Introduction to Machine Learning
5. Support Vector Classification

Alex Smola
Carnegie Mellon University

http://alex.smola.org/teaching/cmu2013-10-701
10-701
Outline

• Support Vector Classification
  Large Margin Separation, optimization problem

• Properties
  Support Vectors, kernel expansion

• Soft margin classifier
  Dual problem, robustness
Support Vector Machines

Linear Separator

Ham

Spam
Linear Separator
Linear Separator

Ham

Spam
Linear Separator

Ham

Spam
Linear Separator

Ham

Spam

Carnegie Mellon University
Large Margin Classifier

\[
\langle w, x \rangle + b \leq -1
\]

\[
\langle w, x \rangle + b \geq 1
\]

linear function

\[
f(x) = \langle w, x \rangle + b
\]
Large Margin Classifier

\[ \langle w, x \rangle + b = -1 \]

\[ \langle w, x \rangle + b = 1 \]

\[
\frac{\langle x_+ - x_-, w \rangle}{2 \| w \|} = \frac{1}{2 \| w \|} \left[ [\langle x_+, w \rangle + b] - [\langle x_-, w \rangle + b] \right] = \frac{1}{\| w \|}
\]
Large Margin Classifier

\[ \langle w, x \rangle + b = -1 \]

\[ \langle w, x \rangle + b = 1 \]

optimization problem

\[
\max_{w, b} \frac{1}{\|w\|} \quad \text{subject to } y_i \left[ \langle x_i, w \rangle + b \right] \geq 1
\]
Large Margin Classifier

\[ \langle w, x \rangle + b = -1 \]

\[ \langle w, x \rangle + b = 1 \]

Optimization problem

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|w\|^2 \\
\text{subject to} & \quad y_i [\langle x_i, w \rangle + b] \geq 1
\end{align*}
\]
Dual Problem

• Primal optimization problem

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|w\|^2 \\
\text{subject to} & \quad y_i \left[ \langle x_i, w \rangle + b \right] \geq 1
\end{align*}
\]

• Lagrange function

\[
L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i \left[ y_i \left[ \langle x_i, w \rangle + b \right] - 1 \right]
\]

Optimality in \( w, b \) is at saddle point with \( \alpha \)

• Derivatives in \( w, b \) need to vanish
Dual Problem

- **Lagrange function**
  \[
  L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i [y_i (\langle x_i, w \rangle + b) - 1]
  \]

- **Derivatives in** \( w, b \) **need to vanish**
  \[
  \frac{\partial}{\partial w} L(w, b, a) = w - \sum_i \alpha_i y_i x_i = 0
  \]
  \[
  \frac{\partial}{\partial b} L(w, b, a) = \sum_i \alpha_i y_i = 0
  \]

- **Plugging terms back into** \( L \) **yields**
  \[
  \text{maximize } -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i
  \]
  subject to \( \sum_i \alpha_i y_i = 0 \) and \( \alpha_i \geq 0 \)

Carnegie Mellon University
Support Vector Machines

\[ \min_{w,b} \frac{1}{2} \|w\|^2 \text{ subject to } y_i [\langle x_i, w \rangle + b] \geq 1 \]

\[ w = \sum_i y_i \alpha_i x_i \]

\[ \max_{\alpha} -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i \]

\[ \text{subject to } \sum \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0 \]
Support Vectors

$$\minimize_{w,b} \frac{1}{2} \|w\|^2 \text{ subject to } y_i [\langle x_i, w \rangle + b] \geq 1$$

$$w = \sum_i y_i \alpha_i x_i$$

Karush Kuhn Tucker
Optimality condition

$$\alpha_i [y_i [\langle w, x_i \rangle + b] - 1] = 0$$

$$\alpha_i = 0 \quad \alpha_i > 0 \implies y_i [\langle w, x_i \rangle + b] = 1$$
Properties

\[ w = \sum_{i} y_i \alpha_i x_i \]

- Weight vector \( w \) as weighted linear combination of instances
- Only points on margin matter (ignore the rest and get same solution)
- Only inner products matter
  - Quadratic program
  - We can replace the inner product by a kernel
- Keeps instances away from the margin
Example
Example

Number of Support Vectors: 3  (-ve: 2, +ve: 1)  Total number of points: 15
Why large margins?

