Exercises
Data Mining: Learning from Large Data Sets
FS 2016

LAS Group, Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `http://las.ethz.ch/courses/dm-f16/`
Email questions to:
Kfir Levy, yehuda.levy@inf.ethz.ch

Series 4, Nov 10th, 2016 (SVM / Kernel)

It is not mandatory to submit solutions and sample solutions will be published after one week. If you choose to submit your solution, please send an e-mail from your `ethz.ch` address with subject `Exercise4` containing a PDF (LaTeX or scan) to `yehuda.levy@inf.ethz.ch` until Wednesday, Nov 16th 2016.

Problem 1 (Support Vector Machines):

The objective of this exercise is to investigate the $L2$-SVM which uses the square sum of the slack variables ξ_i in the objective function instead of the linear sum of the slack variables (i.e. squaring the hinge loss). Let $S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ be a training set of examples and binary labels $y_i \in \{-1, +1\}$. The primal formulation of the $L2$-SVM is as follows

$$\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n} {\xi_i}^2$$

s.t.

$$\begin{aligned} y_i \mathbf{w}^T \mathbf{x}_i &\geq 1 - \xi_i \quad i = 1, \ldots, n \\ \xi_i &\geq 0 \qquad\quad i = 1, \ldots, n \end{aligned} \tag{1}$$

- Reformulate the above optimization as an unconstrained optimization problem.
- Give a step-by-step solution to deriving the optimal parameters using stochastic gradient descent.

Solution 1 (Support Vector Machines):

1. We can get rid of the slack variables by setting

$$\xi_i = \begin{cases} 0; & \text{if} \quad y_i \mathbf{w}^T \mathbf{x}_i \geq 1 \\ 1 - y_i \mathbf{w}^T \mathbf{x}_i; & \text{otherwise.} \end{cases}$$

In other words each $\xi_i = \max(0, 1 - y_i \mathbf{w}^T \mathbf{x}_i)$. This leads to the following unconstrained optimization problem:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n} \underbrace{(\max(0, 1 - y_i \mathbf{w}^T \mathbf{x}_i))^2}_{\ell(\mathbf{w}, \mathbf{x}_i, y_i)}.$$

2. Denote

$$L(\mathbf{w}) = \sum_{i=1}^{n} \left[\frac{\|\mathbf{w}\|^2}{2n} + \frac{C}{2}\ell(\mathbf{w}; \mathbf{x}_i, y_i) \right].$$

Note that $\nabla_{\mathbf{w}}(\frac{\|\mathbf{w}\|^2}{2}) = \mathbf{w}$ and also

$$\nabla_{\mathbf{w}}\ell(\mathbf{w}; \mathbf{x}, y) = \begin{cases} 0; & \text{if} \quad y\mathbf{w}^T\mathbf{x} \geq 1 \\ -2(1 - y\mathbf{w}^T\mathbf{x})y\mathbf{x}; & \text{otherwise.} \end{cases}$$

Hence we obtain the following stochastic gradient descent algorithm.

Stochastic gradient descent

Set $\mathbf{w} \leftarrow 0$ and choose learning rate β_t (For eg. $\beta_t = 1/t$.)
repeat until convergence
Pick $(\mathbf{x}_i, y_i)$ from training set uniformly at random.
if $y_i \mathbf{w}^T \mathbf{x}_i \geq 1$
$\mathbf{w} \leftarrow \mathbf{w} - \beta_t(\frac{\mathbf{w}}{n})$
else
$\mathbf{w} \leftarrow \mathbf{w} - \beta_t(\frac{\mathbf{w}}{n} - C(1 - y_i \mathbf{w}^T \mathbf{x}_i)y_i \mathbf{x}_i)$
end

Problem 2 (Deriving the SVM Dual):
Consider the following SVM formulation:

$$\text{minimize}_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \tag{2}$$

subject to

$$y_i \mathbf{w}^T \mathbf{x} \geq 1 - \xi_i \quad \text{for all} \quad i = 1, \cdots, n \tag{3}$$

and

$$\xi_i \geq 0 \quad \text{for all} \quad i = 1, \cdots, n \tag{4}$$

- Write down the Lagrangian using α_i as the Lagrange multiplier corresponding to constraint 3 and γ_i as the Lagrange multiplier corresponding to constraint 4.
- Compute the derivative of the Lagrangian with respect to $\mathbf{w}$ and ξ_i.
- Solve for the dual and show it is given by

