IE 7374 ST: Machine Learning in Engineering
HW-3

Rohit Bokade
NUID: 001280767

April 8, 2021

1. (a) Class-conditional probability for each class $i \in \{0, 1\}$ is given as

$$p(x|y=i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(x-m_i)^T \Sigma_i^{-1}(x-m_i)\right]$$

where $m_0 = (1, 2), m_1 = (6, 3), \Sigma_0 = \Sigma_1 = \mathbb{I}_2$ and $P(Y = 0) = P(Y = 1) = 1/2$. Also, point x is said to be on the decision surface or boundary if $P(Y = 1|x) = P(Y = 0|x)$. We can use Bayes' theorem to obtain the posterior $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$ for both the classes and equate them to find the optimal decision boundary.

$$\frac{p(x|y=0)p(y=0)}{p(x)} = \frac{p(x|y=1)p(y=1)}{p(x)}$$
$$\therefore p(x|y=0)p(y=0) = p(x|y=1)p(y=1)$$
$$\therefore p(x|y=0)(0.5) = p(x|y=1)(0.5)$$
$$\therefore p(x|y=0) = p(x|y=1)$$

$$\therefore \frac{1}{(2\pi)^{d/2}|\Sigma_0|^{1/2}} \exp\left[-\frac{1}{2}(x-m_0)^T \Sigma_0^{-1}(x-m_0)\right] = \frac{1}{(2\pi)^{d/2}|\Sigma_1|^{1/2}} \exp\left[-\frac{1}{2}(x-m_1)^T \Sigma_1^{-1}(x-m_1)\right]$$

$$\therefore \frac{1}{(2\pi)^{2/2}(1)^{1/2}} \exp\left[-\frac{1}{2}(x-(1,2))^T \begin{pmatrix}1,0\\0,1\end{pmatrix}^{-1} (x-(1,2))\right] = \frac{1}{(2\pi)^{2/2}1^{1/2}} \exp\left[-\frac{1}{2}(x-(6,3))^T \begin{pmatrix}1,0\\0,1\end{pmatrix}^{-1} (x-(6,3))\right]$$

$$\therefore (x_1-1, x_2-2)^T \begin{pmatrix} 1,0 \\ 0,1 \end{pmatrix} (x_1-1, x_2-2) = (x_1-6, x_2-3)^T \begin{pmatrix} 1,0 \\ 0,1 \end{pmatrix} (x_1-6, x_2-3)$$
$$\therefore (x_1-1)^2 + (x_2-2)^2 = (x_1-6)^2 + (x_2-3)^2$$
$$\therefore (x_1-1)^2 - (x_1-6)^2 + (x_2-2)^2 - (x_2-3)^2 = 0$$

$$\therefore (x_1^2 - 2x_1 + 1 - x_1^2 + 12x_1 - 36) + (x_2^2 - 4x_2 + 4 - x_2^2 + 6x_2 - 9) = 0$$
$$\therefore 10x_1 - 35 + 2x_2 - 5 = 0$$
$$\therefore 10x_1 + 2x_2 - 40 = 0$$
$$\therefore 5x_1 + x_2 = 20$$

(b)

$$\therefore \frac{1}{(2\pi)^{d/2}|\Sigma_0|^{1/2}} \exp\left[-\frac{1}{2}(x-m_0)^T\Sigma_0^{-1}(x-m_0)\right] = \frac{1}{(2\pi)^{d/2}|\Sigma_1|^{1/2}} \exp\left[-\frac{1}{2}(x-m_1)^T\Sigma_1^{-1}(x-m_1)\right]$$

$$\therefore \frac{1}{|\Sigma_0|^{1/2}} \exp\left[-\frac{1}{2}(x-m_0)^T\Sigma_0^{-1}(x-m_0)\right] = \frac{1}{|\Sigma_1|^{1/2}} \exp\left[-\frac{1}{2}(x-m_1)^T\Sigma_1^{-1}(x-m_1)\right]$$
$$\therefore \log\left(\frac{1}{|\Sigma_0|^{1/2}}\right) - \frac{1}{2}(x-m_0)^T\Sigma_0^{-1}(x-m_0) = \log\left(\frac{1}{|\Sigma_1|^{1/2}}\right) - \frac{1}{2}(x-m_1)^T\Sigma_1^{-1}(x-m_1)$$
$$\therefore \log\left(\frac{1}{|\Sigma_0|^{1/2}}\right) - \frac{1}{2}(x-m_0)^T\Sigma_0^{-1}(x-m_0) - \log\left(\frac{1}{|\Sigma_1|^{1/2}}\right) + \frac{1}{2}(x-m_1)^T\Sigma_1^{-1}(x-m_1) = 0$$

If $\Sigma_0 = \Sigma_1$, the decision boundary would be linear.

