IAML: Support Vector Machines

Nigel Goddard and Victor Lavrenko
School of Informatics
Semester 1

Overview

Support vector machines are one of the most effective and widely used classification algorithms.

- SVMs are the combination of two ideas
  - Maximum margin classification
  - The “kernel trick”
- SVMs are a linear classifier, like logistic regression and perceptron

Stuff You Need to Remember

\( w^T x \) is length of the projection of \( x \) onto \( w \) (if \( w \) is a unit vector)

i.e., \( b = w^T x \).

(If you do not remember this, see supplementary maths notes on course Web site.)
For any linear classifier

- Training instances \((x_i, y_i), i = 1, \ldots, n, y_i \in \{-1, +1\}\)
- Hyperplane \(w^\top x + w_0 = 0\)
- Notice for this lecture we use \(-1\) rather than \(0\) for negative class. This will be convenient for the maths.

Idea: Maximize the Margin

The **margin** is the distance between the decision boundary (the hyperplane) and the closest training point.

Computing the Margin

- The tricky part will be to get an equation for the margin
- We’ll start by getting the distance from the origin to the hyperplane
- i.e., We want to compute the scalar \(b\) below

\[
 w^\top x + w_0 = 0
\]
Computing the Distance to Origin

▶ Define z as the point on the hyperplane closest to the origin.
▶ z must be proportional to w, because w normal to hyperplane
▶ By definition of b, we have the norm of z given by:
\[ \|z\| = b \]
So
\[ b \frac{w}{\|w\|} = z \]

Computing the Distance to Hyperplane

▶ Now we want c, the distance from x to the hyperplane.
▶ It’s clear that \( c = \|b - a\| \), where a the length of the projection of x onto w. Quiz: What is a?

Computing the Distance to Hyperplane

▶ We know that (a) z on the hyperplane and (b) \( b \frac{w}{\|w\|} = z \).
▶ First (a) means \( w^T z + w_0 = 0 \)
▶ Substituting we get
\[
\begin{align*}
w^T \frac{bw}{\|w\|} + w_0 &= 0 \\
\frac{bw^T w}{\|w\|} + w_0 &= 0 \\
b &= -\frac{w_0}{\|w\|}
\end{align*}
\]
▶ Remember \( \|w\| = \sqrt{w^T w} \).
▶ Now we have the distance from the origin to the hyperplane!
The perpendicular distance from a point \( x \) to the hyperplane \( \mathbf{w}^T \mathbf{x} + w_0 = 0 \) is
\[
\frac{1}{||\mathbf{w}||} |\mathbf{w}^T \mathbf{x} + w_0|
\]

The **margin** is the distance from the closest training point to the hyperplane
\[
\min_i \frac{1}{||\mathbf{w}||} |\mathbf{w}^T \mathbf{x}_i + w_0|
\]

Note that \((\mathbf{w}, w_0)\) and \((\mathbf{c}\mathbf{w}, cw_0)\) define the same hyperplane. The scale is arbitrary.

This is because we predict class \( y = 1 \) if \( \mathbf{w}^T \mathbf{x} + w_0 \geq 0 \). That's the same thing as saying \( \mathbf{c}\mathbf{w}^T \mathbf{x} + cw_0 \geq 0 \)

To remove this freedom, we will put a constraint on \((\mathbf{w}, w_0)\)
\[
\min_i |\mathbf{w}^T \mathbf{x}_i + w_0| = 1
\]

With this constraint, the margin is always \( 1/||\mathbf{w}|| \).

Here is a first version of an optimization problem to maximize the margin (we will simplify)
\[
\max_{\mathbf{w}} \frac{1}{||\mathbf{w}||}
\]
subject to \( \mathbf{w}^T \mathbf{x}_i + w_0 \geq 0 \) for all \( i \) with \( y_i = 1 \)
\( \mathbf{w}^T \mathbf{x}_i + w_0 \leq 0 \) for all \( i \) with \( y_i = -1 \)
\( \min_i |\mathbf{w}^T \mathbf{x}_i + w_0| = 1 \)

The first two constraints are too loose. It's the same thing to say
\[
\max_{\mathbf{w}} \frac{1}{||\mathbf{w}||}
\]
subject to \( \mathbf{w}^T \mathbf{x}_i + w_0 \geq 1 \) for all \( i \) with \( y_i = 1 \)
\( \mathbf{w}^T \mathbf{x}_i + w_0 \leq -1 \) for all \( i \) with \( y_i = -1 \)
\( \min_i |\mathbf{w}^T \mathbf{x}_i + w_0| = 1 \)

Now the third constraint is redundant

That means we can simplify to
\[
\max_{\mathbf{w}} \frac{1}{||\mathbf{w}||}
\]
subject to \( \mathbf{w}^T \mathbf{x}_i + w_0 \geq 1 \) for all \( i \) with \( y_i = 1 \)
\( \mathbf{w}^T \mathbf{x}_i + w_0 \leq -1 \) for all \( i \) with \( y_i = -1 \)

Here's a compact way to write those two constraints
\[
\max_{\mathbf{w}} \frac{1}{||\mathbf{w}||}
\]
subject to \( y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1 \) for all \( i \)

Finally, note that maximizing \( 1/||\mathbf{w}|| \) is the same thing as minimizing \( ||\mathbf{w}||^2 \)
So the SVM weights are determined by solving the optimization problem:

\[
\min_w ||w||^2 \\
\text{s.t. } y_i (w^\top x_i + w_0) \geq +1 \quad \text{for all } i
\]

Solving this will require maths that we don’t have in this course. But I’ll show the form of the solution next time.