Machine Learning: Yoshinari Fujinuma

University of Colorado Boulder

LECTURE 18

Slides adapted from Chenhao Tan, Jordan Boyd-Graber, Chris Ketelsen, and Lecture 12 from Andrew Ng's Coursera class

Logistics

- Homework 3 is due on next Monday March 15th
- Final project proposal is due on Friday March 19th

Learning Objectives

- Introduce Support Vector Machine

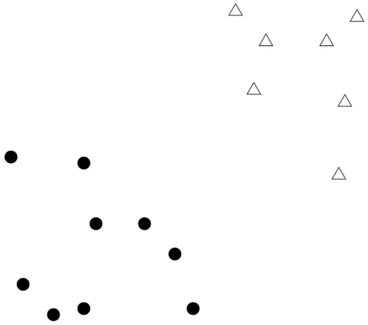
Outline

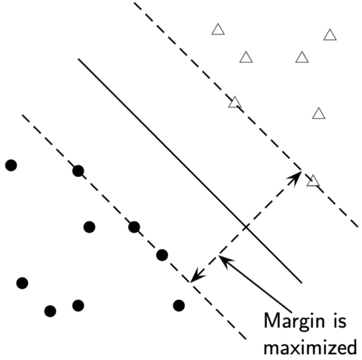
Hard-Margin SVM

Soft-Margin SVM

Support Vector Machines

Assume we want to solve a binary classification problem
Support vector machine is referred to as a **max-margin classifier**





Support Vector Machines

Since a decision boundary is a hyperplane, we classify a given input $\boldsymbol{x}$ by

$$\boldsymbol{w}^T\boldsymbol{x} + b$$

where $\boldsymbol{w}$ is a weight, b is the bias.

- if $\boldsymbol{w}^T\boldsymbol{x} + b \geq 1$, then prediction $\hat{y} = +1$
- if $\boldsymbol{w}^T\boldsymbol{x} + b \leq -1$, then prediction $\hat{y} = -1$

...and we want to maximize the margin.

Optimization Problem for SVM

We want to find a weight vector $\boldsymbol{w}$ and bias b that optimize

$$\max_{\boldsymbol{w},b} \ \textbf{margin}$$

$$\text{subject to } \ \boldsymbol{w}^T\boldsymbol{x}_i + b \geq 1 \text{ if } y_i = 1$$

$$\boldsymbol{w}^T\boldsymbol{x}_i + b \leq -1 \text{ if } y_i = -1$$

given m training examples where i represents each training example, y_i is the gold label of a training example,

So what is **margin**?

Optimization Problem for SVM

So actually **margin** $r = 2/||\boldsymbol{w}||$

Maximizing $r = 2/||\boldsymbol{w}||$ is equivalent to minimizing $||\boldsymbol{w}||$ (and $||\boldsymbol{w}||^2$)

We want to find a weight vector $\boldsymbol{w}$ and bias b that optimize

$$
\begin{aligned}
&\min_{\boldsymbol{w},b} \ \frac{1}{2}||\boldsymbol{w}||^2 \\
\text{subject to } \ &\boldsymbol{w}^T\boldsymbol{x}_i + b \geq 1 \text{ if } y_i = 1 \\
&\boldsymbol{w}^T\boldsymbol{x}_i + b \leq -1 \text{ if } y_i = -1
\end{aligned}
$$

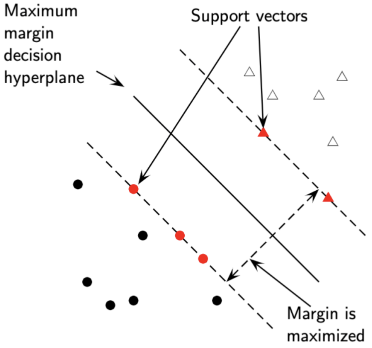
Why is Margin $r = 2/||\boldsymbol{w}||$?

Given two support vectors $\boldsymbol{w}^T\boldsymbol{x}_i + b = 1$ and $\boldsymbol{w}^T\boldsymbol{x}_i + b = -1$,

Let $\boldsymbol{x}^-$ be an example on the support vector $\boldsymbol{w}^T\boldsymbol{x}_i + b = -1$.
Given a unit vector $\frac{\boldsymbol{w}}{||\boldsymbol{w}||}$ (which is perpendicular to the decision hyperplane),

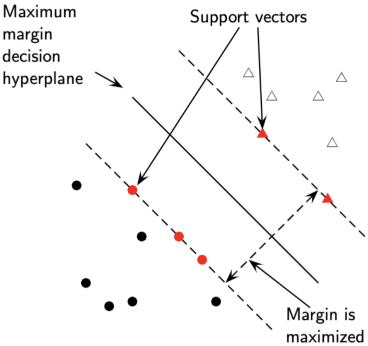
$$\boldsymbol{w}^T(\boldsymbol{x}^- + r\frac{\boldsymbol{w}}{||\boldsymbol{w}||}) + b = 1$$

since moving $\boldsymbol{x}^-$ by the margin r will make $\boldsymbol{x}^-$ be on the other support vector



Why is Margin $r = 2/||w||$?

