Assignment 4
Machine Learning 1, Fall 2016

Dana Kianfar
University of Amsterdam

January 6, 2018

1 Lagrange Multipliers

1. We can observe that $f(x)$ is concave, therefore a unique maximum must exist.

$$L(x_1, x_2, \lambda) = f(x_1, x_2) + \lambda g(x_1, x_2) = 1 - x_2^2 - 2x_2^2 + \lambda(x_1 + x_2 - 1)$$

$$\nabla_{x_1} L(x_1, x_2, \lambda) = -2x_1 + \lambda = 0 \rightarrow \lambda = 2x_1 \rightarrow x_1 = \frac{\lambda}{2}$$

$$\nabla_{x_2} L(x_1, x_2, \lambda) = -4x_2 + \lambda = 0 \rightarrow \lambda = 4x_2 \rightarrow x_2 = \frac{\lambda}{4}$$

$$\nabla_\lambda = x_1 + x_2 - 1 = 0 \rightarrow \frac{\lambda}{2} + \frac{\lambda}{4} = 1 \rightarrow \lambda = \frac{4}{3}$$

$$\text{It follows that: } \mathbf{x_1} = \mathbf{\frac{2}{3}}, \mathbf{x_2} = \mathbf{\frac{1}{3}}$$

2. We can observe that $f(x)$ is concave, therefore a unique maximum must exist.

$$L(\mathbf{x}, \lambda) = 1 - \mathbf{x}^T\mathbf{x} + \lambda(\mathbf{a}^T\mathbf{x} - 1)$$

$$\text{Where } \mathbf{a} = \begin{bmatrix}1\\1\end{bmatrix}, \mathbf{x} = \begin{bmatrix}x_1\\x_2\end{bmatrix}, \mathbf{0} = \begin{bmatrix}0\\0\end{bmatrix}, f(\mathbf{x}) = 1 - \mathbf{x}^T\mathbf{x}, g(\mathbf{x}) = 1 - \mathbf{a}^T\mathbf{x}$$

$$\nabla_{\mathbf{x}} f = -2\mathbf{x} = \mathbf{0} \rightarrow \mathbf{x} = \mathbf{0}$$

We check whether the optima at $\mathbf{x} = \mathbf{0}$ satisfies the constraint $g(\mathbf{x}) = \mathbf{a}^T\mathbf{x} - 1 \geq 0$

$$g(\mathbf{x} = \mathbf{0}) = \mathbf{a}^T\mathbf{0} - 1 = -1$$

We must search for an optima at $g(\mathbf{x}) = 0$ as the constraint is not satisfied at $\mathbf{x} = \mathbf{0}$

i.e. the constraint is active with $\lambda > 0$

$$\nabla_{\mathbf{x}} L = -2\mathbf{x} + \lambda\mathbf{a} = \mathbf{0} \rightarrow \mathbf{x} = \frac{\lambda\mathbf{a}}{2}$$

$$\lambda > 0 \rightarrow g(\mathbf{x}) = 0 \rightarrow \mathbf{a}^T\mathbf{x} - 1 = 0 \rightarrow \mathbf{x} = \frac{1}{2}\mathbf{a}$$

$$\mathbf{x_1} = \mathbf{x_2} = \mathbf{\frac{1}{2}}, \boldsymbol{\lambda} = \mathbf{1}$$

3. The constraint $g(\mathbf{x})$ in this problem is the complement of problem 2. Therefore we expect the constraint here to be inactive, and know that the global maximum of the concave function $f(\mathbf{x})$ is

also the maximum of the Lagrangian function $L(\mathbf{x}, \lambda)$.

$$L(\mathbf{x}, \lambda) = 1 - \mathbf{x}^T\mathbf{x} + \lambda(1 - \mathbf{a}^T\mathbf{x})$$

$$\text{Where } \mathbf{a} = \begin{bmatrix}1\\1\end{bmatrix}, \mathbf{x} = \begin{bmatrix}x_1\\x_2\end{bmatrix}, \mathbf{0} = \begin{bmatrix}0\\0\end{bmatrix}, f(\mathbf{x}) = 1 - \mathbf{x}^T\mathbf{x}, g(\mathbf{x}) = 1 - \mathbf{a}^T\mathbf{x}$$

$$\nabla_{\mathbf{x}} f = -2\mathbf{x} = \mathbf{0} \rightarrow \mathbf{x} = \mathbf{0}$$

$$\text{We check whether the optima at } \mathbf{x} = \mathbf{0} \text{ satisfies the constraint } g(\mathbf{x}) = 1 - \mathbf{a}^T\mathbf{x} \geq 0$$

$$g(\mathbf{x} = \mathbf{0}) = 1 - \mathbf{a}^T\mathbf{0} = 1 > 0$$

$$g(\mathbf{x}) \text{ is satisfied and therefore } \lambda = 0\,, \mathbf{x} = \mathbf{0}$$

$$\mathbf{x_1} = \mathbf{x_2} = \mathbf{0}\,, \lambda = \mathbf{0}$$

4.

