ComS 573: Final Exam

Due on May 5, 2014

Professor De Brabanter at 11:45am

Josh Davis

Problem 6

We want to determine the posterior distribution. We know that the probability $p(\beta \mid X, Y)$ is going to be proportional to this posterior distribution This gives us:

$$p(\beta \mid \mathbf{X}, \mathbf{Y}) \propto f(\mathbf{Y} \mid \mathbf{X}, \beta) p(\beta \mid \mathbf{X}) = f(\mathbf{Y} \mid \mathbf{X}, \beta) p(\beta)$$

We were given the distribution of our β_i, which is:

$$p(\beta) = \prod_{i=1}^{d} p(\beta_i) = \prod_{i=1}^{d} \frac{1}{\sqrt{2c\pi}} \exp\left(-\frac{\beta_i^2}{2c}\right) = \left(\frac{1}{\sqrt{2c\pi}}\right)^d \exp\left(-\frac{1}{2c}\sum_{i=1}^{d}\beta_i^2\right)$$

Using our values:

$$\begin{aligned} f(\mathbf{Y} \mid \mathbf{X}, \beta) p(\beta) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right)\left(\frac{1}{\sqrt{2c\pi}}\right)^d \exp\left(-\frac{1}{2c}\sum_{i=1}^{d}\beta_i^2\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{\sqrt{2c\pi}}\right)^d \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(Y_i - \hat{Y}_i)^2\right] - \frac{1}{2c}\sum_{i=1}^{d}\beta_i^2\right) \end{aligned}$$

Let's take the log to simplify things a bit:

$$\begin{aligned} \log f(\mathbf{Y} \mid \mathbf{X}, \beta) p(\beta) &= \log\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{\sqrt{2c\pi}}\right)^d \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(Y_i - \hat{Y}_i)^2\right] - \frac{1}{2c}\sum_{i=1}^{d}\beta_i^2\right)\right] \\ &= n\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + d\left(\frac{1}{\sqrt{2c\pi}}\right) - \left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(Y_i - \hat{Y}_i)^2\right] + \frac{1}{2c}\sum_{i=1}^{d}\beta_i^2\right) \end{aligned}$$

By maximizing that, it is obvious that we want to minimize the second part, thus we want to minimize:

$$\frac{1}{2\sigma^2}\left(\text{RSS} + \frac{\sigma^2}{c}\sum_{i=1}^{d}\beta_i^2\right)$$

Well all be darned, that's just the Ridge Regression formula with $\lambda = \sigma^2/c$. =]

Problem 7

Suppose you have the following data:

Observation	X_1	X_2	Class
1	2	2	+1
2	2	-2	+1
3	-2	-2	+1
4	-2	2	+1
5	1	1	-1
6	1	1	-1
7	-1	-1	-1
8	-1	1	-1

Part A

Show that the original data is not linearly separable by making a sketch.

Solution

Here is a basic graph of the data where Red represents the +1 class and Green represents the -1 class. As we can see, there is no way to draw a line separating this data.

