MTH 511a - 2020: Lecture 25

Instructor: Dootika Vats

1 A recap of data analysis techniques so far

After this lecture, we shift our focus to computation for Bayesian models, so before we go along there, this lecture is a quick recap of all the techniques which we've learned. We've covered three areas of topics to help us with analyzing data:

1. Estimation (optimization)
2. Model selection (cross-validation)
3. Model inference (bootstrapping)

Optimization in statistics is predominantly motivated from maximizing likelihoods or penalized likelihood functions. In the next month, we will also look into maximizing *posterior* distributions (but more about that later).

When obtaining closed-form expressions of parameters that maximizes the (penalized) likelihood is challenging, numerical optimization techniques are used.

Step 0: Before you venture into any optimization method, it is crucial to *try* to check whether the objective function is concave. If it is concave, then you a global maxima exists. If not concave, you know local maximas will exists.

- Non-stochastic Gradient-Based methods
  - Newton-Raphson
    - Requires explicitly calculating the Hessian and inverting the Hessian.
    - Can be expensive in large dimensions

    - For concave objective functions, converges the fastest and requires little to no tuning.
    - For non-concave objective functions, it may converge to a local maxima, converge to a local minima, or diverge.
    - Examples - logistic regression, probit regression (exam)
  - Gradient-Ascent
    - Requires only the gradient vector.
    - Can be expensive to calculate for large dataset
    - It may converge to a local maxima, or oscillate around a local maxima.
    - In case of oscillations, tuning the stepsize t to t_k is useful.
    - Examples - logistic and probit regression (with careful tuning), Location Cauchy example (with good starting values)
- Non-stochastic non-gradient-based algorithms
  - MM algorithm
    - Useful when objective function can easily be lower bounded by a concave function locally.
    - Guaranteed to converge to a local maxima
    - Examples - Bridge regression
  - EM algorithm
    - A specific case of the EM algorithm
    - Particularly useful for censored/missing data problems.
    - Immensely useful in classification particularly using Gaussian mixture models.
- Stochastic optimization methods
  - Stochastic gradient ascent
    - Useful when original gradient ascent algorithm is expensive to implement
    - or when data is coming in sequentially (like in an online format)
    - Typically a good idea to use mini-batch stochastic gradient for stable answers.
    - Convergence properties are the same as gradient ascent
  - Simulated annealing

    - ∗ No gradient are required.
    - ∗ Most useful for non-concave target distributions since it can escape out of local maximas.

The above methods provide estimators based on certain fix model tuning parameters. This includes penalty term λ, bridge regression penalty term α, and the number of clusters/classes in Gaussian mixture models, C.

In order to choose the these tuning parameters, we use model selection techniques:

- Cross-validation
  - Leave-one-out Cross-validation: expensive, but accurate
  - K-fold cross-validation: cheaper, but not as accurate.
  - effective in regression models
- Akaike Information Criterion (AIC) / Bayesian information criterion (BIC)
  - AIC/BIC is most effective for finding C in Gaussian mixture models.
  - Can also be used in regression problem
  - BIC is more robust than AIC

Note: None of these techniques can be used as black-box. For every implementation you must decide looking at the final prediction error/AIC/BIC values, which value of the tuning parameter may be more reasonable.

Once final estimates using model selection and optimization tools have been obtained, interest is in constructing confidence intervals around parameters of interest. For this we can use two Boostrapping methods:

- Nonparametric Bootstrap
  - Cheaper to implement and can always be implemented
  - Produces higher variance in resulting confidence intervals for low sample size settings
- Parametric Bootstrap
  - Can't always be implemented
  - When it can, it typically works better than nonparametric methods
  - It is more expensive to implement.