Support vector machines

Guide to SVMs from R. Berwick MIT, “Village Idiot”
http://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf

Lagrange Multiplier Examples from Baxter Tyson Smith

https://www.engr.mun.ca/~baxter/Publications/LagrangeForSVMs.pdf

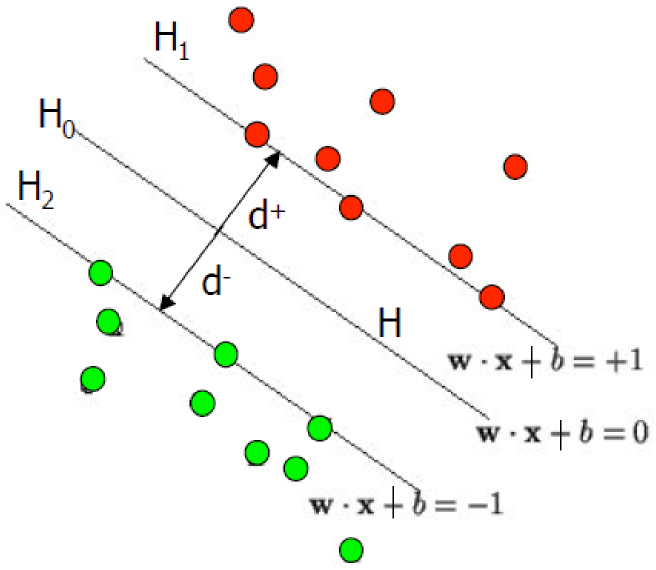
Definitions

Definitions

Define the hyperplanes H such that:

$w \bullet x_i + b \geq +1$ when $y_i = +1$

$w \bullet x_i + b \leq -1$ when $y_i = -1$



H_1 and H_2 are the planes:

H_1: $w \bullet x_i + b = +1$

H_2: $w \bullet x_i + b = -1$

The points on the planes H_1 and H_2 are the tips of the Support Vectors

The plane H_0 is the median in between, where $w \bullet x_i + b = 0$

d+ = the shortest distance to the closest positive point

d- = the shortest distance to the closest negative point

The margin (gutter) of a separating hyperplane is d+ + d–.

Defining the separating Hyperplane

- Form of equation defining the decision surface separating the classes is a hyperplane of the form:

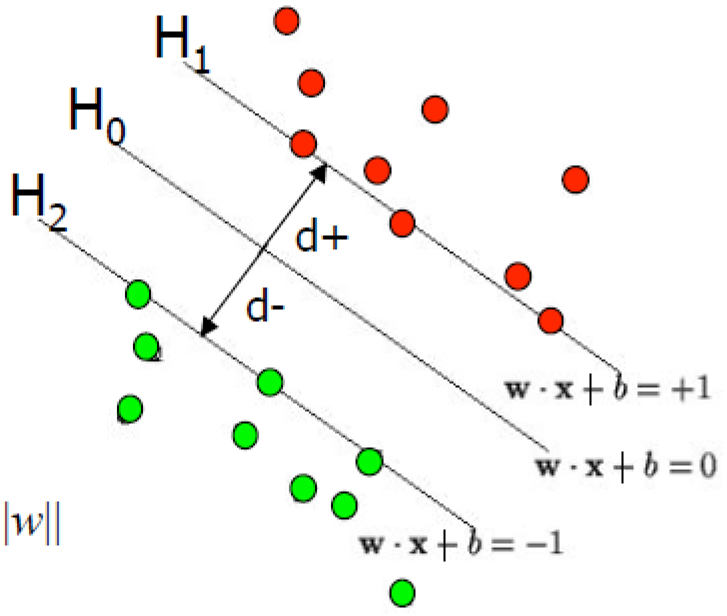
$$\mathbf{w}^T\mathbf{x} + b = 0$$

  - **w is a weight vector**
  - **x is input vector**
  - **b is bias**
- Allows us to write

$$\mathbf{w}^T\mathbf{x} + b \geq 0 \text{ for } d_i = +1$$
$$\mathbf{w}^T\mathbf{x} + b < 0 \text{ for } d_i = -1$$

Maximizing the margin (aka street width)

We want a classifier (linear separator)
with as big a margin as possible.



