Department of Computer Science
Ashoka University

Introduction to Machine Learning

Assignment 2A

Collaborators: None **Name: Rushil Gupta**

Question 1

We compute

$$\varphi(x) \cdot \varphi(x') = \sum_{i=0}^{\infty} \frac{1}{\sqrt{i!}} e^{-x^2/2} x^i \frac{1}{\sqrt{i!}} e^{-x'^2/2} x'^i = e^{-(x^2+x'^2)/2} \sum_{i=0}^{\infty} \frac{(xx')^i}{i!} = e^{-(x^2+x'^2)/2} e^{xx'} = e^{-\frac{(x-x')^2}{2}}$$

Hence $\varphi(x) \cdot \varphi(x') = K(x, x')$, so since K can be written be as a dot product in some feature space, it is a valid kernel.

Question 2

Question 2.1

The hard-margin SVM finds, among all separating hyperplanes, the one that maximizes the margin between the two classes. More precisely, it solves the convex quadratic program

$$\min_{w,b} \|w\|^2 \quad \text{s.t.} \quad y_i\big(w^\top x_i + b\big) \geq 1 \quad \forall i,$$

which is equivalent to maximizing the geometric margin $2/\|w\|$. In contrast, the Perceptron algorithm merely finds *a* separating hyperplane by enforcing

$$y_i\big(w^\top x_i + b\big) > 0 \quad \forall i,$$

but does not attempt to maximize the margin. Consequently:

- **Generalization.** A larger margin yields tighter bounds on the generalization error, whereas the Perceptron solution may have arbitrarily small margin and thus poorer guarantees on unseen data.
- **Uniqueness and Stability.** The SVM optimization is strictly convex, so it yields a unique (w, b) regardless of initialization. The Perceptron can converge to many different solutions depending on the order of updates and the starting point.
- **Robustness.** By maximizing the margin, SVMs are more robust to noise in the training set near the decision boundary than Perceptron-found hyperplanes.

Question 2.3

i) Feasibility of A's optimum for B.

Let (w_A, b_A) be an optimal solution to (A). Then:

$$y_i \left(w_A^\top x_i + b_A\right) \geq 1 \quad \forall i.$$

Then, for any i:

$$y_i \left(w_A^\top x_i + b_A\right) \geq 1 \geq 0 \implies y_i \left(w_A^\top x_i + b_A\right) \geq 0$$

Also, since all $y_i \left(w_A^\top x_i + b_A\right) \geq 1$ and $y_i \in \{-1, 1\}$, we know:

$$\begin{aligned} & \min_i y_i \left(w_A^\top x_i + b_A\right) \geq 1 \\ \implies & \min_i |y_i \left(w_A^\top x_i + b_A\right)| \geq 1 \\ \implies & \min_i |w_A^\top x_i + b_A| \geq 1 \end{aligned}$$

Lastly, since we know that for the support vectors, the margin is equal to 1, we get:

$$\min_i |w_A^\top x_i + b_A| = 1$$

Thus, we can conclude that the optimum of (A) is feasible for (B).

ii) Feasibility of B's optimum for A.

Let (w_B, b_B) be an optimal solution to (B). Then:

$$\begin{aligned} y_i \left(w_B^\top x_i + b_B\right) \geq 0 \quad \forall i \\ \min_i |w_B^\top x_i + b_B| = 1 \end{aligned}$$

Then, for any i:

$$\min_i |w_B^\top x_i + b_B| = 1 \implies |w_B^\top x_i + b_B| \geq 1 \quad \forall i$$

Since $y_i \in \{-1, 1\}$, and $y_i \left(w_B^\top x_i + b_B\right) \geq 0$, we can conclude that:

$$y_i \left(w_B^\top x_i + b_B\right) \geq 1 \quad \forall i$$

Thus, we can conclude that the optimum of (B) is feasible for (A).

iii) Optimality of A's optimum for B.

Let (w_A, b_A) be an optimal solution to (A). Then since (w_A, b_A) is feasible for (B),

$$\min_{(w,b)\in \mathrm{B}} \|w\|^2 \leq \|w_A\|^2 = \min_{(w,b)\in \mathrm{A}} \|w\|^2.$$

Similarly, (w_B, b_B) is feasible for (A), so

$$\min_{(w,b)\in \mathrm{A}} \|w\|^2 \leq \|w_B\|^2 = \min_{(w,b)\in \mathrm{B}} \|w\|^2.$$

Combining these shows the two minima coincide and that any optimal solution to one formulation is optimal for the other, i.e.:

$$\min_{(w,b)\in \mathrm{A}} \|w\|^2 = \min_{(w,b)\in \mathrm{B}} \|w\|^2.$$

Question 3

Question 3.1

The hard-margin SVM does not converge for data that is not linearly separable. This is because the hard-margin SVM tries to find a hyperplane that separates the two classes with maximum margin. If the data is not linearly separable, there will be no hyperplane that can separate the two classes, and thus the optimization problem will not have a solution. In the figure given, there is no hyperplane that can separate the two classes, so the hard-margin SVM will not converge.

Question 3.2

As $C \to \infty$, the soft-margin SVM becomes more similar to the hard-margin SVM. This is because a larger value of C means that we are penalizing misclassifications more heavily, which forces the SVM to find a hyperplane that separates the two classes with maximum margin, similar to the hard-margin SVM.

On the other hand, as $C \to 0^+$, the soft-margin SVM becomes less similar to the hard-margin SVM. In fact, since the weight given to epsilon is very small, the SVM will not care about misclassifications at all. So, it will not give any valuable solution to the data.

