Machine Learning 2
Exercise Sheet 9

Prof. Dr. Dr. Lars Schmidt-Thieme, Brad Baker
Information Systems and Machine Learning Lab
University of Hildesheim

June 27th, 2017

Submission until July 4th, 8:00 AM by learnweb.
Please put your name in all filenames.
Non-pdf submissions for non-programming exercises will not be graded.

Exercise 17: Proximal Gradient Descent (10 points)

a) (3 points) The proximal operator is only useful when applied to a certain class of problem which satisfies certain conditions. Identify these conditions, and provide examples for a problem which

i) satisfies these conditions and

ii) fails these conditions.

b) (7 points) "Elastic Net" regularization is a regularization technique which linearly combines the L1 and L2 regularization terms from LASSO and Ridge regression. The idea is to attempt to overcome some of the limitations of LASSO regression by incorporating an L2 regularized term while still using an L1 term to generate a sparse model.

For a squared-loss optimization, the Elastic Net regularized problem is given as

$$\text{arg min}_\beta ||\mathbf{y} - \mathbf{X}\theta||^2 + \lambda_2||\theta||^2 + \lambda_1||\theta||.$$

- Choose a sensible decomposition for this problem in terms of a convex function h and a differentiable function g. Make sure to show how the terms in your decomposition satisfy these criterion.
- For the differentiable function g compute the gradient ∇g, and write the gradient update step at iteration t.
- Write down the proximal operator on the gradient update step for iteration t, given as

$$prox_{\alpha^{(t)}h}(\theta^{(t)} - \alpha^{(t)}\nabla g(\theta^{(t)}))$$

- Does this proximal operator have an analytic solution? If so, compute it. If not, try another choice of h and g.
- *HINT* Your initial choice of g and h will make the analytic solution of the prox operator more or less difficult. Try to think of problems where the operator already has an easy analytic solution and try to group your decomposition to favor this solution.

Exercise 18: Iterative Shrinkage and Thresholding

a) (2 points) The Barzilai Borwein step-size is computed in order to provide a best-approximation for part of the Proximal Gradient application to Regularized Loss minimization.

- Identify where this approximation appears in the Proximal Gradient approach to Regularized Loss minimization.
- What is the most obvious limitation of using the Barzilai Borwein step-size for optimization?

b) 6 points) Suppose you are given the following predictors X with ground-truth y.

$$X = \begin{pmatrix} 3 & 4 \\ 2 & 6 \\ 7 & 5 \\ 8 & 1 \end{pmatrix}, \quad y = \begin{pmatrix} 1.1 \\ 1.4 \\ 1.7 \\ 1.0 \end{pmatrix}$$

- Initialize your parameters, learning rate, residual, and regularizers in order to perform the Iterative Shrinkage-Thresholding Algorithm (ISTA) given on slide 23 of the lecture.
- Perform two full outer-iterations of ISTA, updating the regularizer after each inner-loop, and reporting the update in the residual. For the inner-loop only perform one iteration for simplicity's sake.

c) (2 points) Nesterov's Accelerated Generalized Gradient Descent (NAGGD) provides a version of Proximal Gradient Descent which converges more quickly than regular Proximal Gradient Descent due to the inclusion of a momentum term. Namely, the parameter update rule for NAGGD is given as

$$\theta^{(t+1)} := prox_{\frac{1}{2\alpha^{(t)}}h}(x^{(t)} + \frac{t-2}{t-1}(\theta^{(t)} - \theta^{(t-1)}) - \alpha^{(t)}\nabla g(x^{(t)}))$$

Using your solution to exercise 17.b given above, provide the parameter update rule for Elastic-Net-regularized NAGGD. You should not need to recompute the solution to the prox-operator.

Bonus: Implementing Proximal Gradient Methods (10 points)

For this problem, you will again work with the IRIS dataset. Your task is to perform an Elastic-Net regularized, multi-class logistic regression in order to predict the species of an Iris given information about the Petal length and width and Sepal length and width.

For problems b) and c), make sure you implement these algorithms yourself, and do not use libraries for ISTA. You can use libraries for the Barzilai Borwein step-update.

a) (1 points) Write down the Elastic-Net regularization problem for multi-class logistic regression. (keep in mind that this MULTI-CLASS classification)

b) (4 points) For the problem from a), implement ISTA using algorithm 13.2 given on slide 23 of the lecture.

c) (2 points) For the problem from a), implement Fast ISTA, which uses the momentum term from NAGGD.

d) (3 points) Plot and compare the convergence for each of these two ISTA implementations. Explain why the convergence rates differ in the way they do.