Exercises
Computational Intelligence Lab
SS 2017

Machine Learning Institute
Dept. of Computer Science, ETH Zürich
Prof. Dr. Thomas Hofmann
Web http://cil.inf.ethz.ch/

Series 4, March 23-24, 2017 (Optimization)

Problem 1 (Lagrange Dual Problem for a Linear Program):

Let $\mathbf{x} \in \mathbb{R}^m, \mathbf{b} \in \mathbb{R}^p$, and $A \in \mathbb{R}^{p \times m}$. Find the Lagrange dual problem of the following standard form Linear Program (LP)

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \mathbf{c}^T \mathbf{x} \\ \text{subject to} \quad & A\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq 0, \end{aligned} \tag{1}$$

1. Form the Lagrangian $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ by introducing the Lagrange multipliers λ_i for each of the m inequality constraints and ν_i for each of the p equality constraints.
2. Write down the dual function $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$.
3. Find an analytical expression of the dual function $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$ by minimizing over $\mathbf{x}$. Use the fact, that a linear function is bounded below only when it is identically zero.
4. Write down the Lagrange dual problem to the primal problem (1). Use the dual function from above. Your objective function should again be linear.

Problem 2 (Dual Function is a Lower Bound on $f(\mathbf{x}^\star)$):

Show that the dual function $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$, for $\boldsymbol{\lambda} \geq 0$, is always a lower bound on the optimal value $f(\mathbf{x}^\star)$ of the primal problem:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & h_j(\mathbf{x}) = 0, \quad j = 1, \dots, p \end{aligned} \tag{2}$$

In other words, show that for any $\boldsymbol{\lambda} \in \mathbb{R}^m$, $\boldsymbol{\lambda} \geq 0$ and $\boldsymbol{\nu} \in \mathbb{R}^p$ we have:

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f(\mathbf{x}^\star)$$

Hints:

1. Write down the Lagrangian of problem (2).
2. Look at a feasible point $\mathbf{x}$ and check what bound this gives you on the weighted sum of the constraint functions.

Problem 3 (Convexity):

Which of the following claims are **true/false**?

a) The intersection of two convex sets is convex.

b) The set $\{\mathbf{u} \in \mathbb{R}^D \mid \|\mathbf{u}\| = 1\}$ is convex.

c) The *epigraph* of a function $f : \mathbb{R}^D \to \mathbb{R}$ is defined as

$$\{(\mathbf{x}, t) \mid \mathbf{x} \in dom\, f, f(\mathbf{x}) \geq t\}$$

d) The function $f(\mathbf{u}, \mathbf{v}) := g(\mathbf{u}\mathbf{v}^T)$ is convex over the set of vectors $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^2 \times \mathbb{R}^2$, when $g : \mathbb{R}^{2\times 2} \to \mathbb{R}$ is defined as $g(\mathbf{X}) = X_{12} + X_{21}$.

Problem 4 (Stochastic Gradient Descent for Collaborative Filtering):

We have seen matrix completion already in Exercise 3, where we approximated a full matrix by an SVD.

In this exercise, we will apply *optimization techniques* to directly minimize the training error for the (unconstrained) matrix factorization formulation $\min_{\mathbf{U}\in Q_1, \mathbf{Z}\in Q_2} f(\mathbf{U}, \mathbf{Z})$, with the objective function being the mean squared error

$$f(\mathbf{U}, \mathbf{Z}) = \frac{1}{|\Omega|} \sum_{(d,n)\in\Omega} \frac{1}{2} \left[\mathbf{X}_{dn} - (\mathbf{U}\mathbf{Z}^T)_{dn}\right]^2 \tag{3}$$

and $\mathbf{U} \in Q_1 := \mathbb{R}^{D\times K}$, $\mathbf{Z} \in Q_2 := \mathbb{R}^{N\times K}$.

Here $\Omega \subseteq [D] \times [N]$ is the set of the indices of the observed ratings of the input matrix $\mathbf{X}$.

Environment setup:

Please use the same setup and data as in Exercise 3, as also explained on the web page for the collaborative filtering project task:

https://inclass.kaggle.com/c/cil-collab-filtering-2017.

Task: Implement Stochastic Gradient Descent

1. Derive the full gradient $\nabla_{(\mathbf{U},\mathbf{Z})} f(\mathbf{U}, \mathbf{Z})$. Note that since we have $(D + N) \times K$ variables, the gradient here can be seen as a $(D + N) \times K$ matrix.
2. Derive a stochastic gradient $\mathbf{G}$ using the sum structure of f over the Ω elements. We want to do this in such a way that $\mathbf{G}$ only depends on a single observed rating $(d, n) \in \Omega$.
3. Implement the Stochastic Gradient Descent algorithm as described in the lecture, for our objective function given in (3).
4. Experimentally find the best stepsize γ to obtain the lowest training error value.
5. Does the test error also decrease monotonically during optimization, or does it increase again after some time?
6. (OPTIONAL: Can you speed up your code, by for example maintaining the set of values $(\mathbf{U}\mathbf{Z}^T)_{dn}$ for the few observed values $(d, n) \in \Omega$, and thereby avoiding the computation of the matrix multiplication $\mathbf{U}\mathbf{Z}^T$ in every step?)

Extensions: Naturally there are many ways to improve your solution. One of them is to use regularization term to avoid over-fitting. Such techniques and other extensions can be found e.g. in the following publications:

- Webb, B. (2006). Netflix Update: Try This at Home. Simon Funk's Personal Blog. `http://sifter.org/~simon/journal/20061211.html`
- Koren Y., Bell R., Volinsky B., "Matrix Factorization Techniques for Recommender Systems" IEEE Computer, Volume 42, Issue 8, p.30-37 (2009); `http://research.yahoo.com/files/ieeecomputer.pdf`
- A. Paterek, "Improving Regularized Singular Value Decomposition for Collaborative Filtering," Proc. KDD Cup and Workshop, ACM Press, 2007, pp. 39-42.