Instituto Superior Técnico

Aprendizgem Automática

Laboratório 2 - Gradiente Descendente

Autores:
Diogo Moura - nº 86976
Diogo Alves- nº 86980

Turno:
Terça feira 17h-18h30m

13 de Outubro de 2019

2.1 Minimização de Funções de uma Variável

1)

Tabela 1: Tabela com o número de passos do algoritmo em função dos parâmetros a e η

η	$a = 0.5$	$a = 1$	$a = 2$	$a = 5$
0.001	> 1000	> 1000	> 1000	990
0.01	760	414	223	97
0.03	252	137	73	31
0.1	75	40	21	8
0.3	24	12	5	8
1	6	1	> 1000	div
3	6	div	div	div
Fastest	$\eta = 2$	$\eta = 1$	$\eta = 0.5$	$\eta = 0.2$
Divergence threshold	$\eta = 5$	$\eta = 2.5$	$\eta = 1.22$	$\eta = 0.5$

2)

$$f(x) = a\frac{x^2}{2} \tag{1}$$

$$x^{(n+1)} = x^{(n)} - \eta \times f'(x^{(n)}) = x^{(n)} - \eta \times a \times x^{(n)} = x^{(n)} \times (1 - \eta a) \tag{2}$$

A otimização mais rápida converge para o valor do mínimo, que é zero, logo:

$$x^{(n+1)} = 0 => \eta a = 1 => \eta = \frac{1}{a} \tag{3}$$

3)

$$x^{(n+1)} = x^{(n)} \times (1 - \eta a) \tag{4}$$

No limiar da convergência, tem-se (ver figura 1):

$$x^{(n+1)} = -x^{(n)} \tag{5}$$

Logo,

$$1 - \eta a = -1 <=> \eta = \frac{2}{a} \tag{6}$$

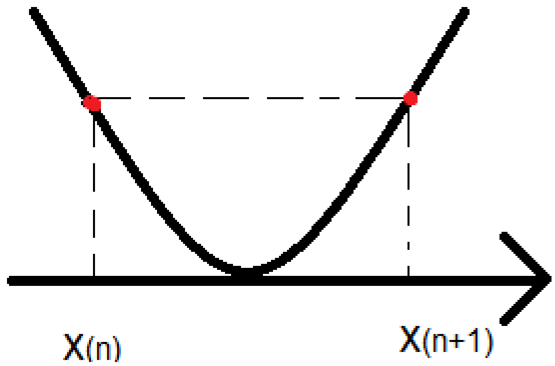

Figura 1: Esquema da situação de limiar de convergência

4)

Exite um valor de η (η_{ideal}) para o qual a convergência é mais rápida. Para valores de η menores que η_{ideal}, existe sempre convergência ($\eta_{ideal} = 1/a$). Existe também um valor de η a partir do qual deixa de existir convergência: $\eta_{threshold} = 2/a$, calculado teóricamente, no entanto, observando a tabela, $\eta_{threshold} \approx 2.5/a$. Isto porque para valores de η superiores mas próximos de $\eta_{threshold-teorico}$ a divergência não é detetada no número finito de iterações do programa.

5)

A optimização mais rápida é sempre feita em 1 iteração.

$$x^{(n+1)} = x^{(n)} - \eta \times f'(x^{(n)}) \tag{7}$$

Para se chegar ao mínimo em 1 iteração, tem que se verificar

$$x_{min} = x^{(1)} = x^{(0)} - \eta \times f'(x^{(0)}) \tag{8}$$

Assim, desde que $f'(x^{(0)})$ seja diferente de 0, existirá sempre um valor de η que irá fazer convergir em uma iteração:

$$\eta = \frac{x^{(0)} - x_{min}}{f'(x^{(0)})} \tag{9}$$

2.2 Minimização de Funções de mais que uma Variável

1)

Tabela 2: Tabela com o número de passos do algoritmo gradiente descente para uma função com vária variáveis em função dos parâmetros a e η

η	$a = 2$	$a = 20$
0.01	414	414
0.03	137	137
0.1	40	> 1000
0.3	12	div
1	> 1000	div
3	div	div
Fastest	$\eta = 0.6$	$\eta = 0.091$
Divergence threshold	$\eta = 1$	$\eta = 0.1$

2)

Quanto menor a largura do vale, menor é o número de iterações que é possivel alcançar, uma vez que as iterações têm sempre direção perpendicular às curvas de nível e se estas forem mais arredondadas, é mais rápida a convergência.

3)

Não, nem sempre é possivel para funções com mais de uma variável convergirem em uma iteração apenas. Neste caso, por exemplo, o número mínimo de iterações é 5 para $a = 2$ e 44 para $a = 20$.

