Project

Last modified: 2023-03-14 17:04

Goals

Throughout this class, you've learned about the foundational methods for probabilistic modeling. The goal of this project is to give you hands-on experience in applying such models to real data.

You'll complete 3 tasks:

- 1) Select a specific data analysis task and suitable baseline model
- 2) Design and implement a promising upgrade to this model
    - Should be *hypothesis driven*: Why is your upgrade promising?
- 3) Evaluate your proposed upgrade via experiments on real data
    - Analyze whether your hypothesis was correct, or not.

In the real world, you'll often iterate between steps 2 and 3 multiple times to get better performance.

Teamwork

You should work in teams of 2. We may make exceptions for you to work alone or in teams of 3, but we do not recommend it. Teams of 4 or more are not allowed.

If you work alone, you will be required to do the same amount of work. If you work in a team of 3, you will be required to do 2 upgrades instead of 1 to appropriately scale the amount of work.

Task and Dataset selection

You should select a dataset that meets these requirements:

- Publicly available
- At least 50 instances
- At least 2 features
- Suitable for your chosen analysis task (regression, classification, clustering, etc.)
- Already available in format for ML analysis (we want you to focus on model building, not data cleaning)

These are **suggested** but not set in stone: if you have an idea you're excited about (e.g. you want to use a private dataset you're already involved with for ongoing research), please discuss with your instructor.

You should be confident that your dataset is appropriate for your desired analysis task.

Baseline method selection

Your "baseline" method is a concrete specification of the following 3 components:

- probabilistic model (likelihood, possibly also a prior)
- optimization problem (e.g. ML estimation or MAP estimation or posterior estimation)
- algorithm (e.g. gradient descent, coordinate descent, or closed-form formula)

Suggestions:

- MAP estimation for Dir-Cat model for unigrams, as in CP1
- MAP estimation for linear regression, as in CP2
- MAP estimation for logistic regression
- Random Walk MCMC for linear regression, as in CP3
- MAP estimation for Gaussian mixtures, as in CP4

If you wish to pursue a model that we do not cover in this course, please discuss with your instructor. The main requirement will be that it needs to be viewable as a *probabilistic model*.

Upgrade method selection

The emphasis of this project will be on creating and testing hypotheses about how to improve model performance. Toward this end, we want you to design, implement, and evaluate one specific, feasible upgrade for your baseline model.

Your proposed upgrade should:

- have a compelling story about why it *could* deliver improvement on your task / dataset
- have a possible performance gain that is concretely measurable
- be completable by you and your team within about 2 weeks of effort
- fit within the concepts of this course
- pursue an idea we have *not already covered* in-depth in homeworks or coding practicals.

Your upgrade does not need to be a *novel* idea (e.g. it can be described already in a textbook or research paper). Your upgrade does not need to succeed (we care more about understanding why it works or does not work than on what your final performance metric is).

For a list of possible ideas, see our Project Brainstorming Google Doc.

There are two major kinds of upgrades, enumerated below.

Upgrade Option 1: Changing the Model (Prior or Likelihood)

For this kind of upgrade, you'd change the concrete PDF/PMF of your prior or likelihood

Examples:

- Change a Normal prior on regression weights to a Laplace prior (corresponds to LASSO regression)
- Change a Normal likelihood for regression to a likelihood for robust regression less sensitive to outliers

Upgrade Option 2: Changing the Estimation Objective and/or Algorithm

Here, we're considering changes to how parameters (or distributions over parameters) are estimated.

Examples:

- Compare a baseline first-order gradient descent for MAP estimation of logistic regression with an upgrade that uses second-order gradient descent.
- Compare the baseline Monte Carlo estimate of logistic regression posterior predictive with an upgrade that uses a closed-form probit approximation.

We especially suggest considering changes to algorithms that enable **scaling** to larger datasets.

Examples:

- Compare the baseline gradient descent for logistic regression to stochastic gradient descent with minibatches
- Compare the baseline kmeans algorithm for clustering to one that is parallelized so that multiple workers each handle a subset of data.

Examples that are *not allowed*:

- Changing from an MLE estimate of the parameters to a MAP estimate is technically a change in the estimation objective, but we will generally not allow this one since we have studied this extensively in CP1 and CP2. We will only consider if the idea is particularly interesting/novel/impactful for your target task.
- Very simple, straightforward code optimzations (for example changing a for loop to a matrix operation) do not count. Your proposed upgrade needs take about a week or two of effort, not a change to a few lines of code.

Other Options

You may wish to pursue an upgrade that doesn't fall neatly into the options enumerated above. If so, please discuss with course staff!

Forming a Hypothesis

We want you to practice forming and testing concrete hypotheses about your modeling task.

Your hypothesis for why you are pursuing your upgrade should have the following structure:

> *We hypothesize that compared to [BASELINE], [UPGRADE] should improve [PERFORMANCE METRIC] on our dataset, especially when [SPECIFIC SCENARIO], because of [PROPERTY OF UPGRADE]*

For example, in CP1, we could have said the following

> *We hypothesize that compared to [the ML estimator for unigrams], [the MAP estimator for unigrams] should improve [heldout log likelihood] on our dataset, especially when [datasets are smaller than a few thousand words], because [MAP offers a smoother way to handle unseen words, while ML is known to overfit and concentrate mass on too few words].*

This clear statement of your hypothesis illuminates exactly what tests we need to run to evaluate. Remember, in science, all good hypotheses are testable and thus **falsifiable**.

Deliverables

- (1) by Mon Mar 27: Complete Team Formation Form on Gradescope
    - Commits you to your team
    - Suggests a chosen task, dataset, and baseline
    - As needed, your data and methods (esp. your upgrade) can still change down the road
- (2) by Tue Apr 4: Turn in Initial Report
    - 1-2 page report
    - Summarize dataset and task.
    - Describe baseline method (model and estimation).
    - Suggest one possible upgrade (very briefly).
- (3) between Apr 10 and May 4: Meet with course staff
    - Discuss your planned upgrade at a whiteboard for 10 min
- (4) by Thu May 11: Turn in Final Report
    - 3-4 page report
    - Describe baseline and upgrade clearly
    - 1 figure/table (or more) evaluating your hypothesis

Grading

- 84% of project grade depends on final report
- 8% of project grade is initial report
- 8% of project grade is participation in the meet with course staff

---