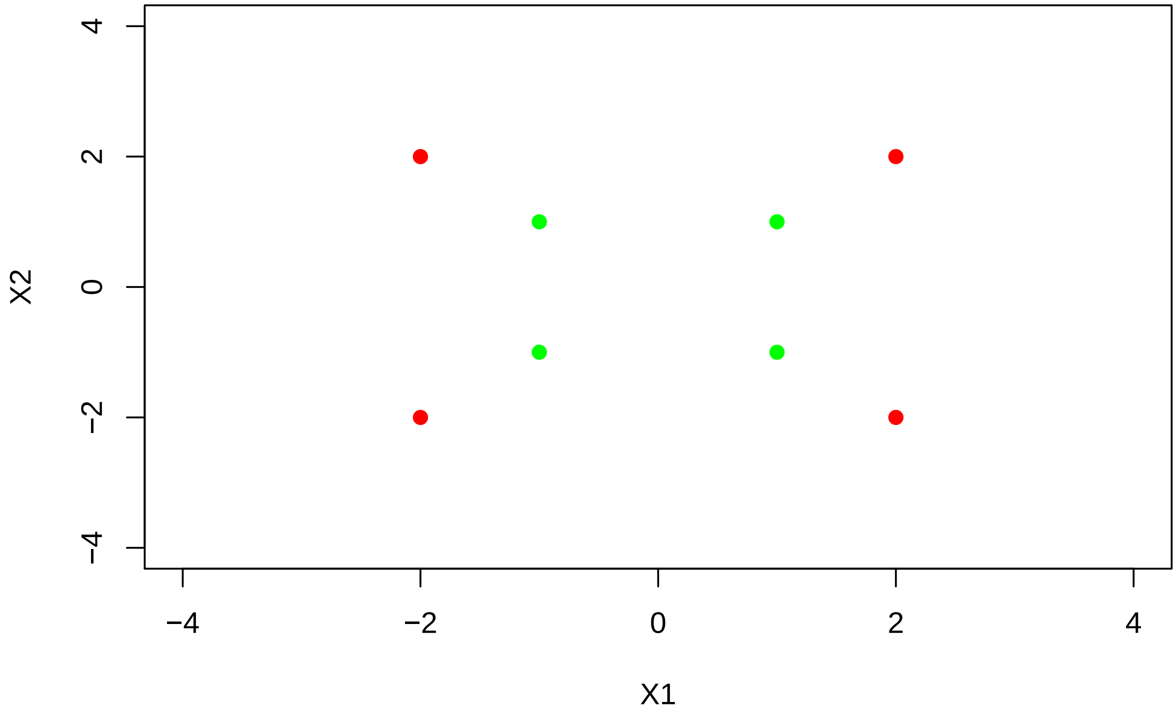

Figure 1: Plot of the data

Problem 7 continued on next page. . .

Part B

Make a sketch of the problem when applying the feature mapping φ to the original data.

Solution

Applying φ to our data, we put our X_1 and X_2 through the equation if the condition holds. If not, we just use the original data points. This gives us the following data:

Observation	X_1	X_2	Class	$\sqrt{X_1^2 + X_2^2} > 2$	new X_1	new X_2
1	2	2	+1	$\sqrt{8} > 2$	2	2
2	2	-2	+1	$\sqrt{8} > 2$	10	6
3	-2	-2	+1	$\sqrt{8} > 2$	6	6
4	-2	2	+1	$\sqrt{8} > 2$	6	10
5	1	1	-1	$\sqrt{2} < 2$	1	1
6	1	-1	-1	$\sqrt{2} < 2$	1	-1
7	-1	-1	-1	$\sqrt{2} < 2$	-1	-1
8	-1	1	-1	$\sqrt{2} < 2$	-1	1

Here is the data after the feature mapping:

