CS613 \ Assignment #3 Solutions

Ariel Stolerman

1)

1.1)

If we use the slack variables, the classification constraints will become:

$$t_n y(\mathrm{x}_n) \geq 1 - \xi_n$$

For $n = 1, \dots, N$, where the slack variables are constrained to satisfy $\xi_n \geq 0$.

1.2)

The error function that we ought to minimize becomes:

$$C \sum_{n=1}^{N} \xi_n + \frac{1}{2} \|\mathrm{w}\|^2$$

Where $C > 0$ is intended to control the tradeoff between the slack variable penalty and the margin (when $C \to \infty$, the above is reduced to SVMs for separable data).

1.3)

Using the above, the corresponding Lagrangian is:

$$L(\mathrm{w}, b, \mathrm{a}) = \frac{1}{2} \|\mathrm{w}\|^2 + C \sum_{n=1}^{N} \xi_n - \sum_{n=1}^{N} a_n \{t_n y(x_n) - 1 + \xi_n\} - \sum_{n=1}^{N} \mu_n \xi_n$$

Where $\{a_n \geq 0\}_{n=1}^{N}$ are the Lagrange multipliers for the classification constraints and $\{a_n \geq 0\}_{n=1}^{N}$ – for the non-negativity constraint on $\{\xi_n\}_{n=1}^{N}$.

Deriving the quadratic error function and its constraints (with KKT conditions) that needs to be minimized to compute $\{a_n\}$:

The set of KKT conditions are:

$$\begin{cases} a_n \geq 0 \\ t_n y(\mathrm{x}_n) - 1 + \xi_n \geq 0 \\ a_n (t_n y(\mathrm{x}_n) - 1 + \xi_n) = 0 \end{cases}$$

$$\begin{cases} \mu_n \geq 0 \\ \xi_n \geq 0 \\ \mu_n \xi_n = 0 \end{cases}$$

For $n = 1, \dots, N$. Recall that $y(\mathrm{x}) = \mathrm{w}^T \phi(\mathrm{x}) + b$, we now optimize out w, b and $\{\xi_n\}$:

The derivative of L w.r.t. w and b is not different than the case without overlapping class distributions, and so we can immediately determine:

$$\frac{\partial L}{\partial \mathrm{w}} = 0 \Rightarrow \mathrm{w} = \boxed{\sum_{n=1}^{N} a_n t_n \phi(\mathrm{x}_n)} \quad (1)$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \underbrace{\boxed{\sum_{n=1}^{N} a_n t_n = 0}}_{(2)}$$

Now for ξ_n:

$$\frac{\partial L}{\partial \xi_n} = \frac{\partial (C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N} a_n\xi_n - \sum_{n=1}^{N}\mu_n\xi_n)}{\partial \xi_n} = \frac{\partial (C\xi_n - a_n\xi_n - \mu_n\xi_n)}{\partial \xi_n} = 0 \Leftrightarrow$$

$$C - a_n - \mu_n = 0 \Leftrightarrow \underbrace{\boxed{a_n = C - \mu_n}}_{(3)}$$

Recall that $k(\mathrm{x}, \mathrm{x}') = \phi(\mathrm{x})^T\phi(\mathrm{x}')$. Using the results above we obtain the dual Lagrangian form:

$$\tilde{L}(\mathrm{a}) = \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N} a_n\{t_n y(x_n) - 1 + \xi_n\} - \sum_{n=1}^{N}\mu_n\xi_n =$$

$$\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m \phi(\mathrm{x}_n)^T\phi(\mathrm{x}_m) + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m \phi(\mathrm{x}_n)^T\phi(\mathrm{x}_m) - \underbrace{\sum_{n=1}^{N} a_n t_n b}_{(2)\Rightarrow=0} + \sum_{n=1}^{N} a_n -$$

$$-\sum_{n=1}^{N} a_n\xi_n - \sum_{n=1}^{N}\mu_n\xi_n = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathrm{x}_n, \mathrm{x}_m) + \sum_{n=1}^{N}\xi_n\left(\underbrace{C - \mu_n - a_n}_{(3)\Rightarrow=0}\right) + \sum_{n=1}^{N} a_n =$$

$$\boxed{\sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathrm{x}_n, \mathrm{x}_m)}$$

With respect to $\{a_n\}$ subject to the constraints:

$$\begin{cases} 0 \le a_n \le C, \forall n = 1, \dots, N \\ \sum_{n=1}^{N} a_n t_n = 0 \end{cases}$$

2)

Let $x_1 \in C_1$ ($t_1 = +1$) and $x_2 \in C_2$ ($t_2 = -1$) be the only two data points from each of the two classes. Following is a proof that irrespective to dimensionality this is sufficient to determine the location of the maximum-margin hyperplane.

