Machine Learning

2º Lab Assignment

Shift: Thursday 17h | Group Number: 21

Number: 80966 | Name: Francisco Azevedo

Number: 81356 | Name: Duarte Dias

2 – The gradient descent method

2.1.1

η	a = 0.5	a = 1	a = 2	a = 5
.001	Div	Div	Div	990
.01	760	414	223	97
.03	252	137	73	31
.1	75	40	21	8
.3	24	12	5	8
1	6	1	Div	Div
3	6	Div	Div	Div
Fastest	η= 1 or 3	η = 1	η = 5	η = .1 or .3
Divergence threslhold	η >0.001	3 > η >0.001	1 > η >0.001	η < 1

2.1.2

We conclude that the parameters that grant the fastest convergence are the eta = 0.1 and a = 1. For values of η too high or too low the gradient generally does not converge. We find that for η = .001 the method does not converge or is slow to converge. For values of η=3 the method only converges for a = 0.5 (which is when the parabola is widest).

The learning rate is a tradeoff between the velocity of convergence and the possibility of the algorithm converging.

2.1.3

For a = 0.5 the fastest optimization corresponds to 6 steps when η = 1 or 3.
For a = 1 the fastest optimization corresponds to 1 step when η = 1.
For a = 2 the fastest optimization corresponds to 5 steps when η = 0.3.
For a = 5 the fastest optimization corresponds to 8 steps when η = 0.3.

If f(x) is differentiable in all its domain then it is continuous in its domain. Given that f(x) is continuous and coercive it has an optimal solution (global minimum).
The value of η that optimizes a differentiable function of one variable and for each given starting point in that number of steps is when η = 1/a.

2.2.1

η	a = 2	a = 20
.01	448	563
.03	148	196
.1	43	Div.
.3	13	Div.
1	Div.	Div
3	Div.	Div.
Fastest	13	196
Divergence threshold	η = 1	η = 0.1

2.2.2

The function does not converge for learning rates η with high values. For a = 2 the function stops converging for η >= 1. For a = 20 the function isn't able to converge for η>= 0.1. This means that this method.

2.2.3

The method does not converge in one step every time because the direction of convergence is perpendicular to the contours which is not necessarily the direction of the minimum. Therefore, the method does not always converge in one step.

A narrower valley implies elliptical curves and not concentric curves. Concentric curves grant that convergence can be done in one iteration given that the direction of the minimum is the same as the direction of convergence (perpendicular to the contours).

Therefore, a narrower valley is associated to higher number of steps needed in order to converge to the desired function minimum.

3 – Momentum term

3.1

η	α = 0	α = 0.5	α = 0.7	α = 0.9	α = 0.95
.003	Div	Div	Div	Div	Div
.01	563	558	552	516	448
.03	186	181	174	115	165
.1	Div	48	35	91	122
.3	Div	Div	29	83	92
1	Div	Div	Div	92	146
3	Div	Div	Div	Div	147
10	Div	Div	Div	Div	Div
Div. Threshold	η= 0.1	η = 0.3	η = 1	η = 3	η = 10

3.2

The momentum term, in general, decreases the number of steps needed for the function to converge.

The alfa that grants the fastest convergence is α = 0.7 and η = 0.3.

The function will converge for a bigger range of different η when the momentum term (alfa) is closer to 1 in value. This is due to maintaining higher momentum from past iterations providing the possibility to converge in a bigger range of η. In case α = 0.95 the convergence is slower than α = 0.7 but still faster than α= 0.

4 – Adaptive step sizes

4.1

# of tests	alpha	η ->	-20%	-10%	best	+10%	+20%
143	0.95	# of iterations ->	153	159	137	147	146

Best at (eta = 0.06, alpha = 0.9))

4.2

In order to find the best alfa and η association that results in the smallest number of steps we need to test for many different pairs of values. This was done using two nested for cycles running different numbers for each variable alfa and η. The complexity is O(n*m) which for high number of alfas (n) and η (m) to be tested creates a computational demanding problem.

4.3

η	a = 0	a = .5	a = .7	a = .9	a = .95	a = .99
.001	596	298	236	140	198	167
.01	565	287	221	190	200	165
.1	769	389	214	183	172	152
1	729	396	233	160	137	173
10	672	383	239	173	124	133

4.4

	# of tests	η	alfa	# of iterations
Without adaptive step sizes	90	-10%	0.9	722
		best		637 (η = 0.0132)
		+10%		>1000
With adaptive step sizes	90	-10%	0.99	312
		best		202 (η = 0.09)
		+10%		329

4.5

Both the techniques improve the convergence rate of the functions. The momentum term technique improves the convergence rate especially when the cost function J(Θ) has a deep valley in which the gradient methods grants slow results.

The adaptive step sizes assumes that the step size η is different for each component of Θ and change each iteration. This technique performs better if the cost function contains valleys aligned with the x axes.