$$L(\mathbf{x}, \lambda) = \mathbf{a}^T\mathbf{x} + \lambda(\mathbf{x}^T\mathbf{x} - 1)$$

$$\text{Where } \mathbf{a} = \begin{bmatrix}1\\2\\-2\end{bmatrix}, \mathbf{x} = \begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}, \mathbf{0} = \begin{bmatrix}0\\0\end{bmatrix}, f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}\,, g(\mathbf{x}) = \mathbf{x}^T\mathbf{x} - 1$$

$$\nabla_{\mathbf{x}} L = \mathbf{a} + 2\lambda\mathbf{x} = \mathbf{0} \rightarrow \mathbf{a} = -2\lambda\mathbf{x} \rightarrow \mathbf{x} = (\frac{-1}{2\lambda}) \cdot \mathbf{a} \rightarrow \mathbf{x} = \begin{bmatrix}\frac{-1}{2\lambda}\\ \frac{-1}{\lambda}\\ \frac{1}{\lambda}\end{bmatrix} \text{ Where: } \lambda \neq 0$$

$$\nabla_\lambda L = \mathbf{x}^T\mathbf{x} - 1 = 0 \rightarrow \mathbf{x}^T\mathbf{x} = 1 \rightarrow \mathbf{x} = (\frac{-1}{2\lambda})^2 + (\frac{-1}{\lambda})^2 + (\frac{1}{\lambda})^2 = 1 \rightarrow \frac{9}{4\lambda^2} = 1$$

$$\lambda^2 = \frac{9}{4} \rightarrow \lambda = \pm\frac{3}{2}$$

We must evaluate both values of λ to see which one leads us to the solution that satisfies the second order constraints of our Lagrangian function, namely that it must be negative (semi) definite for a (local) maximum to exist $\nabla_{\mathbf{xx}} L(\mathbf{x}, \lambda) \leq 0$.

$$\text{Consider } \lambda = \frac{3}{2}$$

$$\mathbf{x} = \begin{bmatrix}\frac{-1}{3}\\ \frac{-2}{3}\\ \frac{2}{3}\end{bmatrix}$$

$$\nabla_{\mathbf{xx}} L(\mathbf{x}, \lambda) = 2\lambda > 0$$

The Lagrangian function is not concave in at this particular value of $\mathbf{x}$. Therefore $\lambda = \frac{3}{2}$ does not provide a maximum.

$$\text{Consider } \lambda = \frac{-3}{2}$$

$$\mathbf{x} = \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{-2}{3} \end{bmatrix}$$

$$\nabla_{\mathbf{xx}} L(\mathbf{x}, \lambda) = 2\lambda < 0$$

The Lagrangian function is concave in at this particular value of $\mathbf{x}$. Therefore $\lambda = \frac{-3}{2}$ provides a maximum. Therefore we have our maximum at

$$\mathbf{x} = \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{-2}{3} \end{bmatrix}$$

5. We observe that the following constraints are present given the nature of the problem and they define the domain of $f(x, y)$, but they are not needed directly in the Lagrangian function. $x \geq 0, y \geq 0$.

$$f(x, y) = 6x^{\frac{2}{3}} y^{\frac{1}{2}}$$

$$g(x, y) = 4x + 3y - 7000 \leq 0$$

$$L(x, y, \lambda) = 6x^{\frac{2}{3}} y^{\frac{1}{2}} + \lambda(4x + 3y - 7000)$$

$$\nabla_x f = 6 \cdot y^{\frac{1}{2}} \cdot \frac{2}{3} \cdot x^{\frac{-1}{3}} = 4 \cdot y^{\frac{1}{2}} \cdot x^{\frac{-1}{3}}$$

$$\nabla_y f = 6 \cdot x^{\frac{2}{3}} \cdot \frac{1}{2} \cdot y^{\frac{-1}{2}} = 3x^{\frac{2}{3}} \cdot y^{\frac{-1}{2}}$$

