IN4320 Machine Learning Exercise

February 17, 2016

Exercises Regularization & Sparsity

We are going to consider L_p-norm ($p > 1$) regularized regression of N outputs y_i onto d-dimensional inputs or feature vectors x_i:

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{N} (w^T x_i - y_i)^2 + \lambda \|w\|_p^p \,, \tag{1}$$

with

$$\|w\|_p^p = \sum_{i=1}^{d} |w_i|^p \,. \tag{2}$$

For $p = 2$, we know that there is a closed-form solution to the minimization of (1).

Through BlackBoard you can find a simple two-class digit classification task in 64 dimensions. It is named `optdigitsubset` and consists of the pixel values of small 8×8 images, which are ordered in rows of 64 columns wide. The first 554 rows contains the values of 8×8 images of zeros, while the remaining block of 571 rows contains the 64 pixel values of 571 ones. The actual feature vectors are obtained by running through the rows of the images of the digits from top to bottom, concatenating all 8 rows of 8 pixels into a 64-dimensional vector.

With this choice of data, the x_i will be the raw pixel values from the data set of your choosing. As we are going to consider a classification task, encode the outputs y_i of the one class by -1 and of the other class by $+1$.

Go through the following exercises and provide the necessary code, outputs, and derivations. At the end of every exercise it is indicated between parentheses what you are, roughly, expected to provide (in print).

a Implement the closed-form solution of regular ridge regression ($p = 2$). (code)

b For different choices of λ, one gets different solutions w. Like all x_i, the ws can be interpreted as images as well. Determine the solution for a couple of choices of λ (e.g. 10^{-3}, 10^{-2}, ..., 10^{+3}), turn these solutions into images (8×8 in case you use the `optdigitsubset`), and plot these solution images. (plots of images)

c Explain that if the two classes are equally sized, with the $-1/+1$ encoding of the output, that for λ large the solution w is basically proportional to the difference between the two class means. Check this by performing the computation for large λ and comparing the outputs. (explanation in words and/or formulas, image plots)

d Minimization of (1) can also be done through *gradient descent*. Show that the derivative of (1) (for $p = 2$) to w_k (i.e., the kth coordinate of w) equals

$$2\sum_{i=1}^{N}\left[\left(\sum_{j=1}^{d} x_{ij}w_j\right)x_{ik} - x_{ik}y_i\right] + 2\lambda w_k\,, \tag{3}$$

where x_{ij} is the value of the jth feature of vector i and $\sum_{j=1}^{d} x_{ij}w_j$ equals $w^T x_i$. (derivation of the gradient)

e Use this gradient to implement an optimization of the regularized least squares by means of gradient descent. You are welcome to exploit optimization toolboxes etc. to accomplish this task. It should, however, not be too difficult to get it up and running without these. (code or relevant code snippets and comments)

f Compare some solutions obtained under **b** with those found by means of the implementation through gradient descent from **e**. (image plots, difference images)

g We are now going to extend the implementation in **e** to other norms. To start with, show that the derivative of $\lambda\|w\|_p^p$ to w_k equals

$$\lambda p w_k |w_k|^{p-2} = \lambda p \,\mathrm{sign}(w_k)|w_k|^{p-1}\,. \tag{4}$$

(Strictly speaking, this equality only holds for $p > 2$.) Check that this answer corresponds with your earlier answer for $p = 2$. (derivation of the gradient)

h By means of the gradient determined under **g**, extend your gradient descent optimization such that it can deal with the general L_p norm regularizer. (code or relevant code snippets and comments)

i Take a fairly large value for λ (e.g. 10^{+3} or 10^{+6}; you might have to experiment a bit) and take only few examples per class (e.g. only 10 per class or so.) Again, plot some images of solutions w, but now with varying values of p between 1 and 2. Show what happens when p get closer and closer to 1. Can you explain why some of the pixel values of w get very close to zero? (image plots, explanation in words and/or formulas)

j **(bonus)** Let us now consider $p = 1$. Explain, based on your experimentation and/or on theoretical or geometric considerations, why *only very few pixels of w can be nonzero* when few training images are used. Explain why this is irrespective of the value of λ (as long as it is positive). (an insightful explanation)