AARHUS UNIVERSITET, SCIENCE, COMPUTER SCIENCE

Machine Learning - Handin 1

Peter Burgaard - 201209175
Marie Louisa T. Berthelsen - 201303610
Nanna Engell Rønde Andersen - 201205671

September 24, 2016

In your report, you should shortly explain your algorithms and the choices you have made. You should include a plot of some of the digits that your classifier makes mistakes on – and discuss why that may be.

To show the performance of your gradient descent implementation include a plot of the cost function as a function of the number of steps taken. You should compare the running time and the convergence/output quality of mini-batch and full-batch gradient descent.

You should provide a figure that shows the parameter vectors for the case of classifying 2s versus 7s, and for the 10 all vs. one classifiers. Furthermore, your report should include results for the pairwise computations (at least two vs seven) as well as the results for the full classifier on the AU data set.

Furthermore you must answer the following theoretical questions (although the bonus question is optional).

- Sanity Check: If we randomly permute the pixels in each image and train the classifier, the classifier will be just as good, given we permutate all later given images the same way as the training data was permutated.

- Linear Separable: If the data is linearly separable, what happens to weights when we implement logistic regression with gradient descent? That is, how do the weights that minimize the negative log likelihood look like?

  Assume that we have full precision (that is, ignore floating point errors). We can run gradient descent on the data set for as long as we want (suppose God helps you). Now what will happen with the weights in the limit?

  Do they converge to some fixed number (fluctuate around it) or do they keep increasing in magnitude (absolute value)? Give a short explanation for your answer. What happens if we add regularization?

  If the data is linearly separable the gradient will converge to infinity, because the difference between the two classes becomes bigger and bigger. That is why it's a good idea to add a regularization parameter, since it will penalize the weight vector from getting too big.

- Bonus Question: Convexity of negative log likelihood. Show that the negative log likelihood function for logistic regression is convex. Is it still convex if we add regularization?