We observe that both $\nabla_x L$ and $\nabla_y L$ are non-negative, therefore the function $f(x, y)$ is monotonically increasing. Therefore there is no global maximum, and we can conclude that our constraint function $g(x, y)$ is active. Therefore we must evaluate the Lagrangian at $g(x, y) = 0$.

$$\nabla_x L = 4 \cdot y^{\frac{1}{2}} \cdot x^{\frac{-1}{3}} + 4\lambda = 0 \rightarrow \lambda = -y^{\frac{1}{2}} \cdot x^{\frac{-1}{3}}$$

$$\nabla_y L = 3x^{\frac{2}{3}} \cdot y^{\frac{-1}{2}} + 3\lambda = 0 \rightarrow \lambda = -x^{\frac{2}{3}} \cdot y^{\frac{-1}{2}}$$

At optimum point x_*, y_*, we have:

$$x_*^{\frac{2}{3}} \cdot y_*^{\frac{-1}{2}} = y_*^{\frac{1}{2}} \cdot x_*^{\frac{-1}{3}} \rightarrow x_*^{\frac{2}{3}} \cdot x_*^{\frac{1}{3}} = y_*^{\frac{1}{2}} \cdot y_*^{\frac{1}{2}} \rightarrow x_* = y_*$$

Now we evaluate our constraint $g(x_*, y_*) = 0$.

$$4x_* + 3y_* - 7000 = 0$$

$$7x_* = 7y_* = 7000 \rightarrow \mathbf{x_*} = \mathbf{y_*} = \mathbf{1000}$$

Therefore we can now estimate the maximum number of dosages we can produce with a budget of 7000 euros.

$$D(x_* = 1000, y_* = 1000) = 6x_*^{\frac{2}{3}} y_*^{\frac{1}{2}} = 6x_*^{\frac{7}{6}} = 6000\sqrt{10}$$

2 Kernel Outlier Detection

1.

$$f(R^2, \xi|\mathbf{C}) = R^2 + C\sum_{i=1}^{N}\xi_i$$

$$g(R^2, \mathbf{a}, \boldsymbol{\alpha}, \xi, \boldsymbol{\mu}|\mathbf{X}) = \sum_{i}^{N}\alpha_i\left(R^2 + \xi_i - \|\mathbf{x}_i - \mathbf{a}\|^2\right) + \sum_{i}^{N}\mu_i\xi_i$$

$$L(R^2, \mathbf{a}, \boldsymbol{\alpha}, \xi, \boldsymbol{\mu}|\mathbf{X}, \mathbf{C}) = R^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i}^{N}\alpha_i\left(R^2 + \xi_i - \|\mathbf{x}_i - \mathbf{a}\|^2\right) - \sum_{i}^{N}\mu_i\xi_i$$

2. We write the first-order, second-order and KKT constraints below.

$$\nabla_{R^2}L = 0 \to 1 - \sum_{i}^{N}\alpha_i = 0 \to \sum_{i}^{N}\alpha_i = 1$$

$$\nabla_{\xi_i}L = 0 \to 0 + C + \alpha_i - \mu_i = 0 \to C = \alpha_i + \mu_i$$

$$\nabla_{\mathbf{a}}L = \sum_{i}^{N}\alpha_i(\mathbf{a} - \mathbf{x}_i) = 0$$

$$\alpha_i \geq 0\,, \mu_i \geq 0$$

$$\alpha_i\left(R^2 + \xi_i - \|\mathbf{x}_i - \mathbf{a}\|^2\right) = 0$$

$$\mu_i\xi_i = 0$$

$$R^2 + \xi_i - \|\mathbf{x}_i - \mathbf{a}\|^2 \geq 0$$

$$\xi_i \geq 0$$

$$\nabla_{R^2R^2}L = 0 \geq 0 \text{ satisfied}$$

3. For $\forall i = 1, \cdots, N$

$$\alpha_i\left(R^2 + \xi_i - \|\mathbf{x}_i - \mathbf{a}\|^2\right) = 0$$

$$\mu_i\xi_i = 0$$

$$\text{Consider } \alpha_i > 0 \to R^2 + \xi_i - \|\mathbf{x}_i - \mathbf{a}\|^2 = 0$$

$$\text{Consider } \mu_i > 0 \to \xi_i = 0$$

no.	α	μ	Interpretation
1	$\alpha_i > 0$	$\mu_i = 0$	$\begin{cases}\xi_i = 0 \to \text{Lies on the boundary} \\ \xi_i > 0 \to \text{Outside the region}\end{cases}$
2	$\alpha_i > 0$	$\mu_i > 0$	$\xi_i = 0 \to$ Lies on the boundary
3	$\alpha_i = 0$	$\mu_i = 0$	Impossible assuming $C > 0$
4	$\alpha_i = 0$	$\mu_i > 0$	$\xi_i = 0 \to$ Inside the region