We are given:

$$y(x_1) = w^T x_1 + b = +1$$
$$y(x_2) = w^T x_2 + b = -1$$

Looking at the dual problem we need to find:

$$\max_a \tilde{L}(a) = \max_a \left(\sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m x_n^T x_m \right)$$

For $N = 2$ and subject to the constraints $a_i \geq 0$ and $\sum_{n=1}^{N} a_n t_n = 0 \Rightarrow$

$$\begin{cases} a_1 - a_2 = 0 & (A) \\ a_1, a_2 \geq 0 & (B) \end{cases}$$

$$\tilde{L}(a) = a_1 + a_2 - \frac{1}{2}\left(a_1^2 t_1^2 x_1^T x_1 + \underbrace{a_1 a_2 t_1 t_2 x_1^T x_2 + a_2 a_1 t_2 t_1 x_2^T x_1}_{equal} + a_2^2 t_2^2 x_2^T x_2 \right) =$$

$$a_1 + a_2 - \frac{1}{2}(a_1^2 x_1^T x_1 - 2a_1 a_2 x_1^T x_2 + a_2^2 x_2^T x_2) = a_1 + a_2 - \frac{1}{2} a_1^2 x_1^T x_1 + a_1 a_2 x_1^T x_2 - \frac{1}{2} a_2^2 x_2^T x_2$$

Denote the kernel: $x_i^T x_j = k(x_i, x_j) = k_{ij}$:

$$\tilde{L}(a) = a_1 + a_2 - \frac{1}{2} a_1^2 k_{11} + a_1 a_2 k_{12} - \frac{1}{2} a_2^2 k_{22}$$

To maximize $\tilde{L}(a)$ we find partial derivatives:

$$\frac{\partial \tilde{L}(a)}{\partial a_1} = 1 - a_1 k_{11} + a_2 k_{12}$$

$$\frac{\partial \tilde{L}(a)}{\partial a_2} = 1 - a_2 k_{22} + a_1 k_{12}$$

Comparing both derivatives to zero and subtracting the first equation from the second:

$$1 - a_2 k_{22} + a_1 k_{12} - (1 - a_1 k_{11} + a_2 k_{12}) = 0 \Leftrightarrow 1 - a_2 k_{22} + a_1 k_{12} - 1 + a_1 k_{11} - a_2 k_{12} = 0 \Leftrightarrow$$

$$a_1 k_{11} - a_2 k_{22} + k_{12} \underbrace{(a_1 - a_2)}_{(A) \Rightarrow 0=} = 0 \Leftrightarrow a_1 = a_2 \frac{k_{22}}{k_{11}}$$

Assigning in the second equation:

$$1 - a_2 k_{22} + a_2 \frac{k_{22}}{k_{11}} k_{12} = 0 \Leftrightarrow a_2 \frac{k_{22} k_{12} - k_{22} k_{11}}{k_{11}} = -1 \Leftrightarrow a_2 = -\frac{1}{k_{22}} \cdot \frac{k_{11}}{k_{12} - k_{11}} \Rightarrow \boxed{a_2 = \frac{1}{k_{22}} \cdot \frac{k_{11}}{k_{11} - k_{12}}}$$

$$\Rightarrow a_1 = \frac{1}{k_{22}} \cdot \frac{k_{11}}{k_{11} - k_{12}} \cdot \frac{k_{22}}{k_{11}} \Rightarrow \boxed{a_1 = \frac{1}{k_{11} - k_{12}}}$$