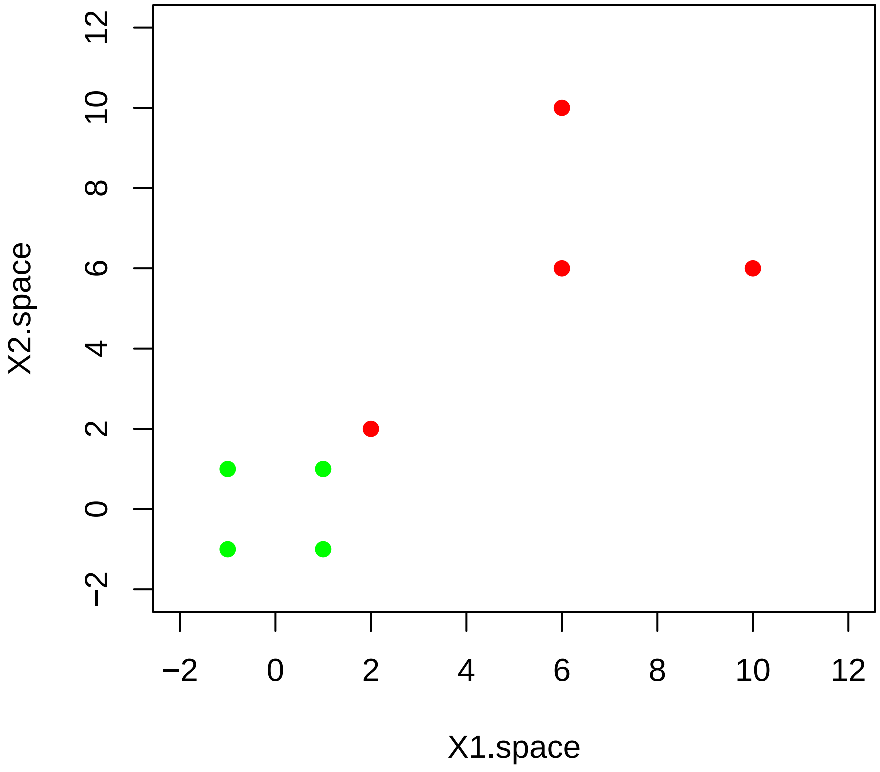

Figure 2: Plot of the data in the feature space

Problem 7 continued on next page. . .

Part C

Find the equation of the separating hyperplane in this feature space and the corresponding Lagrange multiplies α both *without* solving a QP problem. Plot the separating hyperplane on the plot from (b) as well.

Solution

Based on the plot, we know that we don't have to solve a QP problem because it is obvious which points are the SVMs. In this case, we have two:

$$\mathrm{sv}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathrm{sv}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

Our hyperplane is then:

$$H = w^T x + b = 0 = (w_1, w_2) \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + b = 0$$

For all our svs, $Y_i[w^T X_i + b] = 1$. This gives us:

$$(w_1, w_2) \begin{pmatrix} 2 \\ 2 \end{pmatrix} + b = 1 \tag{1}$$

$$(w_1, w_2) \begin{pmatrix} 1 \\ 1 \end{pmatrix} + b = -1 \tag{2}$$

Well shucks, we have three unknowns and only two equations. We'll need another one. As was mentioned in lecture, we can know that the cross product of the vector between our two points is going to be perpendicular to our final plane or also parallel to w. If we go with the latter, we get:

$$\mathbf{A} \times \mathbf{B} = \vec{0}$$

Using our two points, sv_1 and sv_2, we can calculate the vector as follows

$$\mathbf{A} = \begin{pmatrix} \mathrm{sv}_{2x} - \mathrm{sv}_{1x} \\ \mathrm{sv}_{2y} - \mathrm{sv}_{1y} \\ 0 \end{pmatrix} = \begin{pmatrix} 2-1 \\ 2-1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

and then using w, we get our next vector

$$\mathbf{B} = \begin{pmatrix} w_1 \\ w_2 \\ 0 \end{pmatrix}$$

The cross product is then

$$\mathbf{A} \times \mathbf{B} = \begin{pmatrix} \mathbf{A}_2\mathbf{B}_3 - \mathbf{A}_3\mathbf{B}_2 \\ \mathbf{A}_3\mathbf{B}_1 - \mathbf{A}_1\mathbf{B}_3 \\ \mathbf{A}_1\mathbf{B}_2 - \mathbf{A}_2\mathbf{B}_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1w_2 - 1w_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

This gives us our third equation

$$w_2 = w_1 \tag{3}$$

We can then subtract (2) from (1). Which will give us:

$$(2w_1 + 2w_2 - 1) - (w_1 + w_1 + 1) = w_1 + w_2 = 2$$

Then using (3) we can see that $w_1 = 1$ and $w_2 = 1$. Then substituting these back into (1) gives:

$$2w_1 + 2w_2 + b = 2 + 2 + b = 4 + b = 1$$

Problem 7 continued on next page. . .

Thus $b = -3$.

