Machine Learning Homework Sheet 06

Constrained Optimization and SVM

1 Constrained Optimization

Problem 1: Solve the following constrained optimization problem using the recipe described in the lecture (slide 17).

$$\begin{aligned} &\text{minimize} && f_0(\boldsymbol{\theta}) = \theta_1 - \sqrt{3}\theta_2 \\ &\text{subject to} && f_1(\boldsymbol{\theta}) = \theta_1^2 + \theta_2^2 - 4 \leq 0 \end{aligned}$$

Write down the **Lagrangian**

$$L(\boldsymbol{\theta}, \alpha) = \theta_1 - \sqrt{3}\theta_2 + \alpha\left(\theta_1^2 + \theta_2^2 - 4\right) .$$

Calculate the derivative of $L(\boldsymbol{\theta}, \alpha)$ w.r.t. $\boldsymbol{\theta}$ and set it to zero,

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \alpha) = \begin{pmatrix} 1 + 2\alpha\theta_1 \\ -\sqrt{3} + 2\alpha\theta_2 \end{pmatrix} = 0$$

$$\Rightarrow \theta_1 = -\frac{1}{2\alpha}, \quad \theta_2 = \frac{\sqrt{3}}{2\alpha} .$$

Substituting this back into $L(\boldsymbol{\theta}, \alpha)$ gives the **dual function**

$$g(\alpha) = -\frac{1}{2\alpha} - \frac{3}{2\alpha} + \frac{1}{4\alpha} + \frac{3}{4\alpha} - 4\alpha = -\frac{1}{\alpha} - 4\alpha .$$

We must now **maximize the dual function** $g(\alpha)$ subject to $\alpha \geq 0$.
Since $g(\alpha)$ is concave for $\alpha \geq 0$, we set its derivative to zero and solve for α.

$$\begin{aligned} & \frac{dg}{d\alpha} = \frac{1}{\alpha^2} - 4 = 0 \\ \Rightarrow \quad & \alpha^2 = \frac{1}{4} \\ \Rightarrow \quad & \alpha_{1,2} = \pm\frac{1}{2} \\ \Rightarrow \quad & \alpha^* = \frac{1}{2} \end{aligned}$$

since we require $\alpha \geq 0$. The optimization problem is convex and Slater's condition holds, therefore the minimal value of $f_0(\boldsymbol{\theta})$ is

$$\boldsymbol{p}^* = g\left(\frac{1}{2}\right) = -4 .$$

We calculate the minimizer $\boldsymbol{\theta}^*$ by substituting $\alpha = 1/2$ into $\theta_1 = -\frac{1}{2\alpha}$ and $\theta_2 = \frac{\sqrt{3}}{2\alpha}$ and get

$$\theta_1^* = -1, \quad \theta_2^* = \sqrt{3} .$$

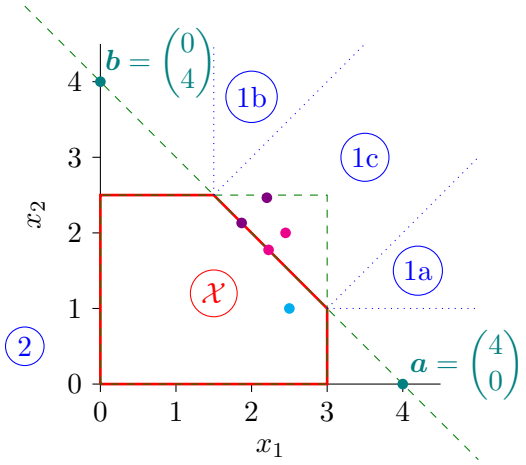
Upload a single PDF file with your solution to Moodle by 2.12.2018, 23:59pm CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted.
If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.

2 Projected Gradient Descent

Problem 2: Given is the following (convex) domain defined by a set of linear constraints

$$\mathcal{X} \subset \mathbb{R}^2 = \{\boldsymbol{x} \in \mathbb{R}^2 : (x_1 + x_2 \leq 4) \wedge (0 \leq x_1 \leq 3) \wedge (0 \leq x_2 \leq 2.5)\}.$$

a) Visualize the set $\mathcal{X}$.



b) Derive a closed form for the projection $\pi_{\mathcal{X}}(\boldsymbol{p}) = \arg\min_{\boldsymbol{x}\in\mathcal{X}} ||\boldsymbol{x} - \boldsymbol{p}||_2^2$. That is, given an arbitrary point $\boldsymbol{p} \in \mathbb{R}^2$, what is its projection on $\mathcal{X}$?