Recall the distance from a point(x_0,y_0) to a line:
$Ax+By+c = 0$ is: $|Ax_0 + By_0 + c|/\text{sqrt}(A^2+B^2)$, so,
The distance between H_0 and H_1 is then:
$|w \bullet x+b|/||w||=1/||w||$, so

The total distance between H_1 and H_2 is thus: $2/||w||$

In order to maximize the margin, we thus need to minimize $||w||$. With the condition that there are no datapoints between H_1 and H_2:

$\mathbf{x}_i \bullet \mathbf{w}+b \geq +1$ when $y_i = +1$
$\mathbf{x}_i \bullet \mathbf{w}+b \leq -1$ when $y_i = -1$

Can be combined into: $y_i(\mathbf{x}_i \bullet \mathbf{w}) \geq 1$

We now must solve a <u>quadratic</u> programming problem

- Problem is: <u>minimize</u> ||**w**||, **s.t.** discrimination boundary is obeyed, i.e., min $f(x)$ s.t. $g(x)=0$, which we can rewrite as:

  min f: ½ $||w||^2$ (Note this is a <u>quadratic</u> function)

  s.t. g: $y_{\mathbf{i}}(\mathbf{w} \bullet \mathbf{x}_{\mathrm{i}}) - \mathbf{b} = \mathbf{1}$ or $[y_{\mathbf{i}}(\mathbf{w} \bullet \mathbf{x}_{\mathrm{i}}) - \mathbf{b}] - \mathbf{1} = \mathbf{0}$

This is a **<u>constrained optimization problem</u>**

It can be solved by the Lagrangian multipler method

Because it is <u>quadratic</u>, the surface is a paraboloid, with just a single global minimum (thus avoiding a problem we had with neural nets!)

In general

Gradient min of f

constraint condition g

$L(x,a) = f(x) + \sum_i a_i g_i(x)$ a function of $n+m$ variables

n for the $x's$, m for the a. Differentiating gives $n+m$ equations, each set to 0. The n eqns differentiated wrt each x_i give the gradient conditions; the m eqns differentiated wrt each a_i recover the constraints g_i

In our case, $f(x)$: $\frac{1}{2}\|\mathbf{w}\|^2$; $g(x)$: $y_i(\mathbf{w}\cdot x_i + b) - 1 = 0$ so Lagrangian is:

$$min\ L = \tfrac{1}{2}\|\mathbf{w}\|^2 - \Sigma a_i[y_i(\mathbf{w}\cdot x_i + \mathrm{b}) - 1] \text{ wrt } \mathbf{w}, b$$

We expand the last to get the following L form:

$$min\ L = \tfrac{1}{2}\|\mathbf{w}\|^2 - \Sigma a_i y_i(\mathbf{w}\cdot x_i + \mathrm{b}) + \Sigma a_i \text{ wrt } \mathbf{w}, b$$

At the heart of Lagrange Multipliers is the following equation:

$$\nabla f(x) = \lambda \nabla g(x) \tag{1}$$

This says that the gradient of f is equal to some multiplier (lagrange multiplier) times the gradient of g. How this equation came about is explained in Section 6. Also, remember the form of g:

$$g(x) = 0 \tag{2}$$

Often, and especially in the context of SVMs, equations 1 and 2 are combined into one equation called the *Lagrangian*:

$$L(x, \lambda) = f(x) - \lambda g(x) \tag{3}$$

Using this equation, we look for points where:

$$\nabla L(x, \lambda) = 0 \tag{4}$$

Problem: Given,

$$
\begin{aligned}
f(x,y) &= 2 - x^2 - 2y^2 &\quad (5)\\
g(x,y) &= x + y - 1 = 0 &\quad (6)
\end{aligned}
$$

Find the extreme values.

