DSC5001: Statistical Machine Learning I

Tutorial 11: Maximum Margin, Soft-Margin SVM, and Kernels

Kangning Cui
City University of Hong Kong (Dongguan)

Semester A, 2026/27

Task 1: Maximum-Margin Sensor Classification

A sensor gives one numerical reading x. The label $y = -1$ means normal; $y = +1$ means an alert.

Case	A	B	C	D
x	-4	-1	2	5
y	-1	-1	$+1$	$+1$

Use $f(x) = wx + b$ and the hard-margin constraints $y_i f(x_i) \geq 1$.

(a) Find the optimal w, b minimizing $w^2/2$ and the decision boundary.

(b) Calculate all four signed scores $y_i f(x_i)$, identify the support vectors, and find the full margin width.

(c) Using $w = \sum_i \alpha_i y_i x_i$, $\sum_i \alpha_i y_i = 0$, and complementary slackness, find the optimal $\alpha_i \geq 0$.

(d) A new normal case arrives at $x = 1$. Refit the hard-margin classifier and compare the margin widths.

Solution 1(a): Margin Constraints

The closest cases of opposite classes are $B = (-1, -1)$ and $C = (2, +1)$.
The separator must increase with x, so $w > 0$.
Their constraints give

$$B: \quad -f(-1) = w - b \geq 1,$$
$$C: \quad f(2) = 2w + b \geq 1.$$

Adding the inequalities eliminates b:

$$\boxed{3w \geq 2 \quad \Longrightarrow \quad w \geq \frac{2}{3}.}$$

Minimizing $\frac{1}{2}w^2$ favors the smallest feasible positive w.

Solution 1(a): Hard-Margin Solution

At the optimum, both closest observations lie on their margin planes:

$$\begin{cases} w - b = 1, \\ 2w + b = 1. \end{cases}$$

Solving gives

$$\boxed{w = \frac{2}{3}, \qquad b = -\frac{1}{3}.}$$

Hence

$$\boxed{f(x) = \frac{2}{3}x - \frac{1}{3}, \qquad f(x) = 0 \iff x = \frac{1}{2}.}$$

The other observations have larger signed scores, so every hard-margin constraint is satisfied.

Solution 1(b): Support Vectors

For $f(x) = (2x - 1)/3$, compute each signed score $y_i f(x_i)$:

Case	A	B	C	D
$f(x_i)$	-3	-1	1	3
$y_i f(x_i)$	3	$\boxed{1}$	$\boxed{1}$	3

The active constraints are those with $y_i f(x_i) = 1$.

Support vectors: B and C.

Cases A and D lie beyond the two margin planes.

Solution 1(b): Margin Width

The weight is $w = 2/3$. In one dimension,

$$\|w\| = |w| = \frac{2}{3}.$$

The **full geometric margin width** is

$$\boxed{\frac{2}{\|w\|} = \frac{2}{2/3} = 3.}$$

The boundary is at $x = 1/2$; the support vectors are at $x = -1$ and $x = 2$.

$$\boxed{\text{Distance to either support vector} = \frac{3}{2}.}$$

Solution 1(c): Active Constraints

Complementary slackness requires

$$\alpha_i[y_i f(x_i) - 1] = 0, \qquad \alpha_i \geq 0.$$

The signed scores are $(3, 1, 1, 3)$. Therefore,

$$\boxed{\alpha_A = \alpha_D = 0.}$$

Only B, C can have nonzero multipliers.
Because $y_B = -1$ and $y_C = +1$, the dual equality becomes

$$-\alpha_B + \alpha_C = 0 \quad \Longrightarrow \quad \boxed{\alpha_B = \alpha_C = \alpha}.$$

Solution 1(c): Multiplier Values

Use the stationarity relation

$$w = \sum_i \alpha_i y_i x_i.$$

Only cases B, C have nonzero multipliers:

$$\begin{aligned} \frac{2}{3} = w &= \alpha_B(-1)(-1) + \alpha_C(+1)(2) \\ &= \alpha + 2\alpha = 3\alpha. \end{aligned}$$

Thus $\boxed{\alpha = 2/9}$, giving

$$\boxed{(\alpha_A, \alpha_B, \alpha_C, \alpha_D) = \left(0, \frac{2}{9}, \frac{2}{9}, 0\right).}$$

All multipliers are nonnegative and satisfy the dual equality.

Solution 1(d): A New Boundary Case

Add $E = (1, -1)$. The closest opposite-class readings are now E at $x = 1$ and C at $x = 2$.

$$-f(1) = -w - b = 1, \qquad f(2) = 2w + b = 1.$$

Solving gives

$$\boxed{w = 2, \qquad b = -3, \qquad x_{\text{boundary}} = \frac{3}{2}.}$$

The full margin shrinks from 3 to

$$\boxed{\frac{2}{|w|} = \frac{2}{2} = 1.}$$

A new case near the other class can sharply reduce the margin.
This motivates allowing limited margin violations.

