Machine Learning for Signal Processing

[5LSL0]

Ruud van Sloun
Rik Vullings

Assignments Optimization and Regularization

May 3, 2019

Optimization and regularization

In this assignment you will investigate several optimization and regularization strategies for training nonlinear models.

Optimization

Consider the following model:

$$f(\mathbf{x}; \mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \mathbf{w}^{(2)}, b^{(2)}) = \left(\mathbf{w}^{(2)}\right)^T \max(0, \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) + b^{(2)}. \quad (.1)$$

In the previous assignment (Activation functions), you learned to derive the gradients of a cross-entropy cost function with respect to the model parameters by back-propagation using the chain rule. We will exploit these gradients to update the parameters through various optimization strategies:

- Stochastic gradient descent, SGD (default settings: $\mu = 0.0001$)
- SGD with momentum (default settings: $\mu = 0.0001, \rho = 0.9$)
- AdaGrad (default settings: $\mu = 0.0001$)
- RMSprop (default settings: $\mu = 0.0001, \rho = 0.999$)
- Adam (default settings: $\mu = 0.0001, \rho_1 = 0.9, \rho_2 = 0.999$)

Q1: Explain how and why the last four methods improve convergence w.r.t. plain SGD. What are their advantages and possible disadvantages?

Q2: Implement all of these solvers for the above model (default parameters), and evaluate the convergence of all parameters, as well as the cost as a function of the number of iterations. Plot these convergence graphs, include them in your report, and discuss the results. Also include a scatter plot of the input data points (x_1 vs x_2) and their classification. Note: use at least 50000 iterations. For implementation of the backpropagation, use the provided Python file "backprop.pyc".

The `backprop.pyc` file needs the following syntax:

```
from backprop import backprop
z_out, J, dJ_db1, dJ_dW1, dJ_db2, dJ_dw2 = backprop(Xbatch,ybatch,W1,b1,w2,b2)
```

where `z_out` is the probability that the data in `Xbatch` belongs to either class 0 (`z_out`<0.5) or class 1 (`z_out`≥ 0.5), `ybatch` are the class labels, `J` is the cost, and `dJ_da` is the gradient of `J` with respect to variable `a`.

Q3: Show what happens if you significantly increase or decrease the learning rates, and discuss your findings.

Regularization

Q4: Implement both ℓ_2 and ℓ_1 regularization strategies, and evaluate their impact on the convergence of the parameters using the Adam solver with $\lambda = 0.001, 0.01$ and 0.1. Plot the convergence graphs, include them in your report.

Q5: Explain your findings (compare to the convergence graphs of Q2), and explain in which cases either of these regularization strategies would be appropriate.

Q6: Explain "dropout" regularization, and explain why it would or would not have been an appropriate strategy for the model described above.