MA477: Data Science

Lesson 14 Outline — 12 February 2026

United States Military Academy, West Point
Instructor: MAJ Patrick Kuiper

1 Administrative

- Quiz 3
- Student review
- LDA / SVM Lecture

2 Support Vector Machine Lesson Objectives

- Understand separating hyperplanes and the maximum margin classifier.
- Understand the details of the support vector classifier (ISLP equations 9.12-9.15).
- Use sklearn modules to assess support vector machine (SVM) classifiers.

Discussion: From Hyperplanes to the Maximal Margin Classifier

Define and describe a hyperplane in two dimensional space?

In two dimensions, a hyperplane is simply a line defined by

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0.$$

In p dimensions, this generalizes to

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0.$$

This equation defines a flat $(p-1)$-dimensional surface in $\mathbb{R}^p$.
For any point X:

- If the left-hand side is greater than zero, the point lies on one side of the hyperplane.
- If it is less than zero, the point lies on the other side.
- If it equals zero, the point lies exactly on the hyperplane.

Thus classification can be based on the sign of

$$f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j.$$

How Can a Hyperplane Be Used for Classification?

Suppose we observe training data $x_1, \dots, x_n \in \mathbb{R}^p$ with class labels $y_i \in \{-1, 1\}$.
A separating hyperplane must satisfy:

$$\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} > 0 \quad \text{if } y_i = 1,$$

$$\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} < 0 \quad \text{if } y_i = -1.$$

These two conditions can be combined into the single requirement

$$y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) > 0,$$

which ensures correct classification for all observations.

If Many Separating Hyperplanes Exist, Which One Should We Choose?

If the data are separable, there are infinitely many hyperplanes satisfying the separation condition.

To choose among them, we introduce the **margin**.

The margin is the smallest perpendicular distance from the hyperplane to any training observation.

Intuitively:

- A large margin means the hyperplane is far from all points.
- A small margin means at least one point is close to the boundary.

We therefore seek the hyperplane that maximizes this minimum distance.

How Do We Express the Margin Mathematically?

The perpendicular distance from a point x_i to the hyperplane is

$$\frac{\left| \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right|}{\sqrt{\sum_{j=1}^{p} \beta_j^2}}.$$

Notice that if we multiply all coefficients by a constant c, the hyperplane does not change, because

$$c(\beta_0 + \sum_{j=1}^{p} \beta_j X_j) = 0$$

defines the same separating surface.

Therefore, the scale of β is arbitrary unless we fix it.

To remove this ambiguity, we impose the normalization constraint

$$\sum_{j=1}^{p} \beta_j^2 = 1.$$

This is the first constraint in the maximal margin optimization problem.

Why is this constraint necessary?

Without it, we could multiply all coefficients by a very large constant and artificially inflate the value of the margin. The normalization fixes the length of the coefficient vector β, ensuring that the quantity

$$y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right)$$

represents the true geometric distance from the hyperplane.

Under this normalization, the perpendicular distance simplifies to

$$y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right).$$

What Is the Maximal Margin Optimization Problem?

We now choose $\beta_0, \beta_1, \ldots, \beta_p$, and M to maximize the margin M subject to two constraints.

Maximal Margin Classifier

$$\text{maximize}_{\beta_0,\beta_1,\ldots,\beta_p,M} \quad M$$

subject to

$$\sum_{j=1}^{p} \beta_j^2 = 1,$$

and

$$y_i \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \right) \geq M \quad \text{for all } i = 1, \ldots, n.$$

Interpretation

- The second constraint ensures that all observations are correctly classified and lie at least distance M from the hyperplane.
- The first constraint removes scaling ambiguity and gives geometric meaning to M.
- Maximizing M yields the widest possible separating slab between the two classes.
- Observations that satisfy the constraint at equality are called **support vectors**, since they determine the position of the hyperplane.