- Maximum robustness relative to uncertainty
- Symmetry breaking
- Independent of correctly classified instances
- Easy to find for easy problems

Figure 7.3

Two-dimensional toy example of a classification problem: Separate 'o' from '+' using a hyperplane. Suppose that we add bounded noise to each pattern. If the optimal margin hyperplane has margin $\rho$, then it is bounded by $r < \rho$, and the hyperplane will correctly separate even the noisy patterns. Conversely, if we run the perceptron algorithm (which finds some separating hyperplane, but not necessarily the optimal one) on the noisy data, then we would recover the optimal hyperplane in the limit $r \rightarrow \rho$.

If we knew $\rho$ beforehand, then this could actually be turned into an optimal margin classifier training algorithm, as follows. If we use a $r$ which is slightly smaller than $\rho$, then the patterns will be separable with a nonzero margin. In this case, the standard perceptron algorithm can be shown to converge.

1. Rosenblatt's perceptron algorithm [423] is one of the simplest conceivable iterative procedures for computing a separating hyperplane. In its simplest form, it proceeds as follows. We start with an arbitrary weight vector $w_0$. At step $n \in \mathbb{N}$, we consider the training example $(x_n, y_n)$. If it is classified correctly using the current weight vector (i.e., if $\text{sgn} \langle x_n, w_n \rangle = y_n$), we set $w_n := w_n - x_n$; otherwise, we set $w_n := w_n + \eta y_n x_i (here, \eta > 0 is a learning rate). We look over all patterns repeatedly, until we can complete one full pass through the training set without a single error. The result in weight vector will thus classify all points correctly. Novikoff [369] proved that this procedure terminates, provided that the training set is separable with a nonzero margin.
Support Vector Machines

CLASSIFIERS
Large Margin Classifier

\[ \langle w, x \rangle + b \leq -1 \]
\[ \langle w, x \rangle + b \geq 1 \]

Linear function

\[ f(x) = \langle w, x \rangle + b \]
Large Margin Classifier

$\langle w, x \rangle + b \leq -1$

$\langle w, x \rangle + b \geq 1$

linear function

$f(x) = \langle w, x \rangle + b$
Large Margin Classifier

\[ \langle w, x \rangle + b \leq -1 \]

\[ \langle w, x \rangle + b \geq 1 \]

linear function

\[ f(x) = \langle w, x \rangle + b \]

linear separator is impossible
Theorem (Minsky & Papert)
Finding the minimum error separating hyperplane is NP hard
Large Margin Classifier

\[ \langle w, x \rangle + b \leq -1 \]

\[ \langle w, x \rangle + b \geq 1 \]

Theorem (Minsky & Papert)
Finding the minimum error separating hyperplane is NP hard
Large Margin Classifier

\[ \langle w, x \rangle + b \leq -1 \]

\[ \langle w, x \rangle + b \geq 1 \]

Theorem (Minsky & Papert)
Finding the minimum error separating hyperplane is NP hard

minimum error separator is impossible
Adding slack variables

\[ \langle w, x \rangle + b \leq -1 + \xi \]

\[ \langle w, x \rangle + b \geq 1 - \xi \]

Convex optimization problem
Adding slack variables

\[ \langle w, x \rangle + b \leq -1 + \xi \]

\[ \langle w, x \rangle + b \geq 1 - \xi \]

Convex optimization problem
Adding slack variables

\[ \langle w, x \rangle + b \leq -1 + \xi \]

\[ \langle w, x \rangle + b \geq 1 - \xi \]