$$w^T(x^- + r\frac{w}{||w||}) + b = 1$$
$$w^T x^- + r\frac{||w||^2}{||w||} + b = 1$$
$$w^T x^- + r||w|| + b = 1$$
$$-1 + r||w|| = 1$$
$$r||w|| = 2$$
$$r = \frac{2}{||w||}$$

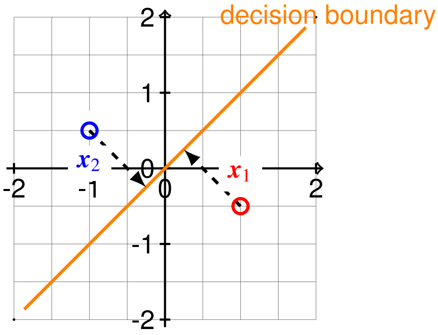


Derivation from `https://math.stackexchange.com/questions/1305925/why-is-the-svm-margin-equal-to-frac2-mathbfw` and `https://nlp.stanford.edu/IR-book/html/htmledition/support-vector-machines-the-linearly-separable-case-1.html`

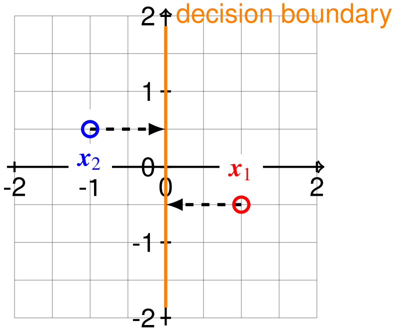
Another view of why minimizing $||w||$

$w^T x = ||w|| p$ where p is a projected vector of x on to the decision boundary

When p is small, $||w||$ is large



When p is large, $||w||$ is small



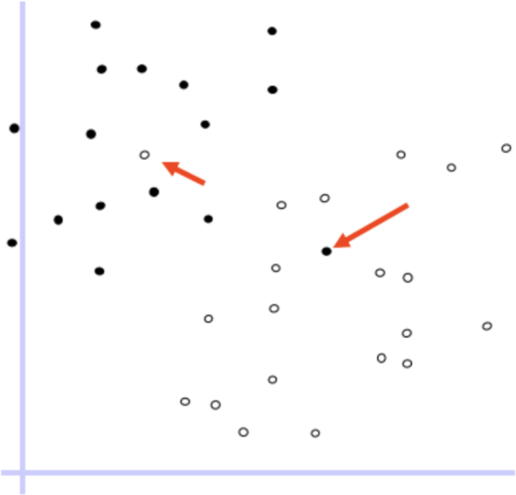
Referred from Andrew Ng's Coursera class Lecture 12 and latex from `https://tex.stackexchange.com/questions/120788/`

Outline

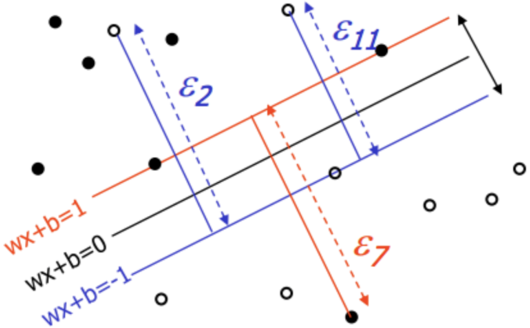
Hard-Margin SVM

Soft-Margin SVM

Can a Hard-Margin SVM work when Outliers Exist?



Allow Outliers by Including Slack Variables ξ_i



New objective function

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i$$

subject to

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i, i \in [1, m]$$
$$\xi_i \geq 0, i \in [1, m]$$

New objective function

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i$$

subject to

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i, i \in [1, m]$$
$$\xi_i \geq 0, i \in [1, m]$$

- Margin

New objective function

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i$$

subject to

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i, i \in [1, m]$$
$$\xi_i \geq 0, i \in [1, m]$$

- Margin
- How wrong a point is (slack variables)

New objective function

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i$$

subject to

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i, i \in [1, m]$$
$$\xi_i \geq 0, i \in [1, m]$$

- Margin
- How wrong a point is (slack variables)
- A hyperparameter which controls the tradeoff between margin and slack variables

New objective function

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i$$

subject to

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i, i \in [1, m]$$
$$\xi_i \geq 0, i \in [1, m]$$

What is ξ_i?

New objective function

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i$$

subject to

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i, i \in [1, m]$$
$$\xi_i \geq 0, i \in [1, m]$$

What is ξ_i?

$$\xi_i = \begin{cases} 0, & \text{if } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 \\ 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b), & \text{otherwise} \end{cases}$$

Hinge loss i.e., $\ell^{(\text{hin})}(y_i, \hat{y}_i) = max(0, 1 - y_i\hat{y}_i)$

Soft-margin SVM

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \ell^{(\text{hin})}(y_i, \boldsymbol{w}^T\boldsymbol{x}_i + b)$$

You can solve this with gradient descent since this is now a unconstrained optimization problem.

Next Lecture

- Problem: Both hard-margin and soft-margin SVMs are still linear classifiers
- Next Lecture: Making SVMs non-linear using “kernels”