Hint: For one part of $\mathbb{R}^2$ you might want to use the line projection $\pi_{\text{line}}(\boldsymbol{p}) = \boldsymbol{a} + \frac{(\boldsymbol{p}-\boldsymbol{a})^T(\boldsymbol{b}-\boldsymbol{a})}{\|\boldsymbol{b}-\boldsymbol{a}\|_2^2}(\boldsymbol{b}-\boldsymbol{a})$, where $\boldsymbol{a} \in \mathbb{R}^2$ and $\boldsymbol{b} \in \mathbb{R}^2$ specify the line.

$$\pi_{\mathcal{X}}(\boldsymbol{p}) = \begin{cases} \boldsymbol{p} & \text{if } \boldsymbol{p} \in \mathcal{X} \\ (3,1)^T & \text{if } \boldsymbol{p} \in \text{1a} = \{\boldsymbol{p} | p_2 \geq 1 \wedge -p_1 + p_2 \leq -2\} \\ (1.5, 2.5)^T & \text{if } \boldsymbol{p} \in \text{1b} = \{\boldsymbol{p} | p_1 \geq 1.5 \wedge -p_1 + p_2 \geq 1\} \\ \pi_{\text{line}}(\boldsymbol{p}) & \text{if } \boldsymbol{p} \in \text{1c} = \{\boldsymbol{p} | -2 < -p_1 + p_2 < 1 \wedge p_1 + p_2 > 4\} \\ \pi_{\text{box}}(\boldsymbol{p}) & \text{if } \boldsymbol{p} \in \text{2} = \{\boldsymbol{p} | \boldsymbol{p} \notin \mathcal{X} \wedge (p_1 < 1.5 \vee p_2 < 1)\} \end{cases}$$

$$\begin{aligned} \pi_{\text{line}}(\boldsymbol{p}) &= \boldsymbol{a} + \frac{(\boldsymbol{p}-\boldsymbol{a})^T(\boldsymbol{b}-\boldsymbol{a})}{\|\boldsymbol{b}-\boldsymbol{a}\|_2^2}(\boldsymbol{b}-\boldsymbol{a}) = \begin{pmatrix}4\\0\end{pmatrix} + \frac{(p_1-4, p_2)\begin{pmatrix}-4\\4\end{pmatrix}}{32}\begin{pmatrix}-4\\4\end{pmatrix} \\ &= \begin{pmatrix}4\\0\end{pmatrix} + \frac{16 - 4p_1 + 4p_2}{32}\begin{pmatrix}-4\\4\end{pmatrix} = \begin{pmatrix}4\\0\end{pmatrix} + \left(2 - \frac{1}{2}p_1 + \frac{1}{2}p_2\right)\begin{pmatrix}-1\\1\end{pmatrix} \\ &= \begin{pmatrix}2 + \frac{1}{2}p_1 - \frac{1}{2}p_2 \\ 2 - \frac{1}{2}p_1 + \frac{1}{2}p_2\end{pmatrix} \\ \pi_{\text{box}}(\boldsymbol{p}) &= \begin{pmatrix}\max(0, \min(3, p_1)) \\ \max(0, \min(2.5, p_2))\end{pmatrix} \end{aligned}$$

Upload a single PDF file with your solution to Moodle by 2.12.2018, 23:59pm CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted.
If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.

c) Given is the following constrained optimization problem:

$$\min_{\boldsymbol{x}} (x_1 - 2)^2 + (2x_2 - 7)^2,$$
$$\text{subject to } \boldsymbol{x} \in \mathcal{X}.$$

Perform two steps of projected gradient descent starting from the point $\boldsymbol{x}^{(0)} = (2.5, 1)^T$. Use a constant learning rate/step size of $\tau = 0.05$.