Solution: First, we put the equations into the form of a Lagrangian:

$$\begin{aligned} L(x, y, \lambda) &= f(x, y) - \lambda g(x, y) &(7)\\ &= 2 - x^2 - 2y^2 - \lambda(x + y - 1) &(8) \end{aligned}$$

and we solve for the gradient of the Lagrangian (Equation 4):

$$\nabla L(x, y, \lambda) = \nabla f(x, y) - \lambda \nabla g(x, y) = 0 \tag{9}$$

which gives us:

$$\begin{aligned} \frac{\partial}{\partial x} L(x, y, \lambda) &= -2x - \lambda = 0 &(10)\\ \frac{\partial}{\partial y} L(x, y, \lambda) &= -4y - \lambda = 0 &(11)\\ \frac{\partial}{\partial \lambda} L(x, y, \lambda) &= x + y - 1 = 0 &(12) \end{aligned}$$

From Equation 10 and 11, we have $x = 2y$. Substituting this into Equation 12 gives $x = \frac{2}{3}$, $y = \frac{1}{3}$. These values give $\lambda = -\frac{4}{3}$ and $f = \frac{4}{3}$.

Problem: Given,

$$
\begin{aligned}
f(x, y) &= x^3 + y^3 && (57)\\
g_1(x, y) &= x^2 - 1 \geq 0 && (58)\\
g_2(x, y) &= y^2 - 1 \leq 0 && (59)
\end{aligned}
$$

Find the extreme values.

Solution: First, we put the equations into the form of a Lagrangian:

$$\begin{aligned} L(x, y, \lambda) &= f(x, y) - \lambda_1 g_1(x, y) - \lambda_2 g_2(x, y) && (60) \\ &= x^3 + y^3 - \lambda_1(x^2 - 1) - \lambda_2(y^2 - 1) && (61) \end{aligned}$$

and we solve for the gradient of the Lagrangian (Equation 4):

$$\nabla L(x, y, \lambda) = \nabla f(x, y) - \lambda_1 \nabla g_1(x, y) - \lambda_2 \nabla g_2(x, y) = 0 \qquad (62)$$

which gives us:

$$\begin{aligned} \frac{\partial}{\partial x} L(x, y, \lambda) &= 3x^2 - 2\lambda_1 x = 0 && (63) \\ \frac{\partial}{\partial y} L(x, y, \lambda) &= 3y^2 - 2\lambda_2 y = 0 && (64) \\ \frac{\partial}{\partial \lambda_1} L(x, y, \lambda) &= x^2 - 1 = 0 && (65) \\ \frac{\partial}{\partial \lambda_2} L(x, y, \lambda) &= y^2 - 1 = 0 && (66) \end{aligned}$$

Furthermore, we require that:

$$\begin{aligned} \lambda_1 &\geq 0 && (67) \\ \lambda_2 &\leq 0 && (68) \end{aligned}$$

since we are dealing with a inequality constraints.

From Equations 65 and 66, we have $x = \pm 1$ and $y = \pm 1$. Substituting $x = \pm 1$ into Equation 63 gives $\lambda_1 = \pm\frac{3}{2}$. Since we require that $\lambda_1 \geq 0$, then $\lambda_1 = \frac{3}{2}$ is the only valid choice for λ_1. Likewise, substituting $y = \pm 1$ into Equation 64 gives $\lambda_2 = \pm\frac{3}{2}$. Since we require that $\lambda_2 \leq 0$, then $\lambda_2 = -\frac{3}{2}$. This gives $x = 1$, $y = -1$ and $f = 0$.

Lagrangian Formulation

- So in the SVM problem the Lagrangian is

$$\min L_P = \tfrac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} a_i y_i \left(\mathbf{x}_i \cdot \mathbf{w} + b\right) + \sum_{i=1}^{l} a_i$$

s.t. $\forall i, a_i \geq 0$ where l is the # of training points

- From the property that the derivatives at min = 0 we get:

$$\frac{\partial L_P}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{l} a_i y_i \mathbf{x}_i = 0$$

$$\frac{\partial L_P}{\partial b} = \sum_{i=1}^{l} a_i y_i = 0 \quad \text{so}$$

$$\mathbf{w} = \sum_{i=1}^{l} a_i y_i \mathbf{x}_i, \quad \sum_{i=1}^{l} a_i y_i = 0$$

Problem: Lets assume that we have two classes of two-dimensional data to separate. Lets also assume that each class consists of only one point. These points are:

$$\begin{aligned} \overline{x}_1 &= A_1 = (1,1) \\ \overline{x}_2 &= B_1 = (2,2) \end{aligned} \tag{74}$$

Find the hyperplane that separates these two classes.