Using our Lagrangian formula where $w = \sum_{i=1}^{n} \alpha_i Y_i x_i$, this gives:

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \alpha_1 Y_1 x_1 + \alpha_2 Y_2 x_2 = -\alpha_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

Knowing this, we can then use our last Lagrangian where $\alpha_1 Y_1 + \alpha_2 + Y_2 = 0$ and get two equations:

$$2\alpha_2 - \alpha_1 = 1 \tag{4}$$

$$\alpha_2 = \alpha_1 \tag{5}$$

Solving these gives us $\alpha_1 = 1$ and $\alpha_2 = 1$.

Graphing our hyperplane on our previous plot we get:

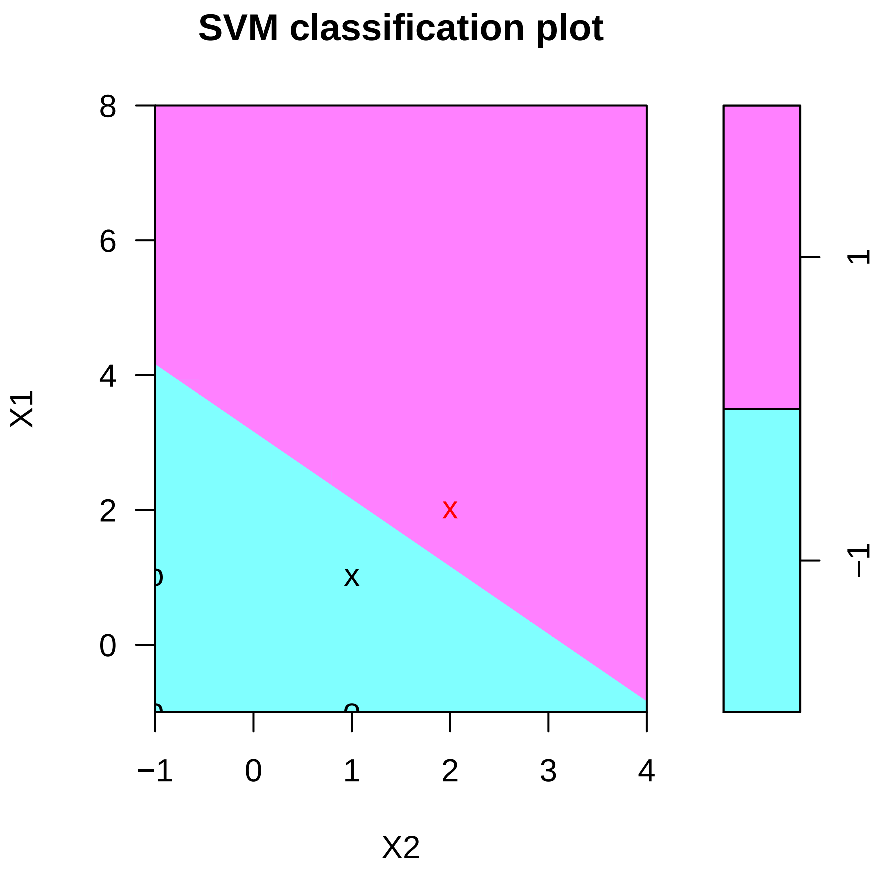

Figure 3: Plot of our SVM

Problem 8

Prove or disprove that least squares support vector machines for regression is a linear smoother. (*Hint:* Start by solving the linear system for b and α).

Solution

Let's derive the LSSVM like we did in class.

Instead of using inequality constraints, let's use equality constraints instead. This becomes:

$$\min_{w,b} \quad \frac{1}{2}w^T w + \gamma \frac{1}{2}\sum_{i=1}^{N} \xi_i^2$$

such that

$$Y_i(w^T \varphi(x_i) + b) = 1 - \xi_i$$

Our problem is no longer QP, we lose the interpretation and now we can solve it with a linear system.

