Machine Learning

Support Vector Machines

Fabio Vandin

November 20, 2017
Classification and Margin

Consider a classification problem with two classes:

- instance set $\mathcal{X} = \mathbb{R}^d$
- label set $\mathcal{Y} = \{-1, 1\}$.

Training data: $S = ((x_1, y_1), \ldots, (x_m, y_m))$

Hypothesis set $\mathcal{H} = \text{halfspaces}$

**Assumption**: data is linearly separable $\Rightarrow$ there exist a halfspace that perfectly classify the training set

**In general**: multiple separating hyperplanes: $\Rightarrow$ which one is the best choice?
Classification and Margin

The last one seems the best choice, since it can tolerate more “noise”.

Informally, for a given separating halfspace we define its margin as its minimum distance to an example in the training set $S$.

**Intuition:** best separating hyperplane is the one with largest margin.

How do we find it?
Training set $S = ((x_1, y_1), \ldots, (x_m, y_m))$ is *linearly separable* if there exists a halfspace $(w, b)$ such that $y_i = \text{sign}(\langle w, x_i \rangle + b)$ for all $i = 1, \ldots, m$.

Equivalent to:

$$\forall i = 1, \ldots, m : y_i(\langle w, x_i \rangle + b) > 0$$

**Informally:** *margin* of a separating hyperplane is its minimum distance to an example in the training set $S$
Given hyperplane defined by $L = \{v : \langle w, v \rangle + b = 0\}$, and given $x$, the distance of $x$ to $L$ is

$$d(x, L) = \min\{||x - v|| : v \in L\}$$

Claim: if $||w|| = 1$ then $d(x, L) = |\langle w, x \rangle + b|$ (Proof: Claim 15.1 [UML])
Margin and Support Vectors

The margin of a separating hyperplane is the distance of the closest example in training set to it. If $||\mathbf{w}|| = 1$ the margin is:

$$\min_{i \in \{1, ..., m\}} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b|$$

The closest examples are called support vectors.
Support Vector Machine (SVM)

**Hard-SVM**: seek for the separating hyperplane with largest margin (only for linearly separable data)

**Computational problem**:

\[
\arg \max_{(w,b): \|w\|=1} \min_{i \in \{1,\ldots,m\}} \left| \langle w, x_i \rangle + b \right|
\]

subject to \( \forall i : y_i (\langle w, x_i \rangle + b) > 0 \)

**Equivalent formulation** (due to separability assumption):

\[
\arg \max_{(w,b): \|w\|=1} \min_{i \in \{1,\ldots,m\}} y_i (\langle w, x_i \rangle + b)
\]
Hard-SVM: Quadratic Programming Formulation

• input: \((x_1, y_1), \ldots, (x_m, y_m)\)
• solve:

\[
(w_0, b_0) = \arg \min_{(w,b)} ||w||^2
\]

subject to \(\forall i : y_i(\langle w, x_i \rangle + b) \geq 1\)
• output: \(\hat{w} = \frac{w_0}{||w_0||}, \hat{b} = \frac{b_0}{||w_0||}\)

How do we get a solution? Quadratic optimization problem:
objective is convex quadratic function, constraints are linear
inequalities \(\Rightarrow\) Quadratic Programming solvers!

Proposition

The output of algorithm above is a solution to the Equivalent
Formulation in the previous slide.

Proof: on the board and in the book (Lemma 15.2 [UML])
Equivalent Formulation and Support Vectors

Equivalent formulation (homogeneous halfspaces): assume first component of $x \in \mathcal{X}$ is 1, then

$$w_0 = \min_w ||w||^2 \text{ subject to } \forall i : y_i \langle w, x_i \rangle \geq 1$$

“Support Vectors” = vectors at minimum distance from $w_0$

The support vectors are the only ones that matter for defining $w_0$!

**Proposition**

Let $w_0$ be as above. Let $I = \{i : |\langle w_0, x_i \rangle| = 1\}$. Then there exist coefficients $\alpha_1, \ldots, \alpha_m$ such that

$$w_0 = \sum_{i \in I} \alpha_i x_i$$

“Support vectors” = $\{x_i : i \in I\}$

**Note**: Solving Hard-SVM is equivalent to find $\alpha_i$ for $i = 1, \ldots, m$, and $\alpha_i \neq 0$ only for support vectors
Hard-SVM works if data is linearly separable.

What if data is not linearly separable? ⇒ soft-SVM

**Idea:** modify constraints of Hard-SVM to allow for some violation, but take into account violations into objective function
Soft-SVM Constraints

Hard-SVM constraints:

\[ y_i(\langle w, x_i \rangle + b) \geq 1 \]

Soft-SVM constraints:

- slack variables: \( \xi_1, \ldots, \xi_m \geq 0 \Rightarrow \) vector \( \xi \)
- for each \( i = 1, \ldots, m: \) \( y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i \)
- \( \xi_i: \) how much constraint \( y_i(\langle w, x_i \rangle + b) \geq 1 \) is violated

Soft-SVM minimizes combinations of

- norm of \( w \)
- average of \( \xi_i \)

Tradeoff among two terms is controlled by a parameter \( \lambda \in \mathbb{R}, \lambda > 0 \)
Soft-SVM: Optimization Problem

