Kernel-Based Learning & Multivariate Modeling
DMKM Master - MIRI Master

Lluís A. Belanche
belanche@cs.upc.edu

Problem Set #2, Sept 21, 2016

Problem 1 Solving simple extremal problems 1

Consider the real two-dimensional space; in this space planes are lines.

1. Find the point $\boldsymbol{p}$ on a given plane π that is closest to the origin; in other words, derive the formula

$$d(\mathbf{0}, \pi) = \frac{|b|}{\|\boldsymbol{w}\|}$$

   where $\pi : \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0$ or $w_2 x_2 + w_1 x_1 + b = 0$.

2. Find the point $\boldsymbol{p}$ on a given plane π that is closest to a point $\boldsymbol{q}$; in other words, derive the formula

$$d(\boldsymbol{q}, \pi) = \frac{|g(\boldsymbol{q})|}{\|\boldsymbol{w}\|}$$

   where $g(\boldsymbol{p}) = \langle \boldsymbol{w}, \boldsymbol{p} \rangle + b$. Note that the previous case corresponds to $\boldsymbol{q} = \mathbf{0}$.

. . . .

Problem 2 Solving simple extremal problems 2

Consider the two-dimensional optimization problem:

$$\textbf{maximize } 2 - x^2 - 2y^2$$
$$\textbf{subject to } x^2 + y^2 = 1$$

Give a graphical/geometrical explanation of the problem and solve it using Lagrange multipliers.

. . . .

Problem 3 Solving simple extremal problems 3

Consider the circle formed by the intersection of the unit sphere with the plane $x+y+z=0.5$. Consider the three-dimensional optimization problem of finding the point on this circle that is closest to the point $(1,2,3)$. Solve it using Lagrange multipliers.

. . . .

Problem 4 SVM Lagrangian (from scratch to dual), 1-norm version

In class, we considered the problem of finding the OSH in different scenarios:

1. In the separable case, we expressed it as:

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{subject to} \quad t_n(\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle + b) \geq 1, \quad n = 1, \dots, N \end{aligned}$$

   (a) Obtain the Lagrangian $\mathcal{L}$ for this linear SVM in primal form.
   (b) Take the gradient of $\mathcal{L}$ with respect to $\boldsymbol{w}, \varepsilon_n$ and b and make it equal to zero; obtain the other conditions on the solution (including the KKT conditions).
   (c) Substitute the resulting equations back into $\mathcal{L}$ to obtain its dual form $\mathcal{L}_D$, just in terms of the dual variables $\boldsymbol{\alpha}$.

2. In the non-separable case, we expressed it as:

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\varepsilon_n \\ &\text{subject to} \quad t_n(\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle + b) \geq 1 - \varepsilon_n, \quad \varepsilon_n \geq 0 \quad n = 1, \dots, N \end{aligned}$$

   (a) Obtain the Lagrangian $\mathcal{L}$ for this linear SVM in primal form.
   (b) Take the gradient of $\mathcal{L}$ with respect to $\boldsymbol{w}, \varepsilon_n$ and b and make it equal to zero; obtain the other conditions on the solution (including the KKT conditions).
   (c) Substitute the resulting equations back into $\mathcal{L}$ to obtain its dual form $\mathcal{L}_D$, just in terms of the dual variables $\boldsymbol{\alpha}$. Argue why this form should be maximized, subject to $0 \leq \alpha_n \leq C$.
   (d) Do the primal and dual forms scale with dimensionality or with training data size? Reason when it will be more convenient to solve one or the other problem.

. . . .

Problem 5 SVM Lagrangian (from primal to dual), 2-norm version

In class, we considered the non-separable case by penalizing the missclassified training examples with a "sum of violations" term which leads to the primal Lagrangian in Problem 4.2. Consider an alternative departing point, where instead we penalize the sum of the *squared* slacks:

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\varepsilon_n^2 \\ &\text{subject to} \quad t_n(\langle \boldsymbol{w}, \boldsymbol{x}_n\rangle + b) \geq 1 - \varepsilon_n, \quad \varepsilon_n \geq 0 \quad n = 1, \ldots, N \end{aligned}$$

1. Starting with this new objective function, write down the new Lagrangian for this soft linear SVM in primal form.
2. Optimize this primal Lagrangian with respect to $\boldsymbol{w}$ and b, plug these solutions back in and write the optimization problem just in terms of the dual variables $\boldsymbol{\alpha}$. How does this compare to the dual formulation for the standard 1-norm soft SVM?

. . . .

Problem 6 Feature maps and kernels

Given a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^D$, we define its associated kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as $k(\boldsymbol{u}, \boldsymbol{v}) = \langle \phi(\boldsymbol{u}), \phi(\boldsymbol{v})\rangle$, $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$. Show that the dual form obtained in Problem 4.2 can be kernelised to yield:

$$\mathcal{L}_D = \sum_{n=1}^{N}\alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m t_n t_m\, k(\boldsymbol{x}_n, \boldsymbol{x}_m)$$

and the dual decision function becomes

$$y_{\text{SVM}}(\boldsymbol{x}) = \text{sgn}\left(\sum_{n=1}^{N}\alpha_n t_n k(\boldsymbol{x}, \boldsymbol{x}_n) + b\right)$$

Is the primal form $y_{\text{SVM}}(\boldsymbol{x}) = \text{sgn}(\langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle + b)$ always computable? What if $D = \infty$? What about $\boldsymbol{w}$? and what about the margin in feature space?

. . . .

Problem 7 Connection to regularization

The SVM formulation has a direct connection to the regularization framework. Consider again a regularized empirical error, expressed in the general form:

$$E_\lambda(\boldsymbol{w}) = \frac{1}{N}\sum_{n=1}^{N} L(t_n, \langle \boldsymbol{w}, \phi(\boldsymbol{x}_n)\rangle) + \lambda\,\langle \boldsymbol{w}, \boldsymbol{w}\rangle\,, \qquad \lambda > 0$$

where $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a suitable *loss* function and $\lambda > 0$. Show that the primal for the soft SVM problem can be cast as a regularization problem $E_\lambda(\boldsymbol{w}, b)$, with the choices $\lambda = (2NC)^{-1}$ and $L(t_n, \langle \boldsymbol{w}, \phi(\boldsymbol{x}_n)\rangle) = \max(1 - t_n(\langle \boldsymbol{w}, \phi(\boldsymbol{x}_n)\rangle + b), 0)$ (the *hinge loss*).

. . . .