Task 2: Soft-Margin Model Comparison

Four inspection readings have labels $y \in \{-1, +1\}$:

Case	A	B	C	D
x	-3	-1	1	2
y	-1	-1	$+1$	-1

For $f(x) = wx$ ($b = 0$), use

$$\xi_i = \max(0, 1 - y_i f(x_i)), \qquad J_C(f) = \tfrac{1}{2}w^2 + C \sum_i \xi_i.$$

Compare $f_0(x) = 0$, $f_1(x) = x$, and $f_{1/3}(x) = x/3$.

(a) Find all hinge losses and their sums for f_0 and f_1.

(b) Compare $J_C(f_0), J_C(f_1)$ at $C = 1/4, 1$; find the value of $C > 0$ where they tie.

(c) Find hinge losses and $J_1(f_{1/3})$; rank all three candidates at $C = 1$.

(d) Compare predicted classes of f_1 and $f_{1/3}$; explain their different objectives.

Solution 2(a): Signed Scores and Hinge Loss

For $y \in \{-1, +1\}$, the hinge loss is

$$\xi = \max(0, 1 - yf(x)).$$

	A	B	C	D	Total
$yf_0(x)$	0	0	0	0	–
$\xi(f_0)$	1	1	1	1	$\boxed{4}$
$yf_1(x)$	3	1	1	-2	–
$\xi(f_1)$	0	0	0	3	$\boxed{3}$

Case D is misclassified by f_1 and has hinge loss 3.

Solution 2(b): Regularized Objectives

For f_0, $w = 0$ and $\sum_i \xi_i = 4$.
For f_1, $w = 1$ and $\sum_i \xi_i = 3$.

$$\boxed{J_C(f_0) = 4C, \qquad J_C(f_1) = \frac{1}{2} + 3C.}$$

Penalty C	$J_C(f_0)$	$J_C(f_1)$	Smaller objective
$1/4$	1	$5/4$	f_0
1	4	$7/2$	f_1

A higher C gives more weight to the difference in hinge loss.

Solution 2(b): Penalty Breakpoint

The candidate objectives tie when

$$4C = \frac{1}{2} + 3C \quad \Longrightarrow \quad \boxed{C^* = \frac{1}{2}}.$$

For these two candidates:

Range of C	Preferred candidate
$0 < C < 1/2$	f_0
$C = 1/2$	tie
$C > 1/2$	f_1

This comparison does not prove global soft-margin optimality.

Solution 2(c): Hinge-Loss Calculation

For $f_{1/3}(x) = x/3$, the signed scores are

$$\left(1, \frac{1}{3}, \frac{1}{3}, -\frac{2}{3}\right).$$

Apply $\xi_i = \max(0, 1 - y_i f(x_i))$:

$$\boxed{(\xi_A, \xi_B, \xi_C, \xi_D) = \left(0, \frac{2}{3}, \frac{2}{3}, \frac{5}{3}\right).}$$

The total is

$$\boxed{\sum_i \xi_i = 0 + \frac{2}{3} + \frac{2}{3} + \frac{5}{3} = 3.}$$

Solution 2(c): Objective Comparison

The third candidate has $w = 1/3$ and total hinge loss 3.
At $C = 1$,

$$J_1(f_{1/3}) = \frac{1}{2}\left(\frac{1}{3}\right)^2 + 3$$
$$= \boxed{\frac{55}{18}}.$$

From part (b),

$$J_1(f_1) = \frac{7}{2}, \qquad J_1(f_0) = 4.$$

Therefore,

$$\boxed{\frac{55}{18} < \frac{7}{2} < 4.}$$

Smallest of the three objectives: $f_{1/3}$.
This only compares the specified candidates.

Solution 2(d): Predicted Classes

Because $f_{1/3}(x) = f_1(x)/3$ and all training inputs are nonzero, both classifiers have the same signs.

Case	A	B	C	D
True label	-1	-1	$+1$	-1
Both predictions	-1	-1	$+1$	$+1$

Both make one classification error, at case D.

Solution 2(d): Regularization Effect

Both candidates have total hinge loss 3, but their weight penalties differ:

$$\frac{1}{2}\left(\frac{1}{3}\right)^2 = \boxed{\frac{1}{18}} \quad < \quad \frac{1}{2}(1)^2 = \boxed{\frac{1}{2}}.$$

Thus, at $C = 1$,

$$\boxed{J_1(f_{1/3}) = \frac{55}{18} < \frac{7}{2} = J_1(f_1).}$$

Conclusion: Equal predicted labels and equal error counts need not give the same SVM objective.

Task 3: Nonlinear Classification with Kernels

Training observations and labels:

Case	A	B	C	D
x	-2	-1	1	2
y	$+1$	-1	-1	$+1$

Use $K(x, z) = (2 + xz)^2$. In part (c), take $\alpha_i = 1/9$ for each case and SVM intercept $b = -5/3$.