Convex optimization problem

minimize amount of slack

Carnegie Mellon University
Intermezzo

Convex Programs for Dummies

• Primal optimization problem

\[
\minimize_x f(x) \text{ subject to } c_i(x) \leq 0
\]

• Lagrange function

\[
L(x, \alpha) = f(x) + \sum_i \alpha_i c_i(x)
\]

• First order optimality conditions in \( x \)

\[
\partial_x L(x, \alpha) = \partial_x f(x) + \sum_i \alpha_i \partial_x c_i(x) = 0
\]

• Solve for \( x \) and plug it back into \( L \)

\[
\max_{\alpha} L(x(\alpha), \alpha)
\]

(keep explicit constraints)
Adding slack variables

\[ \langle w, x \rangle + b \leq -1 + \xi \]

\[ \langle w, x \rangle + b \geq 1 - \xi \]

Convex optimization problem
Adding slack variables

Convex optimization problem

\[ \langle w, x \rangle + b \leq -1 + \xi \]

\[ \langle w, x \rangle + b \geq 1 - \xi \]
Adding slack variables

\[ \langle w, x \rangle + b \leq -1 + \xi \]

\[ \langle w, x \rangle + b \geq 1 - \xi \]

Convex optimization problem

minimize amount of slack
Adding slack variables

• Hard margin problem

 minimizew, b12∥w∥2 subject toyi [⟨w, xi⟩ + b] ≥ 1

• With slack variables

 minimizew, b12∥w∥2 + C ∑iξi subject toyi [⟨w, xi⟩ + b] ≥ 1 − ξi and ξi ≥ 0

Problem is always feasible. Proof:
w = 0 and b = 0 and ξi = 1 (also yields upper bound)
Dual Problem

• Primal optimization problem

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \\
\text{subject to} & \quad y_i [\langle w, x_i \rangle + b] \geq 1 - \xi_i \text{ and } \xi_i \geq 0
\end{align*}
\]

• Lagrange function

\[
L(w, b, \alpha) = \frac{1}{2} \|w\|^2 + C \sum_i \xi_i - \sum_i \alpha_i \left[ y_i [\langle x_i, w \rangle + b] + \xi_i - 1 \right] - \sum_i \eta_i \xi_i
\]

Optimality in \(w, b, \xi\) is at saddle point with \(\alpha, \eta\)

• Derivatives in \(w, b, \xi\) need to vanish
Dual Problem

• Lagrange function

\[ L(w, b, \alpha) = \frac{1}{2} \|w\|^2 + C \sum_i \xi_i - \sum_i \alpha_i [y_i \langle x_i, w \rangle + b] + \xi_i - 1 - \sum_i \eta_i \xi_i \]

• Derivatives in w, b need to vanish

\[ \partial_w L(w, b, \xi, \alpha, \eta) = w - \sum_i \alpha_i y_i x_i = 0 \]

\[ \partial_b L(w, b, \xi, \alpha, \eta) = \sum_i \alpha_i y_i = 0 \]

\[ \partial_{\xi_i} L(w, b, \xi, \alpha, \eta) = C - \alpha_i - \eta_i = 0 \]

• Plugging terms back into L yields

\[
\maximize_{\alpha} - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i \\
\text{subject to } \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \in [0, C] 
\]

bound influence
Karush Kuhn Tucker Conditions

maximize \( \alpha \left( -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i \right) \)

subject to \( \sum_i \alpha_i y_i = 0 \) and \( \alpha_i \in [0, C] \)

\( w = \sum_i y_i \alpha_i x_i \)

\( \alpha_i \left( [y_i \langle w, x_i \rangle + b] + \xi_i - 1 \right) = 0 \)

\( \eta_i \xi_i = 0 \)

\( \alpha_i = 0 \implies y_i \left[ \langle w, x_i \rangle + b \right] \geq 1 \)

\( 0 < \alpha_i < C \implies y_i \left[ \langle w, x_i \rangle + b \right] = 1 \)

