Tutorial: Support Vector Machines

Cheat Sheet and Practice Problems

ES335 - Machine Learning
IIT Gandhinagar

July 23, 2025

1 Summary from Slides

1.1 Key Concepts

What is SVM?

- Binary classification algorithm that finds optimal separating hyperplane
- Maximizes margin between classes for better generalization
- Uses support vectors (closest points to decision boundary)
- Can handle non-linearly separable data using kernel trick

Core Idea: Margin Maximization

- Margin = perpendicular distance between two parallel hyperplanes
- Distance between hyperplanes $\mathbf{w} \cdot \mathbf{x} + b_1 = 0$ and $\mathbf{w} \cdot \mathbf{x} + b_2 = 0$ is $\frac{|b_1 - b_2|}{\|\mathbf{w}\|}$
- Maximum margin classifier: margin $= \frac{2}{\|\mathbf{w}\|}$

1.2 Hard Margin SVM (Linearly Separable)

Primal Formulation:

$$\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 \tag{1}$$

$$\text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad \forall i \tag{2}$$

Dual Formulation:

$$\text{maximize} \quad L(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \tag{3}$$

$$\text{subject to} \quad \sum_{i=1}^{N} \alpha_i y_i = 0, \quad \alpha_i \geq 0 \tag{4}$$

KKT Conditions:

- $\alpha_i(y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1) = 0$ (complementary slackness)
- Support vectors: $\alpha_i \neq 0$ where $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1$
- Non-support vectors: $\alpha_i = 0$ where $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) > 1$

1.3 Kernel Methods

Kernel Trick:

- Transform data to higher dimensional space using $\phi : \mathbb{R}^d \to \mathbb{R}^D$
- Compute kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$ efficiently
- Avoid explicit computation of $\phi(\mathbf{x})$

Common Kernels:

- Linear: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$
- Polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (c + \mathbf{x}_i \cdot \mathbf{x}_j)^d$
- RBF (Gaussian): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$

Properties:

- RBF kernel corresponds to infinite-dimensional feature space
- RBF SVM is non-parametric (model complexity grows with data)
- Linear/polynomial kernels are parametric

1.4 Soft Margin SVM

Motivation: Handle non-separable data with noise and outliers

Primal Formulation:

$$\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i \tag{5}$$

$$\text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{6}$$

Dual Formulation:

$$\text{maximize} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \tag{7}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{8}$$

Hinge Loss Formulation:

$$\text{minimize} \sum_{i=1}^{N} \max[0, 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + b)] + \frac{1}{2C}\|\mathbf{w}\|^2$$

Parameter C:

- Large C: Less tolerance for violations (high variance, low bias)
- Small C: More tolerance for violations (low variance, high bias)
- $C \to \infty$: Hard margin SVM

2 Practice Problems

Exercise 1: Basic SVM Concepts

Consider a 1D dataset with points: $(1, +1), (2, +1), (-1, -1), (-2, -1)$.
Part A: What is the optimal separating hyperplane?
Part B: Calculate the margin of this hyperplane.
Part C: Which points are the support vectors?
Part D: What are the values of α_i for each point?

Exercise 2: Distance Between Hyperplanes

Given two parallel hyperplanes:

- $\mathbf{w} \cdot \mathbf{x} + 5 = 0$
- $\mathbf{w} \cdot \mathbf{x} - 3 = 0$

where $\|\mathbf{w}\| = 4$.
Part A: Calculate the distance between these hyperplanes.
Part B: If this represents the margin in an SVM, what is $\|\mathbf{w}\|$ for the decision boundary $\mathbf{w} \cdot \mathbf{x} + 1 = 0$?

Exercise 3: KKT Conditions

For an SVM with solution $\mathbf{w} = [1, -1]^T$, $b = 0$, consider the following points:

- Point A: $\mathbf{x} = [2, 1]^T$, $y = +1$, $\alpha = 0.5$
- Point B: $\mathbf{x} = [0, 1]^T$, $y = +1$, $\alpha = 0$
- Point C: $\mathbf{x} = [-1, 0]^T$, $y = -1$, $\alpha = 0.5$

Part A: Verify the KKT complementary slackness condition for each point.
Part B: Which points are support vectors?
Part C: Is the constraint $\sum_i \alpha_i y_i = 0$ satisfied?