- **input:** \((x_1, y_1), \ldots, (x_m, y_m)\), parameter \(\lambda > 0\)
- **solve:**
  \[
  \min_{w, b, \xi} \left( \lambda ||w||^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i \right)
  \]
  subject to \(\forall i : y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i\) and \(\xi_i \geq 0\)
- **output:** \(w, b\)

**Equivalent formulation:** consider the **hinge loss**

\[
\ell_{\text{hinge}}((w, b), (x, y)) = \max\{0, 1 - y(\langle w, x \rangle + b)\}
\]

Given \((w, b)\) and a training set \(S\), the empirical risk \(L_S^{\text{hinge}}((w, b))\) is

\[
L_S^{\text{hinge}}((w, b)) = \frac{1}{m} \sum_{i=1}^{m} \ell_{\text{hinge}}((w, b), (x_i, y_i))
\]
Soft-SVM as RLM

Soft-SVM: solve

\[
\min_{w, b, \xi} \left( \lambda \|w\|^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i \right)
\]
subject to \( \forall i : y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i \) and \( \xi_i \geq 0 \)

Equivalent formulation with hinge loss:

\[
\min_{w, b} \left( \lambda \|w\|^2 + L^\text{hinge}_S(w, b) \right)
\]

that is

\[
\min_{w, b} \left( \lambda \|w\|^2 + \frac{1}{m} \sum_{i=1}^{m} \ell^\text{hinge}((w, b), (x_i, y_i)) \right)
\]

Note:

- \( \lambda \|w\|^2 \): \( \ell_2 \) regularization
- \( L^\text{hinge}_S(w, b) \): empirical risk for hinge loss
Soft-SVM: Solution

We need to solve:

$$\min_{w, b} \left( \lambda \|w\|^2 + \frac{1}{m} \sum_{i=1}^{m} \ell_{\text{hinge}}((w, b), (x_i, y_i)) \right)$$

How?

- standard solvers for optimization problems
- **Stochastic Gradient Descent**
Exercise 1

Your friend has developed a new machine learning algorithm for binary classification (i.e., \( y \in \{-1, 1\} \)) with 0-1 loss and tells you that it achieves a generalization error of only 0.05. However, when you look at the learning problem he is working on, you find out that \( \Pr_D[y = 1] = 0.95 \).

- Assume that \( \Pr_D[y = \ell] = p_\ell \). Derive the generalization error of the (dumb) hypothesis/model that always predicts \( \ell \).
- Use the result above to decide if your friend’s algorithm has learned something or not.
Exercise 2

Consider a linear regression problem, where $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$, with mean squared loss. The hypothesis set is the set of constant functions, that is $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$, where $h_a(x) = a$. Let $S = ((x_1, y_1), \ldots, (x_m, y_m))$ denote the training set.

- Derive the hypothesis $h \in \mathcal{H}$ that minimizes the training error.
- Use the result above to explain why, for a given hypothesis $\hat{h}$ from the set of all linear models, the coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(\hat{h}(x_i) - y_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the average of the $y_i, i = 1, \ldots, m$ is a measure of how well $\hat{h}$ performs (on the training set).
Exercise 3

Consider the classification problem with \( \mathcal{X} = \mathbb{R}^2 \), \( \mathcal{Y} = \{0, 1\} \).
Consider the hypothesis class \( \mathcal{H} = \{h_{(c,a)}, c \in \mathbb{R}^2, a \in \mathbb{R}\} \) with

\[
h_{(c,a)}(x) = \begin{cases} 
1 & \text{if } ||x - c|| \leq a \\
0 & \text{otherwise}
\end{cases}
\]

Find the VC-dimension of \( \mathcal{H} \).
Exercise 4

Assuming we have the following dataset \((x_i \in \mathbb{R}^2)\) and by solving the SVM for classification we get the corresponding optimal dual variables:

<table>
<thead>
<tr>
<th>(i)</th>
<th>(x_i^T)</th>
<th>(y_i)</th>
<th>(\alpha_i^*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[0.2, -1.4]</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>[-2.1, 1.7]</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>[0.9, 1]</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>[-1, -3.1]</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>[-0.2, -1]</td>
<td>-1</td>
<td>0.25</td>
</tr>
<tr>
<td>6</td>
<td>[-0.2, 1.3]</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>[2.0, -1]</td>
<td>-1</td>
<td>0.25</td>
</tr>
<tr>
<td>8</td>
<td>[0.5, 2.1]</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Answer to the following:

(A) Which are the support vectors?

(B) Draw a schematic picture reporting the data points (approximately) and the optimal separating hyperplane, and mark the support vectors. Would it be possible, by moving only two data points, to obtain the SAME separating hyperplane with only 2 support vectors? If so, draw the modified configuration (approximately).
Exercise 5

Consider the ridge regression problem
\[ \arg \min_w \lambda ||w||^2 + \sum_{i=1}^{m} (\langle w, x_i \rangle - y_i)^2. \]
Let: \( h_S \) be the hypothesis obtained by ridge regression on with training set \( S \); \( h^* \) be the hypothesis of minimum generalization error among all linear models.

(A) Draw, in the plot below, a typical behaviour of (i) the training error and (ii) the test/generalization error of \( h_S \) as a function of \( \lambda \).

(B) Draw, in the plot below, a typical behaviour of (i) \( L_D(h_S) - L_D(h^*) \) and (ii) \( L_D(h_S) - L_S(h_S) \) as a function of \( \lambda \).