Projected gradient descent is a two-step algorithm consisting of regular gradient descent and projection:

1. $\boldsymbol{p}^{t+1} = \boldsymbol{x}^t - \tau \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^t)$
2. $\boldsymbol{x}^{t+1} = \pi_{\mathcal{X}}(\boldsymbol{p}^{t+1})$

The gradient of this function is

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{pmatrix} 2x_1 - 4 \\ 8x_2 - 28 \end{pmatrix}.$$

Using this gradient and the projection from above we can perform two steps of the algorithm:

1.1 $\boldsymbol{p}^{(1)} = \boldsymbol{x}^{(0)} - \tau \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(0)}) = \begin{pmatrix} 2.5 \\ 1 \end{pmatrix} - 0.05 \begin{pmatrix} 1 \\ -20 \end{pmatrix} = \begin{pmatrix} 2.45 \\ 2 \end{pmatrix}$

1.2 $\boldsymbol{x}^{(1)} = \pi_{\mathcal{X}}(\boldsymbol{p}^{(1)}) = \pi_{\text{line}}(\begin{pmatrix} 2.45 \\ 2 \end{pmatrix}) = \begin{pmatrix} 2.225 \\ 1.775 \end{pmatrix}$

2.1 $\boldsymbol{p}^{(2)} = \boldsymbol{x}^{(1)} - \tau \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(1)}) = \begin{pmatrix} 2.225 \\ 1.775 \end{pmatrix} - 0.05 \begin{pmatrix} 0.45 \\ -13.8 \end{pmatrix} = \begin{pmatrix} 2.2025 \\ 2.465 \end{pmatrix}$

2.2 $\boldsymbol{x}^{(2)} = \pi_{\mathcal{X}}(\boldsymbol{p}^{(2)}) = \pi_{\text{line}}(\begin{pmatrix} 2.2025 \\ 2.465 \end{pmatrix}) = \begin{pmatrix} 1.86875 \\ 2.13125 \end{pmatrix}$

This process is also illustrated in the figure above, with points showing the starting point (cyan), step 1 (magenta) and step 2 (violet).

3 SVM

Problem 3: Explain the similarities and differences between the SVM and perceptron algorithms.

Both algorithms are looking for a hyperplane that separates the two classes.

The difference is that SVM also tries to maximize the margin, while the perceptron only cares about separation.

Problem 4: Show that the duality gap is zero for SVM.

The objective is convex and the constraints are affine in $\boldsymbol{w}$, thus the Slater's condition is satisfied, so the duality gap is zero.

Problem 5: Recall that the dual function for SVM (slide 41) can be written as

$$g(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T \boldsymbol{Q}\boldsymbol{\alpha} + \boldsymbol{\alpha}^T \mathbf{1}_N$$

(a) Show how the matrix $\boldsymbol{Q}$ can be computed. *(Hint: You might want to use Hadamard product, denoted as $\odot$).*

(b) Prove that the matrix $\boldsymbol{Q}$ is negative (semi-)definite.

(c) Explain what the negative (semi-)definiteness means for our optimization problem. Why is this property important?

(a) $\boldsymbol{Q} = (-\boldsymbol{y}\boldsymbol{y}^T \odot \boldsymbol{X}\boldsymbol{X}^T)$

(b) We will first show that $\boldsymbol{M} = (\boldsymbol{y}\boldsymbol{y}^T \odot \boldsymbol{X}\boldsymbol{X}^T) \in \mathbb{R}^{N\times N}$ is a positive (semi-)definite (PSD) matrix.

For $\boldsymbol{M}$ to be PSD, for every $\boldsymbol{a} \in \mathbb{R}^N$ it must hold that $\boldsymbol{a}^T\boldsymbol{M}\boldsymbol{a} \geq 0$.

$$\begin{aligned}
\boldsymbol{a}^T\boldsymbol{M}\boldsymbol{a} &= \sum_{i=1}^{N}\sum_{j=1}^{N} a_i y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j a_j \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} (a_i y_i \boldsymbol{x}_i)^T (a_j y_j \boldsymbol{x}_j) \\
&= (\boldsymbol{a}\odot\boldsymbol{y})^T \boldsymbol{X}\boldsymbol{X}^T (\boldsymbol{a}\odot\boldsymbol{y}) \\
&= (\boldsymbol{X}^T(\boldsymbol{a}\odot\boldsymbol{y}))^T (\boldsymbol{X}^T(\boldsymbol{a}\odot\boldsymbol{y})) \\
&= (\boldsymbol{X}^T(\boldsymbol{a}\odot\boldsymbol{y}))^2 \\
&\geq 0
\end{aligned}$$

As $\boldsymbol{M}$ is PSD, it follows that $\boldsymbol{Q} = -\boldsymbol{M}$ is negative (semi-)definite.

(c) This means that our maximization problem is **concave**, thus every local maximum is a global maximum.

Problem 6: Download the notebook `homework_06_notebook.ipynb` from Piazza. Fill in the missing code and run the notebook. Convert the evaluated notebook to pdf and add it to the printout of your homework (see printing instructions inside the notebook).