E0 270- Machine Learning
Key for Assignment 1

1. (a) The Bayes classifier $f_B : \mathcal{X} \to \{+1, -1\}$ is given by

$$f_B(x) = \mathbf{1}(x \geq 0) - \mathbf{1}(x < 0)$$

and does not depend on the marginal $P(X)$. The Bayes error for all three marginals is also the same here and equal to 0.

(b) The Bayes classifier $f_B : \mathcal{X} \to \{+1, -1\}$ is given by

$$f_B(x) = \mathbf{1}(x \geq 0) - \mathbf{1}(x < 0)$$

and does not depend on the marginal $P(X)$.
The Bayes error in this case depends on the marginals. The Bayes error is given below for the three marginals.

$$\begin{aligned}
R^* &= 2\int_{-h}^{0} \frac{(x+h)}{2h} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx && \text{for } X \sim \mathcal{N}(0,1). \\
R^* &= \frac{h}{4} && \text{for } X \sim \text{Uniform}([-1,1]). \\
R^* &= \begin{cases} \frac{h}{3} & \text{if } h \leq \frac{1}{2} \\ \frac{2h}{3} + \frac{1}{3h}\left(-\frac{3h^2}{2} + \frac{h}{2} - \frac{1}{8}\right) & \text{Otherwise} \end{cases} && \text{for } X \sim \text{Uniform}([-\frac{1}{2},1]).
\end{aligned}$$

(c) The Bayes classifier $f_B : \mathcal{X} \to \{+1, -1\}$ is given by

$$f_B(x) = \begin{cases} 1 & \text{if } x \in [0.3, 0.5] \\ -1 & \text{Otherwise} \end{cases}$$

and does not depend on the marginal $P(X)$.
The Bayes error is given below for the three marginals.

$$\begin{aligned}
R^* &= 0.4(\Phi(0.3) - \Phi(0.1)) + 0.2(\Phi(0.5) - \Phi(0.3)) + 0.4(\Phi(0.7) - \Phi(0.5)) && \text{for } X \sim \mathcal{N}(0,1). \\
R^* &= 0.1 && \text{for } X \sim \text{Uniform}([-1,1]). \\
R^* &= 0.4/3 && \text{for } X \sim \text{Uniform}([-\frac{1}{2},1]).
\end{aligned}$$

2. This least squares problem has the same solution as the Fisher linear discriminant. Refer Bishop chapter 3 for more details.

3. (a) The Synth 3a and 3b datasets differ only in the margin. The Synth 3b dataset has a lesser margin and would require more iterations to converge on average over various permutations of the data.

(b) The perceptron mistake bound of $M \leq R^2/\gamma^2$ yields a value of d/γ^2

4. As the C value was increased from 0.1 to 10, the margin keeps increasing, and eventually finds the separating hyperplane with the largest margin. On both the datasets, the maximum observed value of the optimal lagrange multiplier α was less than 10, when $C = 10$, thus increasing C beyond 10 has no effect .

5. (a) The exact answers vary significantly due to different cross-validation fold splits. But one can reasonable expect a classifier with about 11% error on the test data for $C = 100$.

   (b) Once again the exact answers vary significantly, but the typical error rate numbers look as follows.

   Linear kernel : Test error: 0.13, Best $C = 10$.
   Polynomial kernel: Test error: 0.13, Best parameters: $C = 100, d = 3$.
   RBF kernel : Test error: 0.08, Best parameter: $C = 10, \sigma = \frac{1}{4}$.

6. (a) Both One Vs All, and All pairs perform well on the first dataset. One Vs All would fail to find a good separator for the second dataset due to the inseparability of the class in the middle from the other classes. All pairs would perform well on the second dataset as well.

   (b) Let $W^* = [\mathbf{w}_1^*, \ldots, \mathbf{w}_k^*] \in \mathbb{R}^{d \times k}$.
   Let $W_i \in \mathbb{R}^{d \times k}$ be the weight matrix at the end of the round when the i^{th} mistake was made, and let $x_i, \hat{y}_i y_i$ be the instance, the prediction and the label at this round. Let $U_i \in \mathbb{R}^{d \times k}$ be such that the y_ith column is x_i and the $\hat{y}_i$th column is $-x_i$. All other column of U_i are set to zero. The dot product for the space matrices is just the scalar product of the vector version of these matrices. We have

$$W_i.W^* = (W_{i-1} + U_i).W^* \geq W_{i-1}.W^* + 1 \geq i$$

   We also have

$$W_i.W^* \leq ||W_i||.||W^*|| = ||\sum_{j=1}^{i} U_i||.||W^*|| \leq \sqrt{2i}||W^*||$$

   Hence the number of mistakes is bounded by $2||W^*||^2$.