Exercise 4: Dual Problem Setup

Consider a binary classification problem with 3 training points:

- $(\mathbf{x}_1, y_1) = ([1, 0]^T, +1)$
- $(\mathbf{x}_2, y_2) = ([0, 1]^T, +1)$
- $(\mathbf{x}_3, y_3) = ([-1, -1]^T, -1)$

Part A: Write the dual objective function $L(\boldsymbol{\alpha})$ explicitly.
Part B: What are the constraints on $\alpha_1, \alpha_2, \alpha_3$?
Part C: If the optimal solution is $\alpha_1 = 0.5, \alpha_2 = 0.5, \alpha_3 = 1$, find $\mathbf{w}$.

Exercise 5: Kernel Computation

Consider the polynomial kernel $K(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x} \cdot \mathbf{z})^2$ for $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$.
Part A: For $\mathbf{x} = [2, 1]^T$ and $\mathbf{z} = [1, 3]^T$, compute $K(\mathbf{x}, \mathbf{z})$.
Part B: Find the explicit feature mapping $\phi(\mathbf{x})$ such that $K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$.
Part C: What is the dimensionality of the feature space?
Part D: Verify your answer by computing $\phi(\mathbf{x}) \cdot \phi(\mathbf{z})$ directly.

Exercise 6: RBF Kernel Properties

Consider the RBF kernel $K(\mathbf{x}, \mathbf{z}) = \exp(-\gamma \|\mathbf{x} - \mathbf{z}\|^2)$ with $\gamma = 0.5$.
Part A: Compute $K([0, 0]^T, [1, 1]^T)$.
Part B: What happens to $K(\mathbf{x}, \mathbf{z})$ as $\|\mathbf{x} - \mathbf{z}\| \to \infty$?
Part C: What happens to $K(\mathbf{x}, \mathbf{z})$ as $\|\mathbf{x} - \mathbf{z}\| \to 0$?
Part D: How does increasing γ affect the kernel's behavior?
Part E: Why is RBF SVM considered non-parametric?

Exercise 7: Soft Margin Formulation

Consider a soft margin SVM with regularization parameter $C = 2$.
Part A: Write the primal optimization problem.
Part B: Express the problem in hinge loss form.
Part C: What are the dual constraints?
Part D: If a training point has $\xi_i = 1.5$, what does this mean for the point's classification?

Exercise 8: Support Vector Analysis

In a soft margin SVM, classify the following scenarios for training points:
Part A: Point with $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1.5$ and $\alpha_i = 0$
Part B: Point with $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1$ and $\alpha_i = 0.3$
Part C: Point with $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 0.2$ and $\alpha_i = C$
Part D: Point with $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = -0.5$ and $\alpha_i = C$
For each case, determine: (i) if it's a support vector, (ii) if it's correctly classified, (iii) the value of ξ_i.

Exercise 9: Prediction with Kernels

Given an SVM with RBF kernel $K(\mathbf{x}, \mathbf{z}) = \exp(-\|\mathbf{x} - \mathbf{z}\|^2)$ and the following support vectors:

- $\mathbf{x}_1 = [1, 0]^T$, $y_1 = +1$, $\alpha_1 = 0.5$
- $\mathbf{x}_2 = [-1, 0]^T$, $y_2 = -1$, $\alpha_2 = 0.5$

with $b = 0$.
Part A: Write the decision function $f(\mathbf{x})$.
Part B: Predict the class for test point $\mathbf{x}_{test} = [0.5, 0]^T$.
Part C: Predict the class for test point $\mathbf{x}_{test} = [0, 1]^T$.
Part D: At what point would the decision function output exactly 0?

