Kernelizing the Minimum Volume Ellipsoid

October 1, 2007

Suppose the ellipsoid around the data $(\mathbf{x}_i)_{i=1}^n$ with $\mathbf{x} \in \mathbb{R}^n$ is characterized by $(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) \leq 1$, where $\mu \in \mathbb{R}^n$ and $\mathbf{\Sigma}^1 \in \mathbb{R}^{n\times n}$ is symmetric and positive definite (denoted as $\mathbf{\Sigma} \succeq 0$), the expression for points within the ellipsoid can be written as follows:

$$\begin{aligned}(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) &= (\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{x}-\mu) \\ &= (\mathbf{A}\mathbf{x}-\mathbf{b})^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) \leq 1\end{aligned}$$

where $\mathbf{A} = \mathbf{\Sigma}^{-\frac{1}{2}}$ and $\mathbf{b} = \mathbf{\Sigma}^{-\frac{1}{2}}\mu$. Thus, the problem of finding the minimum volume ellipsoid around $(\mathbf{x}_i)_{i=1}^n$ can be expressed as:

$$\begin{aligned}\min_{\mathbf{A},\mathbf{b}} \quad & -\ln|\mathbf{A}| \\ \text{s.t.} \quad & \|\mathbf{A}\mathbf{x}_i - \mathbf{b}\|^2 \leq 1 \quad \forall 1 \leq i \leq n, \quad \mathbf{A} \succeq 0.\end{aligned}$$

where, $|\mathbf{A}|$ denotes the determinant of the matrix $\mathbf{A}$. The smallest enclosing ellipsoid is the one that has the smallest volume yet encloses all the given points. We measure volume of an ellipsoid via the determinant of $\mathbf{\Sigma}$. Instead of explicitly minimizing the determinant of $\mathbf{\Sigma}$, we equivalently maximize the determinant of the inversely-related matrix $\mathbf{A}$ since $\mathbf{A} = \mathbf{\Sigma}^{-\frac{1}{2}}$. After solving (1) both μ and $\mathbf{\Sigma}$ can be recovered using matrix inversion and simple algebra.

In practice, however, real world data has measurement noise and outliers so the bounding ellipsoid above is not necessarily the most reliable way to estimate the ellipsoidal gap-tolerant classifier. The following relaxed version of the above formulation is thus used in practice[2]:

$$\min_{\mathbf{A},\mathbf{b},\tau} \; -\ln|\mathbf{A}| + E\sum_{i=1}^n \tau_i \tag{1a}$$

$$\text{s.t. } \|\mathbf{A}\mathbf{x}_i - \mathbf{b}\|^2 \leq 1 + \tau_i, \; \tau_i \geq 0 \quad \mathbf{A} \succeq 0 \tag{1b}$$

where τ_i is a penalty on the samples that remain outside the ellipsoid, E is a parameter that trades off the volume of the ellipsoid with the penalty on the

[1] We assume that $\mathbf{\Sigma}$ is of full rank, it is possible to handle the non-full rank $\mathbf{\Sigma}$ with simple modifications.

[2] Sometimes we drop $0 \leq i \leq n$ from the constraint when it is clear

outliers. The quadratic constraint in (1b) can equivalently be expressed as the following semidefinite constraint:

$$\begin{bmatrix} I & (\mathbf{A}\mathbf{x}_i + b) \\ (\mathbf{A}\mathbf{x}_i + b)^\top & 1 + \tau_i \end{bmatrix} \succeq \mathbf{0}.$$

Thus the formulation (1) is an instance of a Semidefinite Programming (SDP) which can be efficiently solved in polynomial time.

Let us now add an extra dimension to the data, that is, $\mathbf{z}_i^\top \leftarrow [\mathbf{x}_i^\top 1]$ and find the ellipsoid enclosing these new points with the origin as the center. By taking the projection of this new ellipsoid with the hyperplane $z_{n+1} = 1$, we should be able to recover the minimum volume ellipsoid for the original problem. Now consider the following formulation:

$$\begin{aligned} &\min_{\mathbf{M},\tau} -\ln|\mathbf{M}| + E\sum_{i=1}^{n}\tau_i \\ &\text{s.t. } \mathbf{z}_i^\top \mathbf{M}\mathbf{z}_i \le 1 + \tau_i, \quad \tau_i \ge 0 \;\; \forall 1 \le i \le n \qquad \mathbf{M} \succeq 0. \end{aligned} \tag{2}$$

where $\mathbf{M} \in \mathbb{R}^{m+1\times m+1}$. Write the above optimization as a Lagrangian:

$$L = -\ln|\mathbf{M}| + E\sum_{i=1}^{n}\tau_i + \sum_{i=1}^{n}\gamma_i(\mathbf{z}_i^\top \mathbf{M}\mathbf{z}_i - 1 - \tau_i) - \sum_{i=1}^{n}\beta_i\tau_i$$

with $\beta_i, \gamma_i \ge 0$. Taking the gradient with respect to $\mathbf{M}$, we get

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{M}} &= -\mathbf{M}^{-1} + \sum_i \gamma_i \mathbf{z}_i \mathbf{z}_i^\top \\ \frac{\partial L}{\partial \tau_i} &= E - \gamma_i - \beta_i \end{aligned}$$

