IN4320 Machine Learning Exercise

February 15, 2017

Exercises Regularization & Sparsity

We are going to consider a regularized version of the nearest mean classifier (NMC) for two-class data. In the general setting, training data consists of pairs of d-dimensional feature vectors x_i and their label y_i, the latter of which we agree to encode with an element from the set $\{+,-\}$. The two means for the two classes are also indexed $+$ and $-$, i.e., m_- belongs to the class with labels $-$, while m_+ belongs to the $+$ class. Now, consider the following loss function (or objective function) L:

$$L(m_-, m_+) := \sum_i^N \|x_i - m_{y_i}\|^2 + \lambda\|m_- - m_+\|_1 \,, \tag{1}$$

with λ the regularization parameter, N the number of samples in the training set, and $\|\cdot\|_1$ the L_1-norm. Minimizing L for both m_- and m_+ gives us the solution to the regularized loss (for that λ). These optimal means, in turn, define the regularized NMC, which classifies new data to the mean nearest to that data point.

Before you start: my assessment is that if your report goes significantly over 4 pages, you are probably on the wrong track.

Some Optima & Some Geometry

1 Assume $d = 1$ and we are in a situation where we know m_- is fixed to 1 and we only have to optimize for m_+. The only observations that we have for that $+$ class are $x_1 = -1$ and $x_2 = 1$.

- **a** Draw the loss function as a function of m_+ for all $\lambda \in \{0, 2, 4, 6\}$. Be precise. A rough sketch or artist impression is not enough.
- **b** Derive for every of the four functions the minimizer and their minimum values. Also determine for every loss the point where the derivative equals 0.

2 We now consider the setting in which *both* means have to be determined through a minimization of the loss. In this case, we have $L : \mathbb{R}^2 \to \mathbb{R}$.

- **a** Generally, what does the regularizer in Equation (1) actually try to enforce and what will, therefore, eventually happen to the means if λ gets larger and larger (i.e., what is the limiting behavior of the two solution means)?

b Precisely describe how the contour lines for the general function L typically look like when we are trying to find two 1-dimensional class means.

Hints: 1) the loss is convex, so the contours are convex as well and 2) the contours consist of the concatenation of two basic geometric shapes.

c Let us consider a handful of data points. For the $+$ class we have the same observations as in Exercise 1. For the $-$ class we now have observations $x_3 = 3$ and $x_4 = -1$. Clearly, if we set λ to 0, we would find as optimal solution $(m_-, m_+) = (1, 0)$. Assume we have a large enough λ as under Exercise 2a: determine the exact solution (m_-, m_+) in that case.

Some Programming & Some Experimenting

Through BlackBoard you can find a two-class digit classification task in 64 dimensions, i.e., $d = 64$. It is named `optdigitsubset` and consists of the pixel values of small 8 × 8 images, which are ordered in $N = 1125$ 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.

3 Implement an optimizer for the regularized NMC from Equation (1) and convince yourself that it indeed optimizes what it should optimize. You can use gradient descent or any other approach of your liking. You are even allowed to use optimization toolboxes and the like. You can either implement a general version of the regularized NMC or one that is completely dedicated to the data set that is given. (Note, however, that the latter is probably not necessarily easier to implement and it is probably harder to debug.)

a Describe *in no more than 200 words* the main ingredients of and/or considerations behind your optimization routine. In particular, sketch the search strategy that you employ and explain how you decide when you have reached the sought-after minimum.

b Similar to all 64-dimensional samples in the data sets, you can restructure the 64-dimensional solution means that you find into an 8 × 8 images again. (For instance, in Matlab you can just use something like `reshape(m,[8 8])`.) Plot the two solution mean images that you find for $\lambda = 0$ and plot the two solution images for a large λ (i.e., one for which the solution does not change anymore with an even further increase of λ).