STAT 535 Homework 5
Out November 7, 2023
Due November 16, 2023
©Marina Meilă
mmp@stat.washington.edu

Reminder: you are allowed and even encouraged to use results from previous homeworks, course notes, lectures withouth proof.

Problem 1 – Descent algorithms for training a neural network
This problem asks you to train a neural network to classify the data sets given on the Assignments web page. The inputs are $d = 2$-dimensional, outputs are ± 1, one data point/line. *Submit the code for this problem.*

Objective to minimize is $\hat{L}_{logit}(\beta, W) = -\frac{1}{n}$log-likelihood$(\mathcal{D}|\beta, W)$ and $\beta \in \mathbb{R}^{m+1}, W \in \mathbb{R}^{(d+1)\times m}$ are the neural net parameters.

Algorithms: steepest descent with fixed step size. You need to implement the algorithm yourself. [Optional, for extra credit: implement Newton, or run Newton, LBFGS quasi Newton from library code.]

Dataset $\mathcal{D}$ given `hw5-nn-train-100.dat`

a. Plot the data set in $\mathbb{R}^2$, representing each class with a different color or symbol.

b. Based on the plot in **a.**, is it possible to get $\hat{L}_{01} = 0$ for $m = 2$? Explain.

c. Choose a number $m \geq 3$ hidden units and train the neural network on the $\mathcal{D}$. Obtain the best empirical $\hat{L}_{logit}$ you can. *Note that larger m values, i.e $m \geq 10$ may be easier to train.*

Explain how you chose the initial points. It's ok to plot the data and look at it or even to make a sketch of the solution you want to find. *If you implement more than one algorithm, start them all from the same initial point.*

The training algorithm will converge to a local optimum. It's OK to look at this local optimum and try other initial points if the found optimum is bad. (Don't forget to use the same initial point for all algos in the results you present in the homework.) It's also recommended to challenge the algorithm by giving it random/uninformative initial points. *Do not start all the parameters at 0 [Why?].*

Chose the stopping criterion $1 - \frac{\hat{L}^{k+1}}{\hat{L}^k} \leq tol$ with $tol = 10^{-4}$. If this tolerance cannot be reached in a reasonable number of steps, set a higher tol and report that value.

d. Describe briefly the implementation details of your algorithms. Size of the fixed step, number of iterations (and if it converged or not) and final value of loss functions $\hat{L}_{logit}$ and $\hat{L}_{01}$. Record also the time each algorithm takes and report it.

[Optional: If you used other algorithms report on those too. If you used line search, report if

you bracketed the min or not in line search, what line search method you used (*you can use code from other sources to bracket the mininmum, and you can implement another line search method than Armijo.)*]

e. Estimate the value of L_{logit}, L_{01} by averaging them on the test set `hw5-nn-test.dat` for the final classifier obtained. Optionally, compute these values at each iteration and plot them in the graphs for **f.**.

f. Plot the values of $\hat{L}_{logit}$, $\hat{L}_{01}$ and the respective costs L_{logit}, L_{01} on the test set vs. the iteration number k. Make two separate plots for the two costs. If you have computed the test set costs at each iteration, plot these too on the respective graphs.

g. Plot the final decision region superimposed on the data.

[**h. Optional but encouraged**] Plot (some of) the β parameters vs k; on a separate plot, show trajectories of β parameters coming from different initializations.

Please make clear, well-scaled, well labeled graphs.

Problem 2 – Regularization is monotonic w.r.t. λ
Let $J_\lambda(w) = \hat{L}(w) + \frac{\lambda}{2}||w||^2$ be a regularized objective functions, where w are the parameters. For example, the linear ridge regression from Problem 3. Let $\lambda_1 > \lambda_2 > 0$ and denote $w_{1,2} = \text{argmin}_w J_{\lambda_{1,2}}$ the optimal solutions for λ_1, respectively λ_2, with $w_1 \neq w_2$, and assume further that $J_{\lambda_{1,2}}$ have *unique global minima.*

a. Prove that $||w_1|| < ||w_2||$ whenever $w_{1,2} \neq 0$.

b. Prove also that $\hat{L}(w_1) > \hat{L}(w_2)$.

In other words, imposing more regularization reduces the regularized quantity $||w||$, and increases the un-regularized one (i.e., the loss).

Problem 3 – Ridge regression
In this problem you will perform ridge regression on the function $f^*(x) = 0.1x^2 + x + 1$ on $[0, 1]$. In the file `hw5_rr.dat` you will find a set of n (x^i, y^i) values with $y^i = f^*(x^i)$.

a Let $f(x) = \beta_0 + \beta_1 x$ be the predictor of y; β_0, β_1 will be estimated by Ridge Regression with regularization parameter λ. Denote $\beta_{0,1}(\lambda)$ the result of this estimation. Let the data matrix be the row vector $X = [x^1 \ \ldots \ x^n]$, and define the column vector $y = [y^1 \ \ldots \ y^n]^T$

Write the expressions of $\beta_0(\lambda)$, $\beta_1(\lambda)$ as functions of X, y, λ.

b Now choose a set of λ values including 0 and n. Calculate $\beta_{0,1}(\lambda)$, $\hat{L}_{LS}(\lambda)$ and $J(\lambda)$. Plot on the same graph $\beta_{0,1}(\lambda)$ vs λ.

c Plot on the same graph $\hat{L}_{LS}(\lambda)$ and $J(\lambda)$ vs λ. Comment on what you observe in the graphs of b, c.

Problem 4 – Online linear regression by Stochastic gradient

Consider the linear regression problem with Least Square loss

$$\min_{\beta} E[(y - \beta^T x)^2] \; = \; \min_{\beta} \; L_{LS} \tag{1}$$

where $y \in \mathbb{R}$, $x \in \mathbb{R}^n$, $\beta \in \mathbb{R}^n$. For simplicity we consider the infinite sample version of the problem, but if you want a variation (ungraded) try also the finite sample version, where we optimize $\hat{L}_{LS}$ instead.

The function in (1) is a quadratic function that has a closed form solution, but we will pretend that we don't know this and investigate the use of (stochastic) gradient descent for this problem.

a. Find the expression of the gradient and Hessian of this problem, i.e $\nabla L_{LS}(\beta)$, $\nabla^2 L_{LS}(\beta)$. Express the Hessian as a function of some well known statistical descriptor(s) of the data distribution.

b. Assume that the covariates x are sampled from a Normal distribution with mean 0 and non-singular covariance Σ (known). Describe and motivate a reasonable way to find the λ parameter of the Stochastic Gradient algorithm based on this assumption.

c. Write the expression of $d = \frac{\partial L_{LS}(y,\beta^T x)}{\partial \beta}$. Show that the direction of descent d is along x, i.e. $d = \alpha x$ for some scalar α, not necessarily positive. What does the scaling of x represent from a statistical modeling point of view?

e. Write in pseudocode the Stochastic Gradient Descent algorithm to optimize this problem. Assume that λ is known.

For practice, ungraded Repeat the problem with an added regularization term $\frac{C}{2}||\beta||^2$.