capstones including:

- teammates
- sponsor
- faculty mentor
- Program Director (Adam Tashman)
- Program Manager (Kylen Baskerville)

Please lean on this system for help!