CSC 411: Lecture 15: Support Vector Machine

Class based on Raquel Urtasun & Rich Zemel's lectures

Sanja Fidler

University of Toronto

March 18, 2016

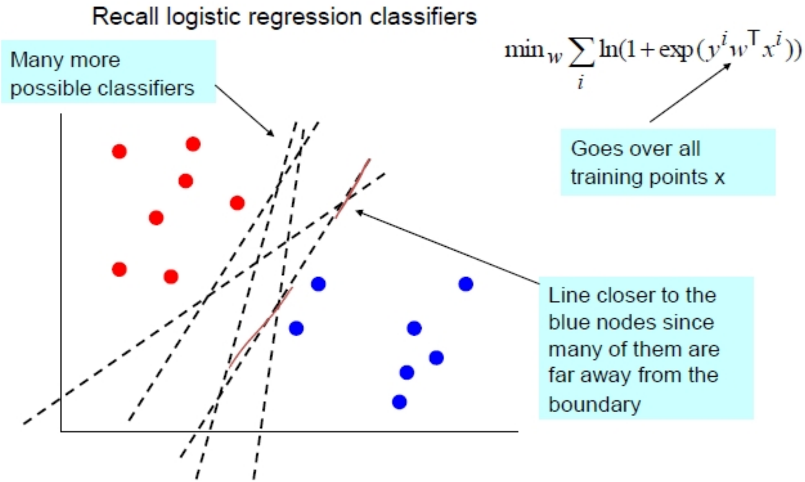
Today

- Margin
- Max-margin classification

Today

- We are back to **supervised** learning
- We are given training data $\{(\mathbf{x}^{(i)}, t^{(i)})\}_{i=1}^{N}$
- We will look at **classification**, so $t^{(i)}$ will represent the class label
- We will focus on **binary** classification (two classes)
- We will consider a **linear** classifier first (next class non-linear decision boundaries)
- Tiny change from before: instead of using $t = 1$ and $t = 0$ for positive and negative class, we will use $t = 1$ for the positive and $t = -1$ for the negative class

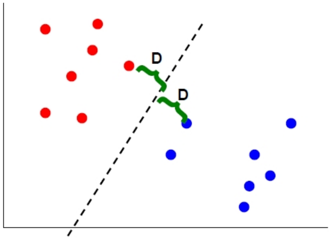
Logistic Regression



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

Max Margin Classification

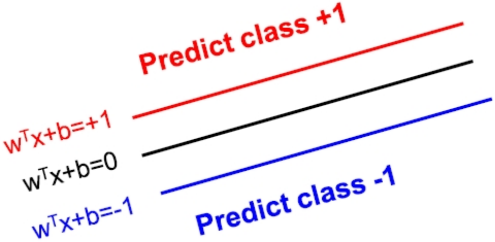
- Instead of fitting all the points, focus on boundary points
- Aim: learn a boundary that leads to the largest margin (buffer) from points on both sides



- Why: intuition; theoretical support; and works well in practice
- Subset of vectors that support (determine boundary) are called the support vectors

Linear SVM

- Max margin classifier: inputs in margin are of unknown class

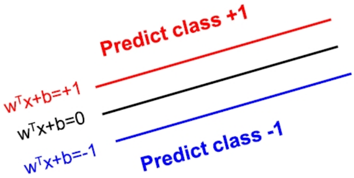


$$y = \begin{cases} 1 & \text{if } \mathbf{w}^T\mathbf{x} + b \geq 1 \\ -1 & \text{if } \mathbf{w}^T\mathbf{x} + b \leq -1 \\ \text{Undefined} & \text{if } -1 \leq \mathbf{w}^T\mathbf{x} + b \leq 1 \end{cases}$$