\( \alpha_i = C \implies y_i \left[ \langle w, x_i \rangle + b \right] \leq 1 \)
C=2
$C=2$
$C=5$
Solving the optimization problem

• Dual problem

\[
\begin{align*}
\text{maximize} & \quad \alpha \left( -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i \right) \\
\text{subject to} & \quad \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \in [0, C]
\end{align*}
\]

• If problem is small enough (1000s of variables) we can use off-the-shelf solver (CVXOPT, CPLEX, OOQP, LOQO)

• For larger problem use fact that only SVs matter and solve in blocks (active set method).
Nonlinear Separation
The Kernel Trick

- **Linear soft margin problem**
  \[
  \min_{w, b} \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \\
  \text{subject to } y_i [\langle w, x_i \rangle + b] \geq 1 - \xi_i \text{ and } \xi_i \geq 0
  \]

- **Dual problem**
  \[
  \max_{\alpha} -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i \\
  \text{subject to } \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \in [0, C]
  \]

- **Support vector expansion**
  \[
  f(x) = \sum_i \alpha_i y_i \langle x_i, x \rangle + b
  \]
The Kernel Trick

• **Linear soft margin problem**

\[
\begin{align*}
\text{minimize} \quad & \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \\
\text{subject to} \quad & y_i [\langle w, \phi(x_i) \rangle + b] \geq 1 - \xi_i \text{ and } \xi_i \geq 0
\end{align*}
\]

• **Dual problem**

\[
\begin{align*}
\text{maximize} \quad & -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j) + \sum_i \alpha_i \\
\text{subject to} \quad & \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \in [0, C]
\end{align*}
\]

• **Support vector expansion**

\[
f(x) = \sum_i \alpha_i y_i k(x_i, x) + b
\]
And now with a narrower kernel
And now with a very wide kernel
Nonlinear separation

- Increasing C allows for more nonlinearities
- Decreases number of errors
- SV boundary need not be contiguous
- Kernel width adjusts function class
Risk and Loss
Loss function point of view

- Constrained quadratic program

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \\
\text{subject to} & \quad y_i [\langle w, x_i \rangle + b] \geq 1 - \xi_i \text{ and } \xi_i \geq 0
\end{align*}
\]

- Risk minimization setting

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|w\|^2 + C \sum_i \max [0, 1 - y_i [\langle w, x_i \rangle + b]]
\end{align*}
\]

Follows from finding minimal slack variable for given \((w,b)\) pair.
Soft margin as proxy for binary

- Soft margin loss: \( \max(0, 1 - yf(x)) \)
- Binary loss: \( \{yf(x) < 0\} \)
More loss functions

- **Logistic** \( \log \left[ 1 + e^{-f(x)} \right] \)

- **Huberized loss**
  \[
  \begin{cases}
  0 & \text{if } f(x) > 1 \\
  \frac{1}{2} (1 - f(x))^2 & \text{if } f(x) \in [0, 1] \\
  \frac{1}{2} - f(x) & \text{if } f(x) < 0
  \end{cases}
  \]

- **Soft margin**
  \[ \max(0, 1 - f(x)) \]
Risk minimization view

- Find function \( f \) minimizing classification error
  \[
  R[f] := \mathbb{E}_{x,y \sim p(x,y)} \{ yf(x) > 0 \}
  \]
- Compute empirical average
  \[
  R_{\text{emp}}[f] := \frac{1}{m} \sum_{i=1}^{m} \{ y_i f(x_i) > 0 \}
  \]
- Minimization is nonconvex
- Overfitting as we minimize empirical error
- Compute convex upper bound on the loss
- Add regularization for capacity control
  \[
  R_{\text{reg}}[f] := \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y_i f(x_i)) + \lambda \Omega[f]
  \]
Summary

• Support Vector Classification
  Large Margin Separation, optimization problem

• Properties
  Support Vectors, kernel expansion

• Soft margin classifier
  Dual problem, robustness