Let's calculate our Lagrangians:

$$\mathcal{L}(w, b, \xi; \alpha) = \frac{1}{2}w^T w + \gamma \frac{1}{2}\sum_{i=1}^{N} \xi_i^2 - \sum_{k=1}^{N}\left[Y_k(w^T\varphi(x_k) + b) - 1 + \xi_k\right]$$

This gives us the following:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w} &= 0 \Rightarrow w = \sum_{k=1}^{N} \alpha_k Y_k \varphi(x_k) \\
\frac{\partial \mathcal{L}}{\partial b} &= 0 \Rightarrow b = \sum_{k=1}^{N} \alpha_k Y_k \\
\frac{\partial \mathcal{L}}{\partial \xi_k} &= 0 \Rightarrow \alpha_k = \gamma \xi_k \\
\frac{\partial \mathcal{L}}{\partial \alpha_k} &= 0 \Rightarrow 1 - \xi_k = Y_i(w^T\varphi(x_i) + b)
\end{aligned}$$

Then the linear system becomes:

$$\begin{bmatrix} 0 & \mathbf{Y}^T \\ \mathbf{Y} & \Omega + \frac{\mathbf{I}}{\gamma} \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 1_n \end{bmatrix}$$

Using Linear Algebra-fu, we get

$$b = \frac{1^T(\Omega + \frac{1}{\gamma}\mathbf{I}_N)^{-1} Y}{1^T(\Omega + \frac{1}{\gamma}\mathbf{I}_N)^{-1} 1_N}$$

$$a = (Y - 1_n b)(\Omega + \frac{1}{\gamma}\mathbf{I}_n)^{-1}$$

Like we mentioned in class, our model becomes:

$$f(x) = \sum_{k=1}^{N} \alpha K(x, x_k) + b$$

where we use a kernel $K(x_k, x_l) = \varphi(x_k)^T \varphi(x_l)$.

I'm in over my head because my linear algebra-fu is lacking. But if we were to continue, we'd see that our model is a linear combination of Y. This is a linear smoother and we'd be done.

Problem 9

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ and $\hat{\mathbf{Y}} = \hat{m}(x_1), \ldots, \hat{m}(x_n))^T$. Show for the linear smoother $\hat{\mathbf{Y}} = \mathbf{W}\mathbf{Y}$ that

$$\sum_{i=1}^{n} \operatorname{Cov}(\hat{Y}_i, Y_i) = \operatorname{trace}(\mathbf{W})\sigma_e^2$$

where σ_e^2 denotes the error variance. Explain in words what you have proved in terms of effective degrees of freedom.

Solution

First let's simplify the left side to see how

$$\begin{aligned}
\sum_{i=1}^{n} \operatorname{Cov}(\hat{Y}_i, Y_i) &= \operatorname{trace}(\operatorname{Cov}(\hat{Y}, Y)) \\
&= \operatorname{trace}(\operatorname{Cov}(Y, Y)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \\
&= \operatorname{trace}(\sigma_e^2\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \\
&= \sigma_e^2 \operatorname{trace}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \\
&= \sigma_e^2 \operatorname{trace}(\mathbf{W})
\end{aligned}$$

This shows that for the linear smoother that $\sum_{i=1}^{n} \operatorname{Cov}(\hat{Y}_i, Y_i) = \operatorname{trace}(\mathbf{W})\sigma_e^2$ are equal.

To see the relationship to effective degrees of freedom, let's calculate what the right side equals:

$$\begin{aligned}
\operatorname{trace}(\mathbf{W})\sigma_e^2 &= \operatorname{trace}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\sigma_e^2 \\
&= \operatorname{trace}((\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1})\sigma_e^2 \qquad \text{by the properties of trace} \\
&= \operatorname{trace}(\mathbf{I})\sigma_e^2 \qquad \text{where } \mathbf{I} \text{ is an identity matrix} \\
&= \sigma_e^2 \sum_{i=1}^{d} \mathbf{I}_{ii} \\
&= \sigma_e^2(1_1 + \cdots + 1_d) \\
&= \sigma_e^2(d)
\end{aligned}$$

This is called effective degrees of freedom because since we are performing regularization on our data, d becomes the number of parameters we are using. This is different than the entire dimensionality of the data because "effectively" only a few parameters are having an effect.