Karush-Kuhn-Tucker Conditions

There are five KKT conditions that affect all of our constraint based optimizations. I won't go into a proof, I'll just present them. They are:

$$
\begin{aligned}
\frac{\partial}{\partial \overline{w}} L(\overline{w}, b, \lambda) = \overline{w} - \sum_i \lambda_i y_i \overline{x}_i &= 0 && (69)\\
\frac{\partial}{\partial b} L(\overline{w}, b, \lambda) = -\sum_i \lambda_i y_i &= 0 && (70)\\
y_i \left[\langle \overline{w}, \overline{x} \rangle + b\right] - 1 &\geq 0 && (71)\\
\lambda_i &\geq 0 && (72)\\
\lambda_i (y_i \left[\langle \overline{w}, \overline{x} \rangle + b\right] - 1) &= 0 && (73)
\end{aligned}
$$

So, anytime we apply a constraint-based optimization, we must ensure that these conditions are satisfied.

Solution: From SVM theory, we know that the equations are:

$$f(\overline{w}) = \frac{1}{2}\|\overline{w}\|^2 \tag{75}$$

$$g_i(\overline{w}, b) = y_i\left[\langle\overline{w}, \overline{x}_i\rangle + b\right] - 1 \geq 0 \tag{76}$$

A common question here is *Why isn't g_i a function of $\overline{x}_i$?*. The answer of course is that x_i isn't a variable - each x_i has a value which we know from Equation 74. We can expand $g_i(\overline{w}, b)$ a bit further:

$$g_1(\overline{w}, b) = \left[\langle\overline{w}, \overline{x}_1\rangle + b\right] - 1 \geq 0 \tag{77}$$

$$g_2(\overline{w}, b) = -\left[\langle\overline{w}, \overline{x}_2\rangle + b\right] - 1 \geq 0 \tag{78}$$

Next, we put the equations into the form of a Lagrangian:

$$\begin{aligned} L(\overline{w}, b, \lambda) &= f(\overline{w}) - \lambda_1 g_1(\overline{w}, b) - \lambda_2 g_2(\overline{w}, b) \\ &= \frac{1}{2}\|\overline{w}\|^2 - \lambda_1(\left[\langle\overline{w}, \overline{x}_1\rangle + b\right] - 1) - \lambda_2(-\left[\langle\overline{w}, \overline{x}_2\rangle + b\right] - 1) \\ &= \frac{1}{2}\|\overline{w}\|^2 - \lambda_1(\left[\langle\overline{w}, \overline{x}_1\rangle + b\right] - 1) + \lambda_2(\left[\langle\overline{w}, \overline{x}_2\rangle + b\right] + 1) \end{aligned} \tag{79}$$

and we solve for the gradient of the Lagrangian (Equation 4):

$$\nabla L(\overline{w}, b, \lambda) = \nabla f(\overline{w}) - \lambda_1 \nabla g_1(\overline{w}, b) - \lambda_2 \nabla g_2(\overline{w}, b) = 0 \tag{80}$$

which gives us:

$$\frac{\partial}{\partial \overline{w}} L(\overline{w}, b, \lambda) = \overline{w} - \lambda_1 \overline{x}_1 + \lambda_2 \overline{x}_2 = 0 \tag{81}$$

$$\frac{\partial}{\partial b} L(\overline{w}, b, \lambda) = -\lambda_1 + \lambda_2 = 0 \tag{82}$$

$$\frac{\partial}{\partial \lambda_1} L(\overline{w}, b, \lambda) = \left[\langle\overline{w}, \overline{x}_1\rangle + b\right] - 1 = 0 \tag{83}$$

$$\frac{\partial}{\partial \lambda_2} L(\overline{w}, b, \lambda) = \left[\langle\overline{w}, \overline{x}_2\rangle + b\right] + 1 = 0 \tag{84}$$