Taking second derivatives to assure maximum:

$$\frac{\partial^2 \tilde{L}(a)}{\partial a_1^2} = -k_{11} < 0$$

Since $k_{11} = x_1^T x_1 = \|x_1\|^2 > 0$ (or simply because the kernel is a positive-definite function). In the same manner:

$$\frac{\partial^2 \tilde{L}(a)}{\partial a_2^2} = -k_{22} < 0$$

Therefore both a_1, a_2 found are maxima. Now all is left to do is to assign a_1, a_2 to w and b:

$$w = \sum_{n=1}^{N} a_n t_n \phi(x_n) = \frac{1}{k_{11} - k_{12}} x_1 - \frac{1}{k_{22}} \cdot \frac{k_{11}}{k_{11} - k_{12}} x_2 \Rightarrow \boxed{w = \frac{1}{k_{11} - k_{12}} \left(x_1 - \frac{k_{11}}{k_{22}} x_2 \right)}$$

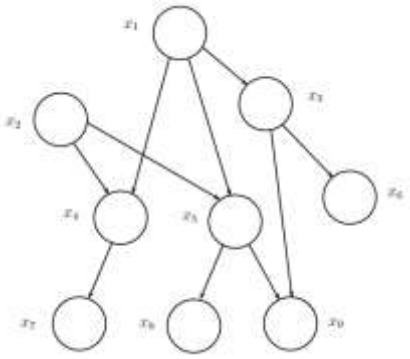
$$b = \frac{1}{N_S} \sum_{n \in S} \left(t_n - \sum_{m \in S} a_m t_m k_{nm} \right) \underset{\substack{x_1, x_2 \text{ are} \\ \text{support} \\ \text{vectors}}}{\Longrightarrow} b = \frac{1}{2} \left(1 - (a_1 k_{11} - a_2 k_{12}) - 1 - (a_1 k_{21} - a_2 k_{22}) \right) =$$

$$= \frac{1}{2} \left(-a_1 k_{11} + \underbrace{a_2 k_{12} - a_1 k_{21}}_{(A) \Rightarrow 0 =} + a_2 k_{22} \right) = \frac{1}{2} \left(\frac{1}{k_{22}} \cdot \frac{k_{11}}{k_{11} - k_{12}} k_{22} - \frac{1}{k_{11} - k_{12}} k_{11} \right) \Rightarrow$$

$$= \frac{1}{2(k_{11} - k_{12})} (k_{11} - k_{11}) \Rightarrow \boxed{b = 0}$$

Therefore we have found the location of the maximum-margin hyperplane with only two data points, and it is irrespective of the dimensionality of the data space (used some general dimension x_1, x_2).

3)



3.1)

The joint probability distribution for the given graphical model:

$$p(x_1)p(x_2)p(x_3|x_1)p(x_4|x_2,x_1)p(x_5|x_2,x_1)p(x_6|x_3)p(x_7|x_4)p(x_8|x_5)p(x_9|x_3,x_5)$$

3.2)

Checking whether the conditional independence holds:

a. $x_2 \perp\!\!\!\perp x_9 | x_3$ when x_7 observed:

The path $x_2 \rightarrow x_5 \rightarrow x_9$ is unblocked, since x_5 is the only intermediate node on the path, it is head-to-tail but not an observed node. Therefore the given conditional independence does **not** hold.

b. $x_6 \perp\!\!\!\perp x_7 | x_5$ when x_1 observed:

All paths from x_6 to x_7:

- $x_6 \rightarrow x_3 \rightarrow x_1 \rightarrow x_4 \rightarrow x_7$: x_1 is tail-to-tail and observed, so it is blocking the path.
- $x_6 \rightarrow x_3 \rightarrow x_1 \rightarrow x_5 \rightarrow x_2 \rightarrow x_4 \rightarrow x_7$: x_1 is tail-to-tail and observed, so it is blocking the path.
- $x_6 \rightarrow x_3 \rightarrow x_9 \rightarrow x_5 \rightarrow x_1 \rightarrow x_4 \rightarrow x_7$: x_5 is head-to-tail and observed, so it is blocking the path (and x_1 also blocks it).
- $x_6 \rightarrow x_3 \rightarrow x_9 \rightarrow x_5 \rightarrow x_2 \rightarrow x_4 \rightarrow x_7$: x_5 is head-to tail and observed, so it is blocking the path.

Since all paths from x_6 to x_7 are blocked, the conditional independence holds.