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T, \mathbf{x}_j. \tag{5}$$

subject to

$$0 \leq \alpha_i \leq C \quad \text{for} \quad i = 1, \cdots, n \tag{6}$$

Solution 2 (Deriving SVM dual):

1. The Lagrangian is

$$\mathcal{L}(w, \xi, \alpha, \gamma) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(y_i w^T x_i - 1 + \xi_i) - \sum_{i=1}^{n}\gamma_i\xi_i$$

2. Setting the derivatives to zero:

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0$$
$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \gamma_i - \alpha_i = 0$$

leads to

$$w = \sum_{i=1}^{n}\alpha_i y_i x_i$$
$$C = \gamma_i + \alpha_i$$

3. Inserting into the primal leads to

$$\begin{aligned}
\mathcal{L}(w, \xi, \alpha, \gamma) &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(y_i w^T x_i - 1 + \xi_i) - \sum_{i=1}^{n}\gamma_i\xi_i \\
&= \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle + C\sum_i \xi_i - \sum_i \gamma_i\xi_i - \sum_i \alpha_i\xi_i \\
&\quad - \sum_i \alpha_i y_i \langle \sum_j \alpha_j y_j x_j, x_i\rangle + \sum_i \alpha_i \\
&= \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle
\end{aligned}$$

Problem 3 (Kernelized Ridge Regression):

Consider the following ridge regression problem:

$$\min_{\mathbf{w}} \sum_{i=1}^{n} ||y_i - \mathbf{w}^T\mathbf{x}_i||_2^2 + \lambda||\mathbf{w}||_2^2,$$

- Compute the derivative of the above objective function with respect to **w**, and derive its closed form solution.

- Show that the closed form solution you derived in the previous step can be written as

$$\mathbf{w} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{y}.$$

- Substitute $\mathbf{w} = \mathbf{X}^T\alpha$ to kernelize the above ridge regression problem.

Solution 3 (Kernelized Ridge Regression):

1.

$$\begin{aligned} L(\mathbf{w}) &= \sum_{i=1}^{n} ||y_i - \mathbf{w}^T\mathbf{x}_i||_2^2 + \lambda||\mathbf{w}||_2^2 \\ &= (\mathbf{y} - X\mathbf{w})^T(\mathbf{y} - X\mathbf{w}) + \lambda||\mathbf{w}||_2^2 \end{aligned}$$

$$\frac{\partial L(w)}{\partial w} = -2\mathbf{X}^T(\mathbf{y} - X\mathbf{w}) + 2\lambda\mathbf{w}$$

Setting the derivative equal to zero, we get

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{7}$$

2. Note that

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{X}^T = \mathbf{X}^T\mathbf{X}\mathbf{X}^T + \lambda\mathbf{X}^T = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})$$

Multiplying each part of the equation by $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}$ at the left and $(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{y}$ at the right, we have

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{y} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{y}$$

Simplifying the equation we get

$$\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{y} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

From Eq. 7 we can see that the right hand side is just the value of $\mathbf{w}^*$, and thus,

$$\mathbf{w}^* = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{y}$$

3. Substituting $\mathbf{w} = \mathbf{X}^T\alpha$ in $L(\mathbf{w})$, we get

$$\begin{aligned} L(\mathbf{X}^T\alpha) &= ||\mathbf{y} - \mathbf{X}\mathbf{X}^T\alpha||_2^2 + \lambda||\mathbf{X}^T\alpha||_2^2 \\ &= ||\mathbf{y} - \mathbf{X}\mathbf{X}^T\alpha||_2^2 + \lambda\alpha^T\mathbf{X}\mathbf{X}^T\alpha \end{aligned}$$

The kernel matrix can be written as $\mathbf{K} = \mathbf{X}\mathbf{X}^T$, which we can substitute into L to get

$$L(\mathbf{X}^T\alpha) = ||\mathbf{y} - \mathbf{K}\alpha||_2^2 + \lambda\alpha^T\mathbf{K}\alpha$$

For a new unseen data point x we predict its target value $\mathbf{y}$ as

$$\begin{aligned}\mathbf{y} = \mathbf{w}^T\mathbf{x} &= (\mathbf{X}^T\alpha)^T\mathbf{x} \\ &= \alpha^T\mathbf{X}\mathbf{x} \\ &= \sum_{i=1}^{n}\alpha_i\mathbf{x}_i^T\mathbf{x} \\ &= \sum_{i=1}^{n}\alpha_i k(\mathbf{x}_i, \mathbf{x})\end{aligned}$$

from which we see that we can also predict using only the kernel, without the need for any operations in the feature space.