Note that support vectors are active when $\alpha_i > 0$

4. We define $\mathbf{k}(.)$ as a kernel function.

$$\begin{aligned}
L(R^2, \mathbf{a}, \boldsymbol{\alpha}, \xi, \boldsymbol{\mu}) &= R^2 + C\sum_i \xi_i - \sum_i \alpha_i \left(R^2 + \xi_i - \|\mathbf{x}_i - \mathbf{a}\|^2\right) - \sum_i \mu_i \xi_i \\
&= \cancel{R2} + C\cancel{\sum_i \xi_i} - \cancel{\sum_i \alpha_i R^2} - \cancel{\sum_i \alpha_i \xi_i} + \sum_i \alpha_i \|\mathbf{x}_i - \mathbf{a}\|^2 - \cancel{\sum_i \mu_i \xi_i} \\
&= \sum_i \alpha_i \|\mathbf{x}_i - \mathbf{a}\|^2 \\
&= \sum_i \alpha_i (\mathbf{x}_i - \mathbf{a})^T (\mathbf{x}_i - \mathbf{a}) \\
&= \sum_i \alpha_i (\mathbf{x}_i^T \mathbf{x}_i - 2\mathbf{x}_i^T \mathbf{a} + \mathbf{a}^T \mathbf{a}) \\
&= \sum_i \alpha_i (\mathbf{x}_i^T \mathbf{x}_i - 2\mathbf{x}_i^T \mathbf{a} + \mathbf{a}^T \mathbf{a}) \\
&= \sum_i \alpha_i \mathbf{x}_i^T \mathbf{x}_i - 2\sum_i \alpha_i \mathbf{x}_i^T \mathbf{a} + \sum_i \alpha_i \mathbf{a}^T \mathbf{a} \\
&= \sum_i \alpha_i \mathbf{x}_i^T \mathbf{x}_i - 2\sum_i \alpha_i \mathbf{x}_i^T \left(\sum_j \alpha_j \mathbf{x}_j\right) + \sum_i \alpha_i \left(\sum_j \alpha_j \mathbf{x}_j\right)^T \left(\sum_j \alpha_j \mathbf{x}_j\right) \\
&= \sum_i \alpha_i \mathbf{x}_i^T \mathbf{x}_i - 2\sum_i \sum_j \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j + \sum_i \sum_j \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j \\
&= \sum_i \alpha_i \mathbf{x}_i^T \mathbf{x}_i - \sum_i \sum_j \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j \\
\tilde{L}(\boldsymbol{\alpha}) &= \sum_i \alpha_i \mathbf{k}(\mathbf{x}_i, \mathbf{x}_i) - \sum_i \sum_j \alpha_i \alpha_j \mathbf{k}(\mathbf{x}_i, \mathbf{x}_j)
\end{aligned}$$

The dual minimization problem is:

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}} \sum_i \alpha_i \mathbf{k}(\mathbf{x}_i, \mathbf{x}_i) - \sum_i \sum_j \alpha_i \alpha_j \mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) \\
&\qquad\qquad \text{s.t.} \sum_i \alpha_i = 1 \qquad \forall i = 1, \cdots, N
\end{aligned}$$

5. From the dual minimization problem we can obtain $\mathbf{a}^*$. We use it to derive the following. Note that to calculate R_*^2, we only need the support vectors lying exactly on the boundary, therefore we only use $\mathbf{x}_i$ where $\alpha_i^* > 0\,, \mu_i^* > 0\,, \xi_i^* = 0$. ξ_i^* is the maximum of 0 and its value when $\alpha*_i > 0$.

$$\begin{aligned}
\mu_i^* &= C - \alpha_i^* \\
\mathbf{a}^* &= \sum_i \alpha_i^* \mathbf{x}_i \\
R_*^2 &= \|\mathbf{x}_i - \mathbf{a}^*\|^2 = \left\| \mathbf{x}_i - \sum_i \alpha_i^* \mathbf{x}_i \right\|^2
\end{aligned}$$

