# Machine Learning HW#1

Evgeny Marshakov

## Problem 1

1. We have an optimization problem

$$\begin{aligned} \min_{w,\xi,b} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i^p \\ \text{subject to} \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned}$$

The lagrangian of this problem

$$\mathcal{L}(w,b,\xi,\alpha) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i^p - \sum_{i=1}^{m}\alpha_i[y_i((w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^{m}\beta_i\xi_i$$

From KKT conditions we can find that $w = \sum_{i=1}^{m}\alpha_i y_i x_i, \sum_{i=1}^{m}\alpha_i y_i = 0$ (the first two conditions are the same as in the case $p = 1$) and $\xi_i = (\frac{\alpha_i+\beta_i}{Cp})^{\frac{1}{p-1}}$, so the dual problem can be formulated as follows

$$\begin{aligned} \max_{\alpha} \quad & -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i \cdot x_j + \sum_{i=1}^{m}\alpha_i - \frac{C(p-1)}{(Cp)^{\frac{p}{p-1}}}\sum_{i=1}^{m}(\alpha_i+\beta_i)^{\frac{p}{p-1}} \\ & \text{subject to} \quad \sum_{i=1}^{m}\alpha_i y_i = 0, \alpha_i, \beta_i \geq 0 \end{aligned}$$

2. In the case $p = 2$ we have the following optimization problem

$$\begin{aligned} \min_{w,\xi,b} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i^2 \\ \text{subject to} \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned}$$

Let us show that we can get rid off the last constraint. Indeed, the following optimization problem is equivalent to the primal one

$$\min_{w,\xi,b} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}|\xi_i|^2$$

$$\text{subject to} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i$$

If we have $\xi_i < 0$ for any $i$, then we can replace it by $\xi_i^* = 0$, because

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \geq 1 - \xi_i^*$$

and

$$|\xi_i|^2 \geq 0 = |\xi_i^*|^2$$

So our primal optimization problem is equivalent to

$$\min_{w,\xi,b} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i^2$$

$$\text{subject to} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i$$

So we can get rid off the term with $\beta_i$ in the dual optimization problem for SVM, so we have the following

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i \cdot x_j + \sum_{i=1}^{m}\alpha_i - \frac{1}{4C}\sum_{i=1}^{m}(\alpha_i)^2$$

$$\text{subject to} \quad \sum_{i=1}^{m}\alpha_i y_i = 0, \alpha_i \geq 0$$

It can be easily seen that Hesse matrix of the objective function is

$$-\{(y_i x_i, y_j x_j)\}_{i,j=1}^{m} - \frac{1}{2C}I_m$$

which is obviously negative definite, because the matrix of scalar products is positive semidefinite and identity matrix is positive definite.

## Sparse SVM

1. We can take the following function

$$\Phi(x) = \{y_j x \cdot x_j\}_{j=1}^{m}$$

Then the constraint of SVM take the form

$$y_i \left( \sum_{j=1}^{m} w_j \cdot \Phi(x_i)_j + b \right) = y_i \left( \sum_{j=1}^{m} w_j y_j x_j \cdot x_i + b \right) \geq 1 - \xi_i$$

Hence, the result follows.

2. So we have an optimization problem

$$\begin{aligned} \min_{\alpha,\xi,b} \quad & \frac{1}{2}\|\alpha\|^2 + C\sum_{i=1}^{m} \xi_i \\ \text{subject to} \quad & y_i(\alpha \cdot \Phi(x_i) + b) \geq 1 - \xi_i \\ & \alpha_i, \xi_i \geq 0 \end{aligned}$$

The lagrangian of this problem is the following

$$\mathcal{L}(\alpha, \xi, b, \lambda, \mu, \nu) = \frac{1}{2}\|\alpha\|^2 + C\sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \lambda_i [y_i(\alpha \cdot \Phi(x_i) + b) - 1 + \xi_i] - \sum_{i=1}^{m} \mu_i \xi_i - \sum_{i=1}^{m} \nu_i \alpha_i$$