Exercise 10: Multi-class SVM

You have a 3-class problem with classes A, B, and C. Using the one-vs-all approach:
Part A: How many binary classifiers do you need?
Part B: For a test point, the classifiers output:

- A vs (B,C): $f_A(\mathbf{x}) = 0.8$
- B vs (A,C): $f_B(\mathbf{x}) = 0.6$
- C vs (A,B): $f_C(\mathbf{x}) = -0.2$

Which class should be predicted?
Part C: What could be a problem with this approach?
Part D: How many classifiers would one-vs-one approach require?

Exercise 11: Regularization Parameter C

Consider the effect of parameter C in soft margin SVM:
Part A: What happens to the decision boundary as $C \to 0$?
Part B: What happens to the decision boundary as $C \to \infty$?
Part C: Given a noisy dataset, should you use high or low C? Justify your answer.
Part D: How does C affect the bias-variance tradeoff?
Part E: If you have 100 training points and C = 10, what's the maximum possible value of $\sum_{i=1}^{100} \xi_i$?

Exercise 12: Hinge Loss Analysis

The hinge loss is defined as $\ell(y, f(\mathbf{x})) = \max[0, 1 - y \cdot f(\mathbf{x})]$.
Part A: For a correctly classified point with $y \cdot f(\mathbf{x}) = 2$, what is the hinge loss?
Part B: For a point on the margin with $y \cdot f(\mathbf{x}) = 1$, what is the hinge loss?
Part C: For a misclassified point with $y \cdot f(\mathbf{x}) = -0.5$, what is the hinge loss?
Part D: At what value of $y \cdot f(\mathbf{x})$ is the hinge loss non-differentiable?
Part E: Compare hinge loss with 0-1 loss and logistic loss for $y \cdot f(\mathbf{x}) \in [-2, 3]$.

Exercise 13: Computational Complexity

Consider the computational aspects of SVM:
Part A: Why is the dual formulation preferred for implementing the kernel trick?
Part B: In the dual problem, what is the dominant computational cost?
Part C: For a dataset with n points and d features, compare the computational complexity of:

- Linear kernel SVM
- RBF kernel SVM
- Explicit feature mapping with degree-2 polynomial kernel

Part D: How does the number of support vectors affect prediction time?
Part E: Why might you prefer linear SVM over RBF SVM for very large datasets?

Exercise 14: Advanced Kernel Design

Design custom kernels for specific scenarios:
Part A: Prove that $K(\mathbf{x}, \mathbf{z}) = K_1(\mathbf{x}, \mathbf{z}) + K_2(\mathbf{x}, \mathbf{z})$ is a valid kernel if K_1 and K_2 are valid kernels.
Part B: Prove that $K(\mathbf{x}, \mathbf{z}) = K_1(\mathbf{x}, \mathbf{z}) \cdot K_2(\mathbf{x}, \mathbf{z})$ is a valid kernel if K_1 and K_2 are valid kernels.
Part C: Design a kernel for text classification where documents are represented as word frequency vectors.
Part D: For time series data, propose a kernel that captures both magnitude and temporal patterns.
Part E: What properties must a function $K(\mathbf{x}, \mathbf{z})$ satisfy to be a valid kernel (Mercer's theorem)?

Exercise 15: Comprehensive SVM Problem

You are given a 2D dataset that forms two concentric circles (inner circle: class +1, outer circle: class -1).
Part A: Why would linear SVM fail on this dataset?
Part B: Propose a feature transformation $\phi(\mathbf{x})$ that could make the data linearly separable.
Part C: Design a custom kernel that directly computes the dot product in your transformed space.
Part D: How would you choose between polynomial and RBF kernels for this problem?
Part E: If there are outliers in the inner circle that are very close to the decision boundary, how would you handle them?
Part F: Describe a complete pipeline for solving this problem, including data preprocessing, model selection, and evaluation.

3 Key Takeaways

- SVM finds the optimal separating hyperplane by maximizing the margin
- Dual formulation enables the kernel trick for non-linear classification
- Support vectors are the critical points that define the decision boundary
- Soft margin allows handling of non-separable data with controlled violations
- Parameter C controls the bias-variance tradeoff in soft margin SVM
- Kernel choice significantly impacts model performance and interpretability
- Hinge loss provides a convex surrogate for 0-1 loss
- SVM is both theoretically grounded and practically effective