This gives us enough equations to solve this analytically. Equating Equations 83 and 84 we get:

$$
\begin{aligned}
[\langle \overline{w}, \overline{x}_1 \rangle + b] - 1 &= [\langle \overline{w}, \overline{x}_2 \rangle + b] + 1 = 0 && (85)\\
\langle \overline{w}, \overline{x}_1 \rangle + b - 1 &= \langle \overline{w}, \overline{x}_2 \rangle + b + 1 && (86)\\
\langle \overline{w}, \overline{x}_1 \rangle - 1 &= \langle \overline{w}, \overline{x}_2 \rangle + 1 && (87)
\end{aligned}
$$

We have the values of $\overline{x}_1$ and $\overline{x}_2$ from Equation 74. This leaves $\overline{w}$ as the unknown. We can break $\overline{w}$ down into its components:

$$\overline{w} = (w_1, w_2) \tag{90}$$

Adding these into the mix we get:

$$
\begin{aligned}
\langle \overline{w}, [\overline{x}_1 - \overline{x}_2] \rangle &= 2 && (91)\\
\langle (w_1, w_2), [(1,1) - (2,2)] \rangle &= 2 && (92)\\
\langle (w_1, w_2), (-1,-1) \rangle &= 2 && (93)\\
-w_1 - w_2 &= 2 && (94)\\
w_1 &= -(2 + w_2) && (95)
\end{aligned}
$$

Adding values to Equation 81 and combining with Equation 82 gives us:

$$
\begin{aligned}
(w_1, w_2) - \lambda_1(1,1) + \lambda_2(2,2) &= 0 && (96)\\
(w_1, w_2) - \lambda_1(1,1) + \lambda_1(2,2) &= 0 && (97)\\
(w_1, w_2) + \lambda_1(1,1) &= 0 && (98)
\end{aligned}
$$

Which yields:

$$w_1 + \lambda_1 = 0 \tag{99}$$

and

$$w_2 + \lambda_1 = 0 \tag{100}$$

Equating these we get:

$$w_1 = w_2 \tag{101}$$

Putting this result back into Equation 95 gives:

$$w_1 = w_2 = -1 \tag{102}$$

Using this in either of Equations 99 or 100 will give:

$$\lambda_1 = \lambda_2 = 1 \tag{103}$$

And finally, using this in Equations 83 and 84 give:

$$
\begin{aligned}
b &= 1 - \langle \overline{w}, \overline{x}_1 \rangle && (104)\\
&= -1 - \langle \overline{w}, \overline{x}_2 \rangle && (105)\\
&= 1 - \langle (-1,-1), (1,1) \rangle && (106)\\
&= -1 - \langle (-1,-1), (2,2) \rangle && (107)\\
&= 3 && (108)
\end{aligned}
$$

Note that this result also satisfies all of the KKT conditions including:

$$\lambda_i(y_i\left[\langle \overline{w}, \overline{x}_i \rangle + b\right] - 1) = 0 \tag{109}$$

ie:

$$\begin{aligned}
\lambda_1([\langle \overline{w}, \overline{x}_1 \rangle + b] - 1) &= 0 && (110)\\
\lambda_2([\langle \overline{w}, \overline{x}_2 \rangle + b] + 1) &= 0 && (111)\\
([\langle(-1,-1),(1,1)\rangle + 3] - 1) &= -2 + 3 - 1 = 0 && (112)\\
([\langle(-1,-1),(2,2)\rangle + 3] + 1) &= -4 + 3 + 1 = 0 && (113)
\end{aligned}$$

and the inequality constaints:

$$\lambda_1 \geq 0 \tag{114}$$

$$\lambda_2 \geq 0 \tag{115}$$

The Lagrangian Dual Problem: instead of minimizing over $\mathbf{w}$, b, subject to constraints involving a's, we can maximize over a (the dual variable) subject to the relations obtained previously for $\mathbf{w}$ and b

Our solution must satisfy these two relations:

$$\mathbf{w} = \sum_{i=1}^{l} a_i y_i \mathbf{x}_i, \quad \sum_{i=1}^{l} a_i y_i = 0$$

By substituting for $\mathbf{w}$ and b back in the original eqn we can get rid of the dependence on $\mathbf{w}$ and b.

