Convex Optimization for Machine Learning and Computer Vision

Lecture: Dr. Virginia Estellers
Exercises: Emanuel Laude
Winter Semester 2017/18

Computer Vision Group
Institut für Informatik
Technische Universität München

Weekly Exercises 6

Room: 02.09.023

Friday, 08.12.2017, 09:15-11:00

Submission deadline theory: Monday, 04.12.2017, 10:15, Room 02.09.023

Theory: Consensus Primal-Dual for Sparse SVMs (16 Points)

Exercise 1 (16 Points)**.** In this exercise you are asked to derive the explicit consensus Primal-Dual (PDHG) for sparse binary SVM training. To this end let $\mathcal{I} = \{1, \ldots, N\}$ denote a set of training samples that are represented by their feature vectors $x_i \in \mathbb{R}^d$, for $i \in \mathcal{I}$. Each training sample i is associated with a binary class label $y_i \in \{-1, 1\}$. The aim is to estimate a linear classifier parameterized by $w^* \in \mathbb{R}^d, b^* \in \mathbb{R}$ so that $y_i = \text{sign } x_i^\top w^* + b^*$ for most training samples i. Like in the logistic regression task from the previous exercise sheet we assume that the feature vectors are degraded by components containing just noise. In order to jointly train the classifier and select the features we attempt to optimize the model

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \sum_{i=1}^{N} \ell(w, b; x_i, y_i) + \lambda\|w\|_1, \tag{1}$$

where $\ell(\cdot, \cdot; x_i, y_i)$ is the hinge loss defined according to

$$\ell(w, b; x_i, y_i) := \max\{0, 1 - (\langle x_i, w\rangle + b)y_i\}. \tag{2}$$

For consensus primal-dual we consider an equivalent "lifted" formulation of the above problem: We introduce for each training sample i a "copy" of the classifier variable $w_i = w$, $b_i = b$ and consider the linearly constrained problem

$$\begin{aligned} \min_{\substack{w\in\mathbb{R}^d, b\in\mathbb{R} \\ \{w_i\},\{b_i\}}} \quad & \sum_{i=1}^{N} \ell(w_i, b_i; x_i, y_i) + \lambda\|w\|_1 \\ \text{subject to} \quad & w_i = w \\ & b_i = b, \forall i. \end{aligned} \tag{3}$$

You are asked to do the following:

1. Bring the consensus model (3) into the standard form

   $$\min_x F(Ax) + G(x). \tag{4}$$

   Identify the operator A, the functions F and G. What is the optimization variable x? What is its dimension?

   Hint: A is a tall matrix of stacked identities.

2. Derive the equivalent Fenchel-Legendre saddle point formulation. Explicitly derive the convex-conjugate F^* of the function F.

   Hint: exploit that F "separates" over the training examples, i.e. F is of the form $F(z) := \sum_i f_i(z_i)$, what is the dimension of the variable z? Note that $\dim(y) \gg \dim(x)$.

3. Explicitly derive closed form prox-operators of the functions F^* and G.

4. Explicitly state the PDHG update scheme for sparse SVM optimization.

5. Argue, why this formulation is well suited to distributed (parallel) optimization in large scale machine learning. What are the shortcomings of this formulation? How can the performance potentially be improved?

   Hint: Suggested Reader: Boyd et al., Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers, Chapters 7.1, 7.2.