CS 335: Support Vector Machines

Dan Sheldon

November 18, 2014

Support Vector Machines (SVMs) Overview

- Linear classifier, **non-linear with "kernel trick"**
- Among the best out-of-box classifiers
- Geometric principles: separating hyperplanes, margins

Alert: Notation Change

- Input: $\mathbf{x} \in \mathbb{R}^n$
- Output: $y \in \{-1, +1\}$
- Hypothesis: $h_{\mathbf{w},b}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$

"Bias term" b treated explicitly (don't add 1 to feature vector)

"Weights" $\mathbf{w}$

Separating Hyperplanes

Recall linear classifier:

$$\begin{aligned} \mathbf{w}^T\mathbf{x} + b < 0 &\quad \Rightarrow \quad \text{predict } -1 \\ \mathbf{w}^T\mathbf{x} + b \geq 0 &\quad \Rightarrow \quad \text{predict } +1 \end{aligned}$$

Assume for now training data is "linearly separable" $\rightarrow$ there is some $\mathbf{w}, b$ that separates positive training examples from negative training examples

$$\begin{aligned} \mathbf{w}^T\mathbf{x}^{(i)} + b < 0 &\quad \text{for} \quad y^{(i)} = -1 \\ \mathbf{w}^T\mathbf{x}^{(i)} + b \geq 0 &\quad \text{for} \quad y^{(i)} = +1 \end{aligned}$$

Picture of training data / separating hyperplane

Margin

Of all separating hyperplanes, which one will lead to best generalization performance?

Intution/illustration about margins

Margin = distance from hyperplane to closest training example

Choose hyperplane $(\mathbf{w}, b)$ to maximize the margin

SVM Optimization Problem

"Find $\mathbf{w}$, b to minimize..."

$$\begin{aligned} &\min_{\mathbf{w},b} \ \frac{1}{2}||\mathbf{w}||^2 \\ \text{subject to} \quad &\mathbf{w}^T\mathbf{x}^{(i)} + b \leq -1 \qquad \text{if } y^{(i)} = -1 \\ &\mathbf{w}^T\mathbf{x}^{(i)} + b \geq +1 \qquad \text{if } y^{(i)} = +1 \end{aligned}$$

Write on board for discussion

(Note: not yet obvious how this maximizes margin...)

Aside: Constrained Optimization

- Constrained optimization
  - Objective function, constraints
  - Assume black-box solver for now

Geometric Interpretation of SVM

MATLAB demo

Good/bad contour plots

Geometric Interpretation Recap

Rewrite problem using functional margin $y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b)$

$$\min_{w,b} \; \frac{1}{2}||\mathbf{w}||^2$$
$$\text{subject to} \quad y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 \quad \text{for all } i$$

- All examples have functional margin at least one: correctly classified "and more"
- On correct side, and at least one contour from decision boundary

Minimize slope/complexity subject to functional margin constraint

Why does this maximize the margin?

Sketch argument on board. First 1D, then 2D

Argument recap

Let $\mathbf{x}^{(i)}$ be training example that is closest to margin. Assume that $y^{(i)} = 1$.

Claim: $\mathbf{w}^T\mathbf{x}^{(i)} + b = 1$
Proof sketch: We know $\mathbf{w}^T\mathbf{x}^{(i)} + b$ is *at least* 1. If it is bigger, shrink $\mathbf{w}$ (multiply by some $\alpha < 1$) until it is exactly 1.

Let γ be the margin, which is the length of the line segment between $\mathbf{x}^{(i)}$ and the closest point on the decision boundary. By our claim, the change in function value along the line segment is one. Thus, the slope along the line segment is

$$\frac{\text{rise}}{\text{run}} = \frac{1}{\gamma}$$

Argument recap

Because the segment connects $\mathbf{x}^{(i)}$ to the *closest* point on the decision boundary, it follows the steepest descent direction and has slope $||w||$ (the gradient/slope of the function $\mathbf{w}^T\mathbf{x} + b$).

So we have two expressions for the slope:

$$||w|| = \frac{1}{\gamma}$$

Hence, by minimizing $||\mathbf{w}||$ (subject to the constraints), we are maximizing the margin γ.

Soft-Margin SVMs

What if training data is not linearly separable?

"Soft margin": allow functional margins that are not big enough, but add a penalty for this in the objective function

$$\begin{aligned}
&\min_{w,b,\xi} \; \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m}\xi_i \\
\text{subject to} \quad & y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - \xi_i \quad \text{for all } i \\
& \xi_i \geq 0, \; i = 1, \ldots, m
\end{aligned}$$

1D picture on board. Revisit MATLAB demo

Summary / What's next

Summary

- Linearly separable data and margins
- "Hard-margin" SVM
  - Constrained optimization
  - Functional margins
  - Why it maximizes the (geometric) margin
- Soft-margin SVMs

What's next

- "Kernel trick" $\rightarrow$ non-linearity
- Connection to logistic regression