HSD5230 High Performance Computing Homework 10

Miao Cai[*]

2019-03-24

1 Continuous Optimization

1) What does the first order derivative tell us about a function?
The first order derivative tells us how fast the function increases or decreases (gradient/first derivative), also known as the instantaneous slope of a point or the rate of change of a function.

2) What does the second order derivative tell us about a function?
The curvature of a function. It tells us how fast the gradient/first derivative increases or decreases.

3) What is the Jacobian matrix?
The Jacobian is the first order partial derivative matrix (the dimension depends on the number of Xs).

4) What is the Hessian matrix?
The Hessian matrix is the second order partial derivative matrix (the dimension depends on the number of Xs).

5) Does iteratively reweighted least squares (IRLS) use only first, or first and second order derivatives?
It uses both the first order and second order derivatives.

6) Why does IRLS tend to converge faster than gradient descent for a GLM?
Because the IRLS uses the Hessian matrix to select an adaptive step size, whereas the gradient descent algorithm need to compute the inverse of the Hessian matrix. The iterative method (IRLS) generally requires fewer iterations to converge than the gradient descent method.

7) What is the difference between gradient descent and stochastic gradient descent?
The stochastic gradient descent calculates the gradient for the i-th sample, not the full sample. In comparison, the gradient descent calculates the gradient at all the points in the sample. Therefore, gradient descent is more accurate but requires more resources while the stochastic gradient descent is less accurate but costs cheaper.

8) Is stochastic gradient descent more efficient than gradient descent in general? Why or why not?
Yes. It calculates the gradient for the i-th sample, not the full sample, so it is less accurate but costs cheaper.

9) What is the difference between coordinate descent and gradient descent?
The coordinate descent change only a single predictor while keeping other variables the same at a time, and it turns out to be the optimal solver for regularized GLMs.

10) Why does IRLS scale poorly with increasing number of predictors?
Because it has a complexity of `O(n**2)`. When the number of predictors increases, the resource it takes will increase exponentially.

11) If I have a large number of predictors (1000's) and want to distribute my calculations using parallel frameworks like H2O or Dask or Apache Spark, should I use IRLS or L-BFGS? Why?
We should use L-BFGS since it does not fully calculate the Hessian matrix (a quasi-Newton methods). Instead, it approximates the Hessian matrix by previous gradient evaluations, which makes it vertically scalable in terms of the number of predictors.

[*] Department of Epidemiology and Biostatistics, College for Public Health and Social Justice, Saint Louis University. Email miao.cai@slu.edu

2 GLMs

12) How does a general linear model differ from a generalized linear model?
A general linear model requires that the conditional mean of the outcome variable should be normally distributed with the variance σ^2, whereas a generalized linear model does not require the conditional mean of the outcome variable to be normally distributed.

13) What are the three components of a generalized linear model?

- the distribution of the outcome variable,
- the linear predictors,
- the link function.

14) What optimization method is typically used for finding solutions to generalized linear models?
Maximium likelihood methods and implemented by iteratively reweighted least squares.

15) What is maximum likelihood and what does it have to do with GLMs?
What are the parameter values that could most likely give us this data.

16) What is the canonical (natural) link for the binomial distribution?
The logit function ($\log \frac{\mu}{1-\mu}$).

17) What is the canonical (natural) link for the Poisson distribution?
The natural log function.

18) What is the canonical (natural) link for the Normal distribution?
The identical distribution.

19) Why is a binomial family with a logit link (logistic regression) sometimes easier to fit than a binomial family with a log link (log-binomial regression)?
Because the range of logit p is from $(-\infty, +\infty)$ while the range of $\log p$ is $(-\infty, 0)$ which is not identical with the range of a normal distribution.