2. (a) There is no linear separator that can achieve a perfect classification score.
   - For $0 \le t \le -\infty$, points $\{(-2,1),(2,1)\}$ will always be misclassified as $H_{0\le t\le -\infty}(x) = 1$
   - And for $2 \le t \le \infty$, points $\{(1,1),(-1,1)\}$ will always be misclassified as $H_{2\le t\le\infty}(x) = -1$
   - For $-2 \le t \le 0$, point $\{(2,1)\}$ will always be misclassified as $H_{2\le t\le 0}(x) = 1$
   -
   - For $0 \le t \le 2$, point $\{(-1,1)\}$ will always be misclassified as $H_{0\le t\le 2}(x) = -1$

   (b)
$$S' = \{(\phi(x), y) : (x,y) \in S\}$$

   We can visually tell that the linear separation would be possible with the transformed data. The plane of maximal separation would be halfway between the two classes. That

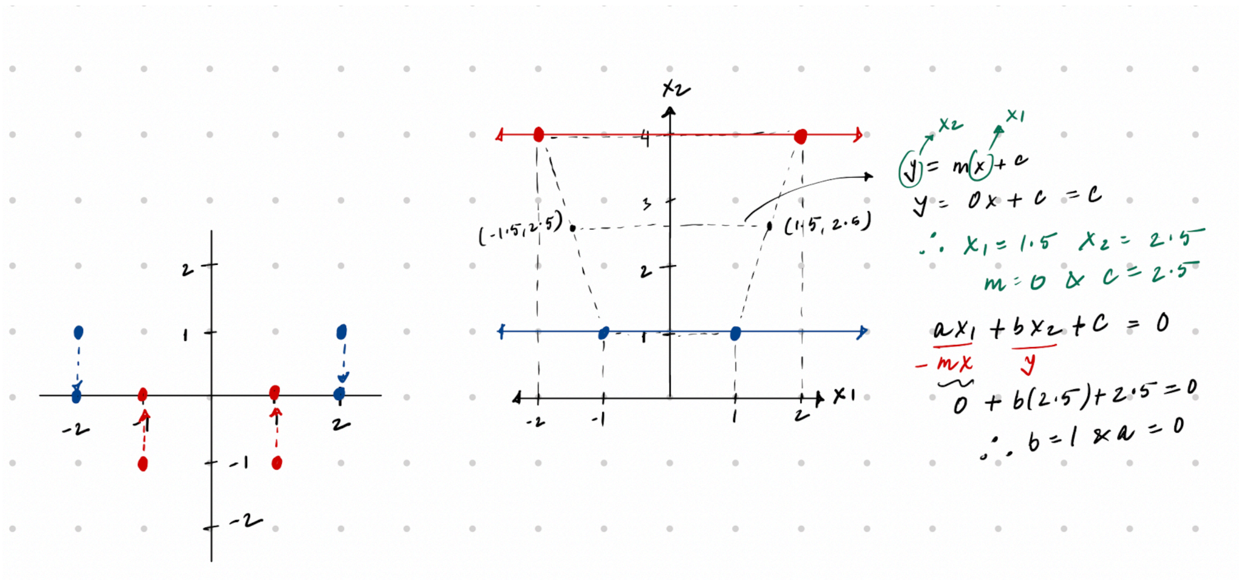

Figure 1: Transformed values of x

would be the line which passes through the midpoints {(-2, 4) and (-1, 2)} and {(2, 4) and (1, 2)}.
We can visually tell that the slope of the line would be zero.

$$x_2 = mx_1 + c$$
$$x_2 = c$$

Thus we have $x_1 = 1.5, x_2 = 2.5, m = 0, c = 2.5$.

$$\begin{aligned} H' &= \{ax_1 + bx_2 + c \geq 0 : a^2 + b^2 \neq 0\} \\ ax_1 + bx_2 + c &= 0 \\ 0 + b(2.5) + 2.5 &= 0 && \text{(plugging values from above)} \\ b &= 1 \end{aligned}$$

Thus, $x_2 \geq 2.5$ for a class to be classified as 1.