Note first that we already now have our answer for what the weights $\mathbf{w}$ must be: they are a linear combination of the training inputs and the training outputs, x_i and y_i and the values of a. We will now solve for the a's by differentiating the dual problem wrt a, and setting it to zero. Most of the a's will turn out to have the value zero. The non-zero a's will correspond to the support vectors

Primal problem:

$$\min L_P = \tfrac{1}{2}\left\|\mathbf{w}\right\|^2 - \sum_{i=1}^{l} a_i y_i \left(\mathbf{x}_i \cdot \mathbf{w} + b\right) + \sum_{i=1}^{l} a_i$$

s.t. $\forall i \ a_i \geq 0$

$$\mathbf{w} = \sum_{i=1}^{l} a_i y_i \mathbf{x}_i, \quad \sum_{i=1}^{l} a_i y_i = 0$$

Dual problem:

$$\max L_D(a_i) = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l} a_i a_j y_i y_j \left(\mathbf{x}_i \cdot \mathbf{x}_j\right)$$

s.t. $\sum_{i=1}^{l} a_i y_i = 0 \ \& \ a_i \geq 0$

(note that we have removed the dependence on **w** and b)

The Dual problem

- Kuhn-Tucker theorem: the solution we find here will be the same as the solution to the original problem
- Q: But why are we doing this???? (why not just solve the original problem????)
- Ans: Because this will let us solve the problem by computing the just the inner products of x_i, x_j (which will be very important later on when we want to solve non-linearly separable classification problems)

The Dual Problem

Dual problem:

$$\max L_D(a_i) = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l} a_i a_j y_i y_j \left(\mathbf{x}_i \cdot \mathbf{x}_j\right)$$

$$\text{s.t. } \sum_{i=1}^{l} a_i y_i = 0 \ \&\ a_i \geq 0$$

Notice that all we have are the dot products of x_i, x_j

If we take the derivative wrt a and set it equal to zero, we get the following solution, so we can solve for a_i:

$$\sum_{i=1}^{l} a_i y_i = 0$$

$$0 \leq a_i \leq C$$

Problem: Lets assume that we have two classes of two-dimensional data to separate. Lets also assume that each class consists of only one point. These points are:

$$\begin{aligned} \overline{x}_1 &= A_1 = (1,1) \\ \overline{x}_2 &= B_1 = (2,2) \end{aligned} \tag{116}$$

Find the hyperplane that separates these two classes.

Solution: Note that this is the same problem as the previous one, but we are going to solve it in a different way. This time we are going to use the Wolfe dual of the Lagrangian to to it. It is supposed to make things simpler. Lets see if it does. The equation for the primal representation of the Lagrangian for a SVM is:

$$L(\overline{w}, b, \lambda) = \frac{1}{2}\|w\|^2 - \lambda_1([\langle \overline{w}, \overline{x}_1 \rangle + b] - 1) + \lambda_2([\langle \overline{w}, \overline{x}_2 \rangle + b] + 1) \quad (117)$$

$$= \frac{1}{2}\|w\|^2 - \lambda_1(\langle \overline{w}, \overline{x}_1 \rangle + b - 1) + \lambda_2(\langle \overline{w}, \overline{x}_2 \rangle + b + 1) \quad (118)$$

$$= \frac{1}{2}\|w\|^2 - \lambda_1 \langle \overline{w}, \overline{x}_1 \rangle - \lambda_1 b + \lambda_1 + \lambda_2 \langle \overline{w}, \overline{x}_2 \rangle + \lambda_2 b + \lambda_2) \quad (119)$$

$$= \frac{1}{2}\|w\|^2 - \lambda_1 \langle \overline{w}, \overline{x}_1 \rangle + \lambda_2 \langle \overline{w}, \overline{x}_2 \rangle - \lambda_1 b + \lambda_2 b + \lambda_1 + \lambda_2 \quad (120)$$