From KKT conditions we can find that

- $\nabla_\alpha \mathcal{L} = \alpha - \sum_{i=1}^{m} \lambda_i y_i \Phi(x_i) - \nu = 0 \Leftrightarrow \alpha = \sum_{i=1}^{m} \lambda_i y_i \Phi(x_i) + \nu$
- $\nabla_b \mathcal{L} = -\sum_{i=1}^{m} \lambda_i y_i = 0$
- $\nabla_{\xi_i} \mathcal{L} = C - \lambda_i - \mu_i \Leftrightarrow \lambda_i + \mu_i = C$

Plugging $\alpha$ we obtain the dual optimization problem

$$\begin{aligned} \max_{\mu,\gamma} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j \Phi(x_i) \cdot (\Phi(x_j) + \nu) - \frac{1}{2}\sum_{i=1}^{m} \lambda_i y_i \gamma \cdot \Phi(x_i) - \frac{1}{2}\sum_{i=1}^{m} \gamma_i^2 \\ \text{subject to} \quad & \sum_{i=1}^{m} \lambda_i y_i = 0 \\ & 0 \leq \lambda_i \leq C, \nu_i \geq 0 \end{aligned}$$

3. In the case $p = 1$ we have the following optimization problem

$$\begin{aligned} \min_{\alpha,\xi,b} \quad & \sum_{i=1}^{m} \alpha_i + C \sum_{i=1}^{m} \xi_i \\ \text{subject to} \quad & y_i(\alpha \cdot \Phi(x_i) + b) \geq 1 - \xi_i \\ & \alpha_i, \xi_i \geq 0 \end{aligned}$$

The lagrangian of this problem is as follows:

$$\mathcal{L}(\lambda, \mu.\nu, \alpha, b, \xi) = \sum_{i=1}^{m} \alpha_i + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \lambda_i [y_i(\alpha \cdot \Phi(x_i) + b) - 1 + \xi_i] - \sum_{i=1}^{m} \mu_i \alpha_i - \sum_{i=1}^{m} \nu_i \xi_i$$

From KKT conditions we can find that

- $\nabla_{\alpha_j} \mathcal{L} = \mathbf{1} - \sum_{i=1}^{m} \lambda_i y_i \Phi(x_i) - \mu = 0$
- $\nabla_b \mathcal{L} = - \sum_{i=1}^{m} \lambda_i y_i = 0$
- $\nabla_{\xi_i} \mathcal{L} = C - \lambda_i - \nu_i$

So the dual optimization problem is as follows

$$\begin{aligned} \max_{\lambda} \quad & \sum_{i=1}^{m} \lambda_i \\ \text{subject to} \quad & \sum_{i=1}^{m} \lambda_i y_i = 0 \\ & 0 \leq \lambda_i \leq C \end{aligned}$$

# Problem 2

Let us formulate WSVM optimization problem as follows

$$\begin{aligned} \min_{w,\xi,b} \quad & \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} p_i \xi_i \\ \text{subject to} \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned}$$

The lagrangian of this problem is as follows

$$\mathcal{L}(\alpha, \xi, b, w) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} p_i\xi_i - \sum_{i=1}^{m} \alpha_i[y_i(w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^{m} \beta_i\xi_i$$

From KKT conditions we can find that

- $\nabla_\alpha \mathcal{L} = w - \sum_{i=1}^m \alpha_i y_i x_i - \nu = 0 \Leftrightarrow w = \sum_{i=1}^m \alpha_i y_i x_i$
- $\nabla_b \mathcal{L} = -\sum_{i=1}^m \alpha_i y_i = 0$
- $\nabla_{\xi_i} \mathcal{L} = Cp_i - \alpha_i - \beta_i \Leftrightarrow \alpha_i + \beta_i = Cp_i$

So we have the dual optimization problem

$$\begin{aligned} \max_{\alpha} \quad & -\frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j x_i \cdot x_j + \sum_{i=1}^{m} \alpha_i \\ \text{subject to} \quad & \sum_{i=1}^{m} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq Cp_i \end{aligned}$$

