Solutions Sheet

Nina Fischer and Yannick Zelle

November 23, 2021

Exercise 1

Given: Let $m \leq n \leq k, y \in \mathbb{R}^m, b \in \mathbb{R}^k$ and $A \in \mathbb{R}^{mxn}, B \in \mathbb{R}^{kxn}$ We are considering the follwoing optimization optimization problem:

$$\min_{x \in \mathbb{R}^n} ||Ax - y||_2^2$$
$$\text{s.t.} Bx = b$$

Task: Find a matrix $P \in \mathbb{R}^{(n+k)x(n+k)}$ and a vector $p \in \mathbb{R}^{n+k}$ such that solving :

$$P \begin{bmatrix} x \\ \lambda \end{bmatrix} = p$$

gives a critical point for the optimization problem.

Proof. **Solution:** We will start by defining the Langragian function associated to this problem:

$$L(\lambda) = ||Ax - y||_2^2 + \lambda^T \cdot (Bx - b)$$

We will now search for the derivatives with respect to x and λ by using Matrix differential calculus:

- We will start by calculating $D_x L$

$$\begin{aligned} dL &= d||Ax - y||_2^2 + d\lambda^T (Bx - b) \\ &= d(Ax - y)^T (Ax - y) + \lambda^T B dx \\ &= 2(Ax - y)^T d(Ax - y) + \lambda^T B dx \\ &= 2(Ax - y)^T A dx + \lambda^T B dx \end{aligned}$$

So :

$$D_x L = 2(Ax - y)^T A + \lambda^T B$$

- We will now calculate $D_\lambda L$:

$$\begin{aligned} dL &= d||Ax-y||_2^2 + d\lambda^T(Bx-b) \\ &= (Bx-b)^T d\lambda \end{aligned}$$

So we have :

$$\nabla L = (2(Ax-y)^T A dx + \lambda^T B, (Bx-b)^T)$$

- For $\nabla L = 0$ we have :

$$\begin{aligned} \begin{bmatrix} 0 \\ 0 \end{bmatrix} &= \begin{bmatrix} (Bx-b)^T \\ 2(Ax-y)^T A + \lambda^T B \end{bmatrix} \\ \leftrightarrow \begin{bmatrix} b \\ 2y^T A \end{bmatrix} &= \begin{bmatrix} Bx \\ 2(Ax)^T A^T A + \lambda^T B \end{bmatrix} \\ \leftrightarrow \begin{bmatrix} b \\ 2(y^T A)^T \end{bmatrix} &= \begin{bmatrix} Bx \\ 2A^T Ax + B^T \lambda \end{bmatrix} \\ \leftrightarrow \begin{bmatrix} b \\ 2(y^T A)^T \end{bmatrix} &= \begin{bmatrix} B & 0^{kxn} \\ 2A^T A & B^T \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} \end{aligned}$$

So with $p = \begin{bmatrix} b \\ 2(y^T A)^T \end{bmatrix}$ and $P = \begin{bmatrix} B & 0^{kxn} \\ 2A^T A & B^T \end{bmatrix}$ solving

$$P \begin{bmatrix} x \\ \lambda \end{bmatrix} = p$$

will give a critical point to the optimization problem.

□

Exercise 2

Given: Let $v \in [0,1]$. We then have the follwoing optimization problem:

$$\begin{aligned} &\min_{w,b,\xi,p} \frac{1}{2}||w||^2 - vp + \frac{1}{n}\sum_{i=1}^{n} \xi_i \\ &\text{s.t} y_i(w^T x_i + b) \geq p - \xi_i \forall i \\ &\xi \geq 0 \forall i \\ &p \geq 0 \end{aligned}$$

(a) The Langragian is given by]

$$
\begin{aligned}
L(w,b,\xi,p,\alpha,\beta,\delta) &= \frac{1}{2}||w||^2 - vp + \frac{1}{n}\sum_i \xi_i - \sum_i \alpha_i(y_i(< x_i, w > +b) - p + \xi_i) - \sum_i \xi_i\beta_i - \delta p \\
&= \frac{1}{2}||w||^2 - p(v + \delta - \sum_i \alpha_i) - \sum_i (\frac{1}{n} - \alpha_i + \beta_i)\xi_i - < \sum_i \alpha_i y_i x_i, w > -(\sum \alpha_i y_i)b
\end{aligned}
$$

(b) The corresponding partial derrivatives with respect to w, b, ξ, p are given by:

$$\frac{\partial L}{\partial w} = w - \sum_i \alpha_i y_i x_i \tag{1}$$

$$\frac{\partial L}{\partial b} = -\sum_i \alpha_i y_i \tag{2}$$

$$\frac{\partial L}{\partial \xi} = \sum \frac{1}{n} - \alpha_i + \beta_i \tag{3}$$

$$\frac{\partial L}{\partial p} = v + \delta - \sum_i \alpha_i \tag{4}$$

c Setting the partial derrivatives to zero gives us :

$$L(w_0, b_0, \xi_0, p_0, \alpha, \beta, \delta)\frac{1}{2} < \sum_i \alpha_i y_i x_i, \sum_\alpha {}_i y_i x_i >$$

]

d Because **(3)** we have :

$$\frac{1}{n} - lpha_i \beta_i = 0$$

and since $\alpha_i, \beta_i \geq 0$ we have :

$$0 \leq \alpha_i \leq n$$

Furthermore we get from **(4)** set to 0 :

$$\sum \alpha_i = v + \delta$$

and since $\delta \geq 0$ we have

$$\sum_i \alpha_i \geq v$$

(e) Thus we have the dual problem given by :

$$\max_{\alpha} \frac{1}{2} < \sum_i \alpha_i y_i x_i, \sum_i \alpha_i y_i x_i >$$
$$\text{s.t.} \frac{1}{n} \geq \alpha_i \forall i$$
$$\sum_i \alpha_i \geq v \forall i$$
$$\alpha_i \geq 0 \forall$$

Exercise 3

In the first case where we would replace the 1 with a 0, the SVM wouldn't work anymore because to minimize w we could just plug in the zero vector no matter the training examples. Thus if our data is seperated is a question of poor luck.

The second case would generate us a seperating plane but it would not be optimal.

Exercise 4