If we substitute Equations 81 and 82 into this formulation we get:

$$\begin{aligned}
L(\lambda) &= \frac{1}{2}\left\|\lambda_1\overline{x}_1 - \lambda_2\overline{x}_2\right\|^2 \\
&\quad - \lambda_1\left\langle\lambda_1\overline{x}_1 - \lambda_2\overline{x}_2, \overline{x}_1\right\rangle + \lambda_2\left\langle\lambda_1\overline{x}_1 - \lambda_2\overline{x}_2, \overline{x}_2\right\rangle \qquad (121)\\
&\quad + b(\lambda_2 - \lambda_1) \\
&\quad + \lambda_1 + \lambda_2 \\
&= \frac{1}{2}(\lambda_1^2\left\langle\overline{x}_1, \overline{x}_1\right\rangle - 2\lambda_1\lambda_2\left\langle\overline{x}_1, \overline{x}_2\right\rangle + \lambda_2^2\left\langle\overline{x}_2, \overline{x}_2\right\rangle) \\
&\quad - \lambda_1^2\left\langle\overline{x}_1, \overline{x}_1\right\rangle + \lambda_1\lambda_2\left\langle\overline{x}_1, \overline{x}_2\right\rangle + \lambda_1\lambda_2\left\langle\overline{x}_1, \overline{x}_2\right\rangle - \lambda_2^2\left\langle\overline{x}_2, \overline{x}_2\right\rangle \qquad (122)\\
&\quad + b(0) \\
&\quad + \lambda_1 + \lambda_2 \\
&= \lambda_1 + \lambda_2 + \lambda_1\lambda_2\left\langle\overline{x}_1, \overline{x}_2\right\rangle - \frac{1}{2}\lambda_1^2\left\langle\overline{x}_1, \overline{x}_1\right\rangle - \frac{1}{2}\lambda_2^2\left\langle\overline{x}_2, \overline{x}_2\right\rangle \qquad (123)
\end{aligned}$$

which is the equation for the Wolfe Dual Lagrangian. Keep in mind that this is also subject to $\lambda_i \geq 0$ and all the KKT constraints [69-73]. Equation 82, which is a KKT constraint, can be rewritten as:

$$-\sum_i \lambda_i y_i = 0 \qquad (124)$$

and since it is a constraint, we must also take it into account when taking the gradient of the Lagrangian. We add it the same way we add any constraint. Thus the Dual Lagrangian becomes:

$$
\begin{aligned}
L(\lambda,\gamma) &= \lambda_1 + \lambda_2 + \lambda_1\lambda_2 \langle \overline{x}_1, \overline{x}_2 \rangle - \frac{1}{2}\lambda_1^2 \langle \overline{x}_1, \overline{x}_1 \rangle - \frac{1}{2}\lambda_2^2 \langle \overline{x}_2, \overline{x}_2 \rangle \\
&\quad - \gamma(\lambda_1 - \lambda_2) \qquad (125)
\end{aligned}
$$

$$
\begin{aligned}
&= \lambda_1 + \lambda_2 + \lambda_1\lambda_2 \langle \overline{x}_1, \overline{x}_2 \rangle - \frac{1}{2}\lambda_1^2 \langle \overline{x}_1, \overline{x}_1 \rangle - \frac{1}{2}\lambda_2^2 \langle \overline{x}_2, \overline{x}_2 \rangle \\
&\quad - \gamma\lambda_1 + \gamma\lambda_2 \qquad (126)
\end{aligned}
$$

Now we just need to find the Lagrange Multipliers, λ_1 and λ_2. To do this we solve for the gradient of the Dual Lagrangian which gives us:

$$
\begin{aligned}
\frac{\partial}{\partial \lambda_1} L(\lambda,\gamma) &= 1 + \lambda_2 \langle \overline{x}_1, \overline{x}_2 \rangle - \lambda_1 \langle \overline{x}_1, \overline{x}_1 \rangle - \gamma = 0 \qquad (127) \\
\frac{\partial}{\partial \lambda_2} L(\lambda,\gamma) &= 1 + \lambda_1 \langle \overline{x}_1, \overline{x}_2 \rangle - \lambda_2 \langle \overline{x}_2, \overline{x}_2 \rangle + \gamma = 0 \qquad (128) \\
\frac{\partial}{\partial \gamma} L(\lambda,\gamma) &= -\lambda_1 + \lambda_2 = 0 \qquad (129)
\end{aligned}
$$

Solving this gives $\lambda_1 = \lambda_2 = 1$ and $\gamma = 3$. It is interesting to note that we don't need to solve for γ explicitly. That is, we don't need it to solve for $\overline{w}$ or b. We can use KKT conditions 69 and 73 to solve for these.