## Problem 3

1. $K(x, y) = \cos(x - y)$ over $\mathbb{R} \times \mathbb{R}$. We see that $\cos(x - y) = \cos(x)\cos(y) + \sin(x)\sin(y)$, hence $K(x, y) = \Phi(x)^T\Phi(y)$, where $\Phi(x) = \begin{bmatrix} \cos(x) \\ \sin(x) \end{bmatrix}$.
2. $K(x, y) = (x + y)^{-1}$ over $(0, \infty) \times (0, \infty)$. Consider the following function

$$f(a) = \sum_{i,j} c_i c_j \frac{a^{x_i + x_j}}{x_i + x_j}$$

It can be easily seen that if $a \geq 0$ then

$$\frac{df}{da} = \sum_{i,j} c_i c_j a^{x_i + x_j - 1} = \frac{1}{a}\|\{c_i a^{x_i}\}_i\|^2 \geq 0$$

So function $f$ is non decreasing over $a \geq 0$, so $f(1) \geq f(0) \leftrightarrow \sum_{i,j} c_i c_j (x_i + x_j)^{-1} \geq 0$

3. $K(x,y) = \exp\{-\lambda \sin^2(x-y)\}$ with $\lambda > 0$ over $\mathbb{R} \times \mathbb{R}$. We note that

$$K(x,y) = \exp\{-\lambda \sin^2(x-y)\} = \exp\{-\lambda\} \cdot \exp\{\lambda \cos^2(x-y)\}$$

We know that power of PDS is PDS and that if $K$ is PDS then $\exp\{\lambda K\}$ is PDS. Hence the result follows.

## Problem 4

1. $K(x,y) = \sin^2(x-y)$ over $\mathbb{R} \times \mathbb{R}$. For any $x_1, \dots, x_n \in \mathbb{R}$ and $c_1, \dots, c_n \in \mathbb{R}$, s.t $\sum_i c_i = 0$:

$$\sum_{i,j} c_i c_j \sin^2(x_i - x_j) = \sum_{i,j} c_i c_j - \sum_{i,j} c_i c_j \cos^2(x_i - x_j) = -\sum_{i,j} c_i c_j \cos^2(x_i - x_j) \leq 0$$

So it is NDS.

2. $K(x,y) = \log(x+y)$ over $(0,\infty) \times (0,\infty)$, We know that $K$ is NDS iff $\exp\{-tK\}$ is PDS for all $t > 0$. Obviously, since $(x+y)^{-1}$ is PDS, the function $(x+y)^{-n}$ is PDS for all $n \in \mathbb{N}$. Hence, the kernel $(x+y)^{-t}$ is PDS for all $t > 0$, so the result follows.

## Problem 5

$$K(x,y) = (x^T y + c)^d = \sum_{k_0 + k_1 + \cdots + k_N = d} \binom{d}{k_0, k_1, \dots, k_N} c^{k_0} \prod_{i=1}^{N} (x_i y_i)^{k_i} =$$

$$= \sum_{k_0 + k_1 + \ldots + k_N = d} \binom{d}{k_0, k_2, \dots, k_N} c^{k_0} \prod_{i=1}^{N} (x_i)^{k_i} \prod_{i=1}^{N} (y_i)^{k_i} =$$

$$= \sum_{k_1 + \ldots + k_n \leq d}^{d} \binom{d}{d - k_1 - \ldots - k_N, k_1, \dots, k_N} c^{d - k_1 - \ldots - k_N} \prod_{i=1}^{N} (x_i)^{k_i} \prod_{i=1}^{N} (y_i)^{k_i} = \Phi(x)^T \Phi(y)$$

where

$$\Phi(x) = \left\{ p(k_1, k_2, \dots, k_N) x_1^{k_1} x_2^{k_2} \dots x_N^{k_N} \right\}_{k_1 + \ldots + k_N \leq d}$$

with $p(k_1, \dots, k_N) = \left( \binom{d}{d - k_1 - \ldots - k_N, k_1, \dots, k_N} c^{d - k_1 - \ldots - k_N} \right)^{\frac{1}{2}}$ So the dimension of the feature space associated to kernel $K$ is the number of monomials of $n+1$ variables of degree $d$. Thus, the dimension is $\binom{N+1+d-1}{N} =$

$\binom{N+d}{d}$. The kernek $K$ can be expressed in terms of $k_i(x, y) = (x \cdot y)^i$ as follows

$$(x^T y + c)^d = \sum_{i=0}^{d} \binom{d}{i} c^{d-i} k_i(x, y)$$

So the weight of kernel $k_i$ depends on $c$ and has the form $\binom{d}{i}c^{d-i}$.