Question 3.3

Let (w, b) be a solution to the hard-margin SVM. Then, by formulation 2, we have:

$$\begin{aligned} y_i \left(w^\top x_i + b\right) &\geq 0 \quad \forall i \\ \min_i |w^\top x_i + b| &= 1 \end{aligned}$$

The condition 2 implies that there exists at least one i such that:

$$\begin{aligned} & |w^\top x_i + b| = 1 \\ \implies & w^\top x_i + b = 1 \quad \text{or} \quad w^\top x_i + b = -1 \end{aligned}$$

So, we can conclude that at least one training data point lies on each of these margin hyperplanes.

Question 4

Question 4.1

Consider the l_2-soft-margin problem

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 \quad \text{s.t.} \quad y_i\big(w^T x_i + b\big) \geq 1 - \xi_i, \quad i = 1, \ldots, n,$$

with or without the additional constraints $\xi_i \geq 0$. Suppose (w^*, b^*, ξ^*) is optimal for the version without $\xi_i \geq 0$. If for some i, $\xi_i^* < 0$, then

$$y_i\big(w^{*T} x_i + b^*\big) \geq 1 - \xi_i^* > 1,$$

so replacing ξ_i^* by $\tilde{\xi}_i = \max\{\xi_i^*, 0\} = 0$ preserves feasibility and decreases the objective (since $\xi_i^{*2} > 0$ but $\tilde{\xi}_i^2 = 0$). Hence at optimum $\xi_i^* \geq 0$ for all such i, and the two formulations have the same optimal value.

Question 4.2

Introduce multipliers $\alpha_i \geq 0$ for the constraints $y_i(w^T x_i + b) \geq 1 - \xi_i$. The Lagrangian is

$$L(w, b, \xi, \alpha) = \frac{1}{2}\, w^T w + \frac{C}{2}\, \xi^T \xi - \sum_{i=1}^{n} \alpha_i\big(y_i(w^T x_i + b) - 1 + \xi_i\big),$$

where $\xi = (\xi_1, \ldots, \xi_n)^T$.

Question 4.3

Setting derivatives to zero gives

$$\nabla_w L: \quad w - \sum_{i=1}^{n} \alpha_i\, y_i\, x_i = 0 \quad \Longrightarrow \quad w = \sum_{i=1}^{n} \alpha_i\, y_i\, x_i,$$

$$\frac{\partial L}{\partial b}: \quad -\sum_{i=1}^{n} \alpha_i\, y_i = 0,$$

$$\nabla_\xi L: \quad C\,\xi - \alpha = 0 \quad \Longrightarrow \quad \xi = \frac{1}{C}\,\alpha,$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)^T$.

Question 5

Question 5.1

Let $X \in \mathbb{R}^{n \times d}$ be the matrix whose i-th row is x_i^T, and let $y \in \mathbb{R}^n$ be the label vector.

The regression objective is

$$J(\theta) = \sum_{i=1}^{n} \big(y_i - \theta^\top x_i\big)^2 + \frac{\lambda}{2}\,\theta^\top \theta.$$

Taking the gradient with respect to θ and setting it to zero gives:

$$\begin{aligned} \nabla_\theta J(\theta) &= -2 \sum_{i=1}^{n} \big(y_i - \theta^\top x_i\big) x_i + \lambda\theta \\ &= -2X^\top (y - X\theta) + \lambda\theta \\ &= 0 \end{aligned}$$

Rearranging gives:

$$\begin{aligned} \lambda\theta &= 2X^\top (y - X\theta) \\ \implies \lambda\theta + 2X^\top X\theta &= 2X^\top y \\ \implies (\lambda I_d + 2X^\top X)\theta &= 2X^\top y \end{aligned}$$

Assuming $X^\top X + \frac{1}{2}\lambda I_d$ is invertible, the unique minimiser is

$$\hat{\theta} = \left(X^\top X + \frac{1}{2}\lambda I_d\right)^{-1} X^\top y$$

Question 5.2

Let $\phi(x)$ be any feature map and define

$$\Phi = \begin{bmatrix} \phi(x_1)^\top \\ \vdots \\ \phi(x_n)^\top \end{bmatrix} \in \mathbb{R}^{n \times D}, \qquad K = \Phi\Phi^\top \in \mathbb{R}^{n \times n}, \quad K_{ij} = k(x_i, x_j) = \phi(x_i)^\top \phi(x_j).$$

The ridge objective in feature space is

$$J(\theta) = \|y - \Phi\theta\|^2 + \frac{\lambda}{2}\,\theta^\top \theta.$$

Using the identity $(\lambda I + BA)^{-1}B = B(\lambda I + AB)^{-1}$ with $A = \Phi^\top, B = \Phi$ gives

$$\theta^\star = \big(\Phi^\top\Phi + \tfrac{\lambda}{2} I_D\big)^{-1}\Phi^\top y = \Phi^\top \Big(K + \tfrac{\lambda}{2} I_n\Big)^{-1} y.$$

Therefore $\theta^\star$ lies in the span of the training data:

$$\theta^\star = \sum_{i=1}^{n} \alpha_i\, \phi(x_i) \quad \text{where} \quad \alpha = \Big(K + \tfrac{\lambda}{2} I_n\Big)^{-1} y.$$

For a new point x_{new} define

$$k_{\text{new}} = \begin{bmatrix} k(x_1, x_{\text{new}}) \\ \vdots \\ k(x_n, x_{\text{new}}) \end{bmatrix} \in \mathbb{R}^n.$$

The prediction uses only kernel evaluations:

$$\hat{y}_{\text{new}} = \theta^{\star\top}\phi(x_{\text{new}}) = k_{\text{new}}^\top \alpha = k_{\text{new}}^\top \Big(K + \tfrac{\lambda}{2} I_n\Big)^{-1} y.$$

This closed-form expression avoids ever computing $\phi(x)$ explicitly.