3 Termo do Momento

1)

Tabela 3: Tabela com o número de passos do algoritmo com termo do momento em função dos parâmetros α e η

η	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 0.7$	$a = 0.9$	$\alpha = 0.95$
0.003	1383	1380	1375	1354	1320
0.01	414	411	406	382	338
0.03	137	134	129	96	171
0.1	> 1000	36	31	85	122
0.3	div	> 1000	31	67	148
1	div	div	div	74	146
3	div	div	div	div	172
10	div	div	div	div	div
Divergence threshold	$\eta = 0.1$	$\eta = 0.3$	$\eta = 0.6$	$\eta = 1.9$	$\eta = 3.9$

2)

A partir da tabela, verifica-se que para cada valor de η existe um valor de α que faz com que o número de iterações seja mínimo. Para valores de η pequenos, este valor de α é próximo de 1, e vai decrescendo à medida que η aumenta até cerca de 0.3 e depois volta a subir. O limiar de convergência ($\eta_{threshold}$) também aumenta à medida que α aumenta.

4 Tamanhos dos "steps"adaptativos

1)

Tabela 4: Tabela com o número de passos do algoritmo gradiente descente para uma função com vária variáveis e passos adaptativos em função dos parâmetros α e η

N. of tests	α	η	-20%	-10%	Best	$+10\%$	$+20\%$
6	0.95	N. of iterations $\rightarrow$	196	232	148	220	219

2)

É dificil encontrar valores de parâmetros que resultem em poucas iterações porque o número de iterações varia de forma imprevisivel à medida que se varia η, isto é, não existe um valor de η que seja claramente ideal e que ao afastarmos-nos desse valor obtenhamos sucessivamente um pior valor para o número de iterações.

3)

Tabela 5: Tabela com o número de passos do algoritmo gradiente descente para uma função com vária variáveis em função dos parâmetros α e η

η	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 0.7$	$a = 0.9$	$\alpha = 0.95$	$\alpha = 0.99$
0.001	401	215	171	101	160	158
0.01	384	201	168	165	145	139
0.1	575	306	159	149	138	144
1	522	305	169	135	132	123
10	479	292	190	146	113	108

4)

Tabela 6: Tabela com o número de passos do algoritmo gradiente descente para uma função com vária variáveis e passos adaptativos em função dos parâmetros α e η

	N. of tests	η	α	N. of iterations
Without adaptative step sizes	6	-10%	0.99	1090
		final $\eta = 0.01$		490
		$+10\%$		717
With adaptative step sizes	20	-10%	0.99	205
		final $\eta = 0.165$		97
		$+10\%$		228

5 Notas finais

O método do gradiente é utilizado para minimizar funções de custo de uma ou mais variáveis, sendo assim um método utilizado para problemas de otimização. Quando estudamos o método do gradiente descendente, notamos que os fatores mais importantes são a convergência e a sua rapidez (número de passos).
A convergência do método não está assegurada e está dependente do parâmetro η (tamanho dos passos adaptativos) escolhido e da própria função que se pretende estudar, pelo que para cada função podemos definir o "threshold"da divergência, que é o valor dos passos η a partir do qual a função diverge.
O método do gradiente descendente apresenta limitações, nomeadamente para funções com "vales"estreitos, em que a divergência se torna oscilatória, pelo que se acrescentarmos um termo do momento α, a convergência se torna mais

rápida, dado que o termo do momento atenua as oscilações ao usar, em cada iteração, uma fração, correspondente ao momento, da iteração anterior.
Assim, verificámos que o termo do momento provoca a redução do número de passos (ver tabela 3), sendo que existe um termo de momento ótimo para a rapidez da convergência. Por exemplo, para $\eta = 0.1$ sem termo do momento o método percorre mais de mil iterações até convergir, ao passo que com $\alpha = 0.5$ converge em 36 iterações. Também assegura a convergência do método do gradiente em alguns casos em que o mesmo não era convergente para um dado tamanho de passos adaptativos, como se pode verificar pelo aumento do "threshold"dos passos adaptativos para a divergência.
Ainda assim, o método ainda poderá apresentar algumas limitações, nomeadamente em funções de custo complexas. Outra optimização que utilizámos foi o uso de passos adaptativos (ver tabelas 4, 5 e 6).
Este trabalho foi bastante proveitoso no sentido de estudarmos o método do gradiente descente e na forma como o podemos otimizar. Compreendemos agora que para melhorarmos o método do gradiente descendente devemos usar um termo de momento de tamanho variável e que em funções complexas é vantajoso usar "steps"adaptativos para assegurar a convergência e de forma mais rápida.