(c) Kernel function $K(x, z) = \phi(x)^T\phi(z) = (x, x^2)^T(z, z^2) = xz + x^2z^2$

3. (a) The upper bound on the number of misclassified instances can be given as

$$\sum_{i=1}^{n} \xi_i$$

(b) C [2] is the variable that controls the trade-off between the classification accuracy and the margin of the linear separators. In other words, it determines the influence of misclassification on the objective function.
As $C \to \infty$, the resulting hyperplane would have relatively smaller margin given it is able to better separate the classes.
As $C \to 0$, the optimizer would choose a relatively smaller margin hyperplane even if that hyperplane misclassifies a few samples. It would act as a regularization effect on the optimization.

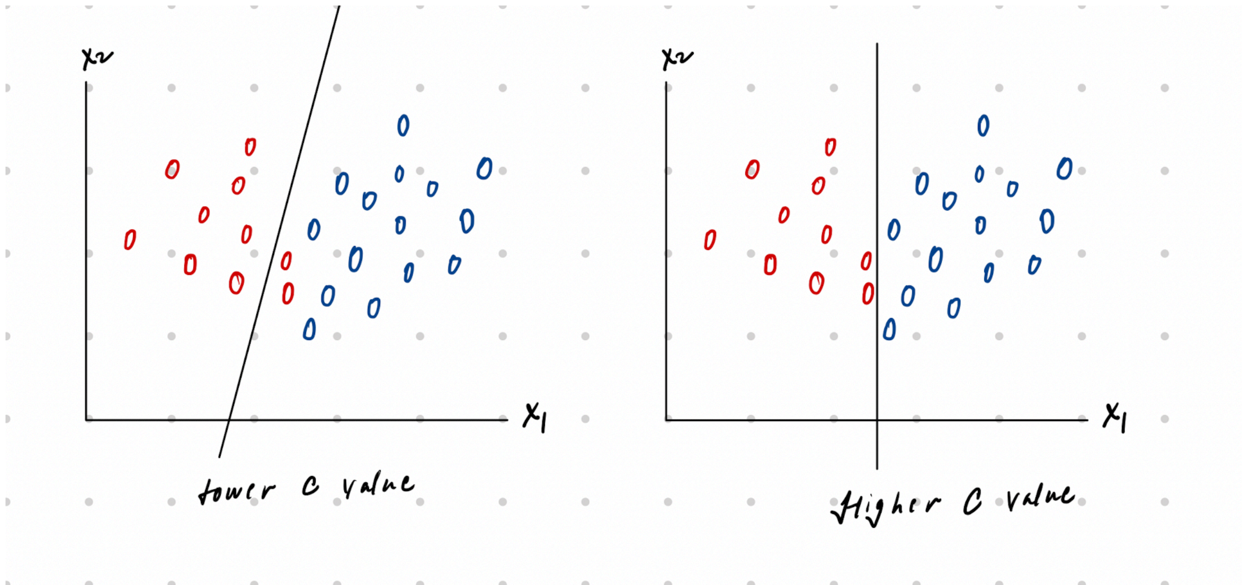

Figure 2: Influence of C

(c) We can consider the dual primal relationship

$$\phi(x_i)^T\phi(x_j) = k(x_i, x_j)$$

. Now, the estimate for a sample can be given as

$$\begin{aligned}\hat{y} &= sign(w^T\phi(x)) \\ &= sign(\sum_{i=1}^{n} \alpha_i y_i \phi(x_i)^T\phi(x)) \\ &= sign(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x_j))\end{aligned}$$

The kernel trick uses $k(x_i, x_j)$ instead of $\phi(x)^T\phi(x_j)$. Therefore, predictions can be made using $k(x_i, x_j)$ instead of using the $\phi(x)$ function.

4.

$$J(w) = \|Xw - y\| + \lambda\|w\|_2^2$$

We have positive semidefinite kernel k

(a) The kernelized version of the objective for a given kernel $k_{ij} = k(x_i, x_j)$ can be given as

$$J(\alpha) = \|k\alpha - y\| + \lambda\alpha^T k\alpha$$

(b) The prediction using a new point x^* can be given as

$$f_\alpha(x^*) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$$