02-750 Week 5

Automation of Scientific Research

Aidan Jan

February 10, 2026

Linear Classifiers

- Let $x \in \mathbb{R}^n$ be a point in an n-dimensional space
- Let $w \in \mathbb{R}^n$ be a weight vector
- Let $b \in \mathbb{R}$ be a bias term
- The set of points satisfying $w \cdot x + b = 0$ form a **hyperplane**
- A hyperplane can be used as a binary classifier for instance x by simply determining whether $w \cdot x + b < 0$ or $w \cdot x + b \geq 0$

There are many different algorithms for learning linear classifiers

- Ex. Naive Bayes; LDA; Logistic Regression; SVMs

Support Vector Machines (SVMs)

- SVM learning algorithms find the hyperplane that maximizes the **margin** between the two classes (aka structural risk minimization)
  - The margin is distance between hyperplane and the nearest training points
  - The points achieving this distance are called the support vectors

For SVMs,

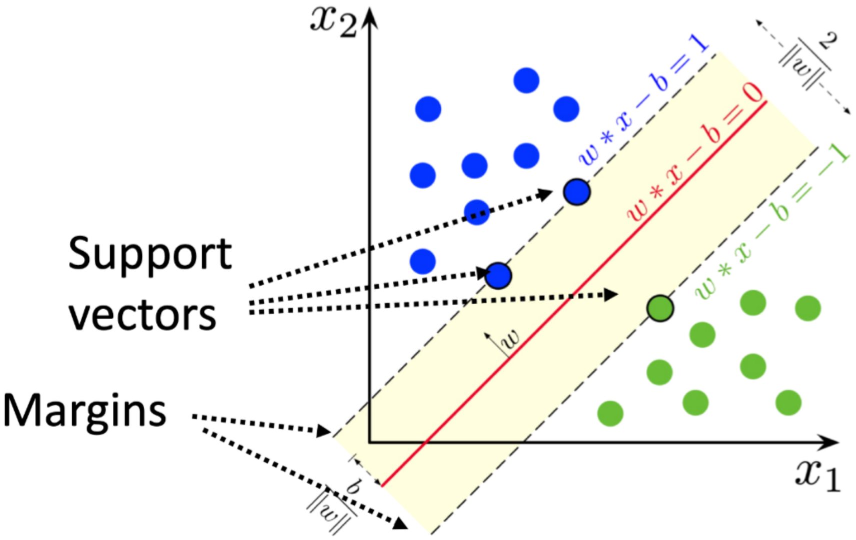
- Hypothesis class $\mathcal{H}$: linear classifiers
- Loss function: **Hinge Loss**

$$\updownarrow(x_i; w, b) = \max(0, 1 - y \cdot (w \cdot x_i - b))$$

- Risk objective: structural risk minimization

$$argmin_{w,b} = \frac{!}{n} \sum \updownarrow(x_i; w, b) + \lambda \|w\|^2$$

- Search function: quadratic programming



- Given the hyperplane, the **signed distance** between point x to the plane is:

$$\frac{w \cdot x + b}{\|w\|_2}$$

- The **unsigned distance** is:

$$\left| \frac{w \cdot x + b}{\|w\|_2} \right|$$

Experimental Set-up

- Source data: DuPont Thrombin data from KDD 2001
  - Binary features: $x \in \{0, 1\}^d$, where $d = 139, 351$
- Initialization: random batches until the 1st active compound is found
- Batch query selection
  - Batch size: 5% of the unlabeled samples
  - 4 query selection strategies were used (see next slide)
- Two rounds of experiments were performed (Rounds 0 and 1)
  - Pool size: *Round 0*: 1316 (40 active); *Round 1*: 643 (150 active)
- Experiments are repeated 10 times to compute average performance

Query Selection Strategies

- Random Sampling
- Proximity-Based
  - i.e., points closest to known active compounds
- Select points with the largest positive distance from hyperplane
  - Assuming positive distance means active
- Near boundary selection (i.e., margin sampling).