The soft-margin support vector machine

Nuno Vasconcelos

ECE Department, UCSD
Classification

- A classification problem has two types of variables:
  - $X$ - vector of observations (features) in the world
  - $Y$ - state (class) of the world

- E.g.
  - $x \in X \subset \mathbb{R}^2 = \text{(fever, blood pressure)}$
  - $y \in Y = \{\text{disease, no disease}\}$

- $X$, $Y$ related by an (unknown) function $f$:

$$X \xrightarrow{f(.)} Y = f(X)$$

- Goal: design a classifier $h: X \rightarrow Y$ such that $h(x) = f(x)$ $\forall x$
Linear classifier

- implements the decision rule
  \[ h^*(x) = \begin{cases} 
  1 & \text{if } g(x) > 0 \\
  -1 & \text{if } g(x) < 0 
\end{cases} = \text{sgn}[g(x)] \]
  with \[ g(x) = w^T x + b \]

- has the properties
  - it divides \( \mathcal{X} \) into two “half-spaces”
  - boundary is the plane with:
    - normal \( w \)
    - distance to the origin \( b/\|w\| \)
  - \( g(x)/\|w\| \) distance from \( x \) to boundary
  - for a linearly separable training set \( D = \{(x_1,y_1), \ldots, (x_n,y_n)\} \) we have zero empirical risk when
    \[ y_i(w^T x_i + b) > 0, \forall i \]
The margin

- is the distance from the boundary to the closest point

\[ \gamma = \min_i \frac{|w^T x_i + b|}{||w||} \]

- there will be no error if it is strictly greater than zero

\[ y_i(w^T x_i + b) > 0, \ \forall i \iff \gamma > 0 \]

- note that this is ill-defined in the sense that \( \gamma \) does not change if both \( w \) and \( b \) are scaled by \( \lambda \)

- we need a normalization
Normalization

- a convenient normalization is to make $|g(x)| = 1$ for the closest point, i.e.

$$\min_i |w^T x_i + b| \equiv 1$$

under which

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

- the SVM is the classifier that maximizes the margin under these constraints

$$\min_{w,b} ||w||^2 \text{ subject to } y_i(w^T x_i + b) \geq 1 \ \forall i$$
The dual problem

► no duality gap, the dual problem is

\[
\max_{\alpha \geq 0} \