(a) Explain the failure of a linear threshold. Find $f(x) = ax^2 + b$ with $y_i f(x_i) = 1$ for all cases.

(b) Find a 3-D map ϕ for K; compute $K(-2, 1)$ and $K(1, 2)$.

(c) Simplify $\sum_i \alpha_i y_i K(x_i, x) + b$; predict at $x = 0, 3$.

(d) Compare $K_1(0, 2)$ and $K_2(0, 2)$ for $K_\ell(x, z) = e^{-(x-z)^2/(2\ell^2)}$; interpret ℓ.

Solution 3(a): Linear Separability

The labels, ordered by the input, are

x	-2	-1	1	2
y	$+1$	-1	-1	$+1$

A linear score $wx + b$ changes sign at most once.
Here, labels change from positive to negative and back to positive.

No single linear threshold separates these cases.

The squared feature x^2 is equal for x and $-x$, so it can distinguish inner from outer readings.

Solution 3(a): Quadratic Separator

Let $f(x) = ax^2 + b$. Set the signed training scores to 1.
For the positive outer points, $x = \pm 2$:

$$4a + b = 1.$$

For the negative inner points, $x = \pm 1$:

$$a + b = -1.$$

Subtracting gives $3a = 2$, hence

$$\boxed{a = \frac{2}{3}, \quad b = -\frac{5}{3}, \quad f(x) = \frac{2x^2 - 5}{3}.}$$

Check: $f(\pm 2) = 1$ and $f(\pm 1) = -1$.

Solution 3(b): Feature-Map Expansion

Expand the polynomial kernel:

$$K(x, z) = (2 + xz)^2 = 4 + 4xz + x^2 z^2.$$

One 3-dimensional feature map is

$$\boxed{\phi(x) = \begin{pmatrix} 2 \\ 2x \\ x^2 \end{pmatrix}.}$$

Indeed,

$$\phi(x)^\top \phi(z) = 4 + 4xz + x^2 z^2 = K(x, z).$$

Solution 3(b): Kernel Values

Evaluate the two kernel values directly:

$$\boxed{K(-2,1) = (2-2)^2 = 0.}$$

$$\boxed{K(1,2) = (2+2)^2 = 16.}$$

The inner products in the three-dimensional feature space can be obtained without explicitly forming the vectors $\phi(x)$.

Solution 3(c): Kernel Prediction Formula

Use $\alpha_i = 1/9$, labels $(+1, -1, -1, +1)$, and the kernel intercept $b = -5/3$.

$$\begin{aligned} f(x) &= \sum_i \alpha_i y_i K(x_i, x) + b \\ &= \frac{1}{9}\big[K(-2, x) - K(-1, x) \\ &\qquad - K(1, x) + K(2, x)\big] - \frac{5}{3}. \end{aligned}$$

Pair the outer and inner observations to simplify the score.

Solution 3(c): Symmetry Simplification

The kernel values in the two symmetric pairs satisfy

$$K(-2, x) + K(2, x) = 8 + 8x^2,$$
$$K(-1, x) + K(1, x) = 8 + 2x^2.$$

Their difference is $6x^2$. Substitute into the score:

$$f(x) = \frac{1}{9}(6x^2) - \frac{5}{3}$$
$$= \boxed{\frac{2x^2 - 5}{3}}.$$

This agrees with the quadratic score from part (a).

Solution 3(c): New Sensor Predictions

For the fitted score $f(x) = (2x^2 - 5)/3$:

$$f(0) = -\frac{5}{3} < 0 \qquad \Longrightarrow \qquad \boxed{\widehat{y}(0) = -1},$$

$$f(3) = \frac{18 - 5}{3} = \frac{13}{3} > 0 \qquad \Longrightarrow \qquad \boxed{\widehat{y}(3) = +1}.$$

The polynomial kernel predicts through a nonlinear quadratic boundary without explicitly computing the mapped features.

Solution 3(d): RBF Similarities

The RBF kernel is

$$K_\ell(x, z) = \exp\left(-\frac{(x-z)^2}{2\ell^2}\right).$$

The squared distance from 0 to 2 is 4.

$$\boxed{K_1(0,2) = e^{-2}, \qquad K_2(0,2) = e^{-1/2}.}$$

Because $e^{-2} < e^{-1/2}$, a larger length scale assigns greater similarity to the same two inputs.

Solution 3(d): Length-Scale Interpretation

For two distinct inputs, the length scale ℓ controls how quickly similarity falls with distance.

Small ℓ : rapid decay; more local influence,

Large ℓ : slower decay; broader influence.

For either length scale, self-similarity is always

$$\boxed{K_\ell(x, x) = 1.}$$

The kernel width therefore controls the spatial scale of interactions, not the labels themselves.

Tutorial 11: margins, regularization, and nonlinear kernels

Questions and discussion