Equating the above to zero, we get $\mathbf{M}^{-1} = \sum_i \gamma_i \mathbf{z}_i \mathbf{z}_i^\top$. Now consider,

$$\begin{aligned} \sum_{i=1}^{n}\gamma_i \mathbf{z}_i^\top \mathbf{M}\mathbf{z}_i &= \sum_{i=1}^{n}\gamma_i tr(\mathbf{z}_i^\top \mathbf{M}\mathbf{z}_i) \\ &= \sum_{i=1}^{n}\gamma_i tr(\mathbf{M}\mathbf{z}_i\mathbf{z}_i^\top) \\ &= tr(\mathbf{M}\sum_{i=1}^{n}\gamma_i \mathbf{z}_i\mathbf{z}_i^\top) \\ &= tr(\mathbf{I}) = m + 1. \end{aligned}$$

Substituting these results obtained from equating the partial derivatives to zero back in the Lagrangean, we get the dual optimization problem:

$$\begin{aligned} &\max_{\gamma_i} \ln|\sum_{i=1}^{n}\gamma_i \mathbf{z}_i\mathbf{z}_i^\top| - \sum_{i=1}^{n}\gamma_i + m + 1 \\ &\text{s.t. } 0 \le \gamma_i \le E \qquad\qquad \forall i \le i \le n. \end{aligned} \tag{3}$$

We note that $m+1$ from the objective can be removed without changing the values of γ_i at the solution. Now let us retrieve $\mathbf{\Sigma}$ and μ from γ. We know that,

$$\mathbf{M}^{-1} = \sum_i \gamma_i \mathbf{z}_i \mathbf{z}_i^\top = \begin{bmatrix} \mathbf{X\Gamma X}^\top & \mathbf{X}\gamma \\ \gamma^\top \mathbf{X}^\top & \gamma^\top \mathbf{1} \end{bmatrix}$$

where $\mathbf{\Gamma}$ is a diagonal matrix containing γ_i as the diagonal entries and γ is a vector containing γ_i as the entries. We can rewrite $\mathbf{M}^{-1}$ as

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{X\Gamma X}^\top & \mathbf{X}\gamma \\ \gamma^\top \mathbf{X}^\top & \gamma^\top \mathbf{1} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \frac{\mathbf{X}\gamma}{s} \\ 0 & s \end{bmatrix} \begin{bmatrix} \mathbf{X\Gamma X}^\top - \frac{\mathbf{X}\gamma\gamma^\top \mathbf{X}^\top}{s^2} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I} & 0 \\ \frac{\gamma^\top \mathbf{X}}{s}^\top & s \end{bmatrix}$$

where $s = \gamma^\top \mathbf{1}$. Inverting on both the sides,

$$\mathbf{M} = \begin{bmatrix} \mathbf{I} & 0 \\ -\frac{\gamma^\top \mathbf{X}}{s^2}^\top & \frac{1}{s} \end{bmatrix} \begin{bmatrix} (\mathbf{X\Gamma X}^\top - \frac{\mathbf{X}\gamma\gamma^\top \mathbf{X}^\top}{s^2})^{-1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\frac{\mathbf{X}\gamma}{s^2} \\ 0 & \frac{1}{s} \end{bmatrix}$$

Substituting $\mathbf{M}$ in $\mathbf{z}_i^\top \mathbf{M} \mathbf{z}_i \leq 1$ and identifying the resultant equation with (1a), it can be shown that:

$$\mu = \frac{\mathbf{X}\gamma}{\gamma^\top \mathbf{1}}$$

and,

$$\mathbf{\Sigma} = \mathbf{X\Gamma X}^\top - \frac{\mathbf{X}\gamma\gamma^\top \mathbf{X}^\top}{\gamma^\top \mathbf{1}}.$$

Now, to solve (3) in kernel defined feature space, we merely make the following substitution:

$$\sum_{i=1}^{n} \gamma_i \mathbf{z}_i \mathbf{z}_i^\top = \begin{bmatrix} \mathbf{K}^{\frac{1}{2}} \mathbf{\Gamma} \mathbf{K}^{\frac{1}{2}} & \mathbf{K}^{\frac{1}{2}} \gamma \\ \gamma^\top \mathbf{K}^{\frac{1}{2}} & \gamma^\top \mathbf{1} \end{bmatrix}.$$

To calculate $\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}$ assuming $\mathbf{w} = \mathbf{X}\alpha$:

$$\begin{aligned} \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w} &= \alpha^\top \mathbf{X}^\top (\mathbf{X\Gamma X}^\top - \frac{\mathbf{X}\gamma\gamma^\top \mathbf{X}^\top}{\gamma^\top \mathbf{1}}) \mathbf{X}\alpha \\ &= \alpha^\top (\mathbf{K\Gamma K} - \frac{\mathbf{K}\gamma\gamma^\top \mathbf{K}}{\gamma^\top \mathbf{1}}) \alpha. \end{aligned}$$

A simpler (and mathematically more sound) option is to centralize the data in Hilbert space around the origin before solving the MVE (1).