We know $\xi_i^* \mu_i = 0 \rightarrow \xi_i^* = 0 \quad or \quad \xi_i^* > 0$

if $\xi_i^* > 0 \rightarrow C = \alpha_i^*$

$\alpha_i^* > 0 \rightarrow R^2 + \xi_i^* - \|\mathbf{x}_i - \mathbf{a}^*\|^2 = 0 \rightarrow \xi_i^* = -R^2 + \|\mathbf{x}_i - \mathbf{a}^*\|^2$ Outside the boundary

6. If a data point is an outlier, then case 1 from the table in question 3 applies, therefore $\xi_i = 0, \alpha_i > 0, \mu_i = 0$. We can then devise a simple test where if the squared distance of a point $\mathbf{z}$ from the center $\mathbf{a}$ is larger than R^2 then it is an outlier. $\mathbf{x}_i$ is any vector that lies on the boundary, i.e. a support vector.

$$\begin{aligned}
\|\mathbf{z} - \mathbf{a}^*\|^2 - R_*^2 &> 0 \\
(\mathbf{z} - \mathbf{a}^*)^T(\mathbf{z} - \mathbf{a}^*) - (\mathbf{x}_i - \mathbf{a}^*)^T(\mathbf{x}_i - \mathbf{a}^*) &> 0 \\
(\mathbf{z}^T\mathbf{z} - 2\mathbf{z}^T\mathbf{a}^* + \cancel{(\mathbf{a}^*)^T\mathbf{a}^*}) - (\mathbf{x}_i^T\mathbf{x}_i - 2\mathbf{x}_i^T\mathbf{a}^* + \cancel{(\mathbf{a}^*)^T\mathbf{a}^*})) &> 0 \\
\mathbf{k}(\mathbf{z}, \mathbf{z}) - 2\mathbf{k}(\mathbf{z}, \mathbf{a}^*) - \mathbf{k}(\mathbf{x}_i, \mathbf{x}_i) + 2\mathbf{k}(\mathbf{x}_i, \mathbf{a}^*) &> 0 \\
\mathbf{k}(\mathbf{z}, \mathbf{z}) - \mathbf{k}(\mathbf{x}_i, \mathbf{x}_i) - 2\sum_j \alpha_j^*(\mathbf{k}(\mathbf{z}, \mathbf{x}_j) - \mathbf{k}(\mathbf{x}_i, \mathbf{x}_j)) &> 0
\end{aligned}$$

7. If $C \to 0$, then we have $\alpha_i + \mu_i = 0$. In effect, our minimization problem is unconstrained and as a result the value of R^2 is greatly minimized. There is no penalization in this settings and in an extreme case most or all points will be treated as outliers as the boundary effectively doesn't exist outside the origin. If $C \to \infty$ then the minimization procedure will focus on shrinking ξ_i. The penalization in this case in infinite and as a results all points will be inliers. As the value of R tries to compensate for the penalization, it will become the maximum distance of any point from the origin. This also implies that the origin will be placed at the center of two points with the largest distance.

8. An RBF kernel with a small sigma produces a very tight fit on the data and as a result the boundary will fit exactly to the data. There will be many support vectors, as the kernel evaluated at any two points is close to one if those two points are almost the same. The kernel value will be close to zero for any two points whose distance is a very small value close to zero. The boundary will be non-spherical as opposed to a Linear Kernel which has a spherical boundary due to the euclidean norm. This RBF kernel will overfit and adding extra data points can cause extreme changes in the boundary.

   An RBF kernel with a large sigma will allow a degree of flexibility in fitting a boundary on the data, and will look more spherical. Note that the hyperparameter C also plays a role in the boundary, as it controls the size of the boundary. The sigma dictates the shape of the boundary.

9. We have $y = 1$ when a datapoint is an outlier and $y = -1$ when it is an inlier or on the decision boundary. Therefore we can reformulate our findings into

$$\min_{\mathbf{a}, R, \xi} R^2 + C \sum_i \xi_i$$

$$\text{s.t } y_i \left(\|\mathbf{x}_i - \mathbf{a}\|^2 - R^2\right) + \xi_i \geq 0, \quad \xi_i \geq 0 \ \forall i = 1, \cdots, N.$$