- Can write above condition as:

$$(\mathbf{w}^T\mathbf{x} + b)y \geq 1$$

Geometry of the Problem



- The vector $\mathbf{w}$ is orthogonal to the $+1$ plane.
  If $\mathbf{u}$ and $\mathbf{v}$ are two points on that plane, then

$$\mathbf{w}^T(\mathbf{u} - \mathbf{v}) = 0$$

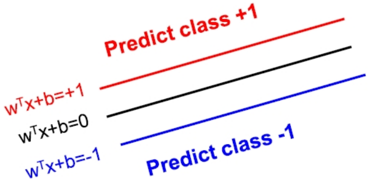
- Same is true for -1 plane
- Also: for point $\mathbf{x}_+$ on $+1$ plane and $\mathbf{x}_-$ nearest point on -1 plane:

$$\mathbf{x}_+ = \lambda \mathbf{w} + \mathbf{x}_-$$

Computing the Margin

- Also: for point $\mathbf{x}_+$ on $+1$ plane and $\mathbf{x}_-$ nearest point on -1 plane:

$$\mathbf{x}_+ = \lambda \mathbf{w} + \mathbf{x}_-$$



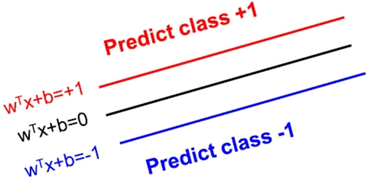
$$\begin{aligned} \mathbf{w}^T\mathbf{x}_+ + b &= 1 \\ \mathbf{w}^T(\lambda\mathbf{w} + \mathbf{x}_-) + b &= 1 \\ \mathbf{w}^T\mathbf{x}_- + b + \lambda\mathbf{w}^T\mathbf{w} &= 1 \\ -1 + \lambda\mathbf{w}^T\mathbf{w} &= 1 \end{aligned}$$

Therefore

$$\lambda = \frac{2}{\mathbf{w}^T\mathbf{w}}$$

Computing the Margin

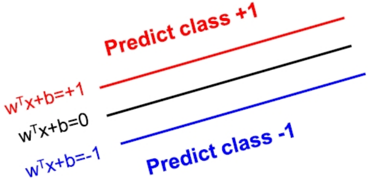
- Define the margin M to be the distance between the $+1$ and -1 planes
- We can now express this in terms of **w** to maximize the margin we minimize the length of **w**



$$
\begin{aligned}
M &= ||\mathbf{x}_+ - \mathbf{x}_-|| \\
&= ||\lambda \mathbf{w}|| = \lambda\sqrt{\mathbf{w}^T\mathbf{w}} \\
&= 2\frac{\sqrt{\mathbf{w}^T\mathbf{w}}}{\mathbf{w}^T\mathbf{w}} = \frac{2}{\sqrt{\mathbf{w}^T\mathbf{w}}} = \frac{2}{||\mathbf{w}||}
\end{aligned}
$$

Learning a Margin-Based Classifier

- We can search for the optimal parameters ($\mathbf{w}$ and b) by finding a solution that:
  1. Correctly classifies the training examples: $\{(\mathbf{x}^{(i)}, t^{(i)})\}_{i=1}^{N}$
  2. Maximizes the margin (same as minimizing $\mathbf{w}^T\mathbf{w}$)



$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2$$

$$s.t. \forall i \quad (\mathbf{w}^T\mathbf{x}^{(i)} + b)t^{(i)} \geq 1,$$

- This is called the primal formulation of Support Vector Machine (SVM)
- Can optimize via projective gradient descent, etc.
- Apply Lagrange multipliers: formulate equivalent problem

Learning a Linear SVM

- Convert the constrained minimization to an unconstrained optimization problem: represent constraints as penalty terms:

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + \text{penalty\_term}$$

- For data $\{(\mathbf{x}^{(i)}, t^{(i)})\}_{i=1}^N$, use the following penalty

$$\max_{\alpha_i \geq 0} \quad \alpha_i[1 - (\mathbf{w}^T\mathbf{x}^{(i)} + b)t^{(i)}] = \begin{cases} 0 & \text{if} \quad (\mathbf{w}^T\mathbf{x}^{(i)} + b)t^{(i)} \geq 1 \\ \infty & \text{otherwise} \end{cases}$$

- Rewrite the minimization problem

$$\min_{\mathbf{w},b}\{\frac{1}{2}||\mathbf{w}||^2 + \sum_{i=1}^N \max_{\alpha_i \geq 0} \alpha_i[1 - (\mathbf{w}^T\mathbf{x}^{(i)} + b)t^{(i)}]\}$$

where α_i are the Lagrange multipliers

$$= \min_{\mathbf{w},b} \max_{\alpha_i \geq 0}\{\frac{1}{2}||\mathbf{w}||^2 + \sum_{i=1}^N \alpha_i[1 - (\mathbf{w}^T\mathbf{x}^{(i)} + b)t^{(i)}]\}$$

Solution to Linear SVM

- Let:

$$J(\mathbf{w}, b; \alpha) = \frac{1}{2}||\mathbf{w}||^2 + \sum_{i=1}^{N} \alpha_i [1 - (\mathbf{w}^T \mathbf{x}^{(i)} + b) t^{(i)}]$$

- Swap the "max" and "min": This is a lower bound

$$\max_{\alpha_i \geq 0} \min_{\mathbf{w}, b} J(\mathbf{w}, b; \alpha) \leq \min_{\mathbf{w}, b} \max_{\alpha_i \geq 0} J(\mathbf{w}, b; \alpha)$$

- Equality holds in certain conditions

Solution to Linear SVM

- Solving:

$$\max_{\alpha_i \geq 0} \min_{\mathbf{w},b} J(\mathbf{w}, b; \alpha) = \max_{\alpha_i \geq 0} \min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + \sum_{i=1}^{N} \alpha_i [1 - (\mathbf{w}^T \mathbf{x}^{(i)} + b) t^{(i)}]$$

- First minimize $J()$ w.r.t. $\mathbf{w}, b$ for fixed Lagrange multipliers:

$$\begin{aligned} \frac{\partial J(\mathbf{w}, b; \alpha)}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{N} \alpha_i \mathbf{x}^{(i)} t^{(i)} = 0 \\ \frac{\partial J(\mathbf{w}, b; \alpha)}{\partial b} &= -\sum_{i=1}^{N} \alpha_i t^{(i)} = 0 \end{aligned}$$

- We obtain

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i t^{(i)} \mathbf{x}^{(i)}$$

- Then substitute back to get final optimization:

$$L = \max_{\alpha_i \geq 0} \{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} t^{(i)} t^{(j)} \alpha_i \alpha_j (\mathbf{x}^{(i)^T} \cdot \mathbf{x}^{(j)}) \}$$

Summary of Linear SVM

- Binary and linear separable classification
- Linear classifier with maximal margin
- Training SVM by maximizing

$$\max_{\alpha_i \geq 0} \{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} t^{(i)} t^{(j)} \alpha_i \alpha_j (\mathbf{x}^{(i)^T} \cdot \mathbf{x}^{(j)}) \}$$

$$\text{subject to} \quad \alpha_i \geq 0; \quad \sum_{i=1}^{N} \alpha_i t^{(i)} = 0$$

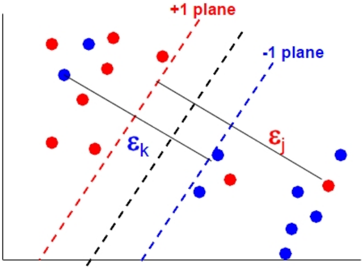
- The weights are

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i t^{(i)} \mathbf{x}^{(i)}$$

- Only a small subset of α_i's will be nonzero, and the corresponding $\mathbf{x}^{(i)}$'s are the support vectors $\mathbf{S}$
- Prediction on a new example:

$$y = \text{sign}[b + \mathbf{x} \cdot (\sum_{i=1}^{N} \alpha_i t^{(i)} \mathbf{x}^{(i)})] = \text{sign}[b + \mathbf{x} \cdot (\sum_{i \in \mathbf{S}} \alpha_i t^{(i)} \mathbf{x}^{(i)})]$$

What if data is not linearly separable?



- Introduce slack variables ξ_i

$$\min \frac{1}{2}||\mathbf{w}||^2 + \lambda \sum_{i=1}^{N} \xi_i$$

$$\text{s.t} \quad \xi_i \geq 0; \quad \forall i \quad t^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)} + b) \geq 1 - \xi_i$$

- Example lies on wrong side of hyperplane $\xi_i > 1$
- Therefore $\sum_i \xi_i$ upper bounds the number of training errors
- λ trades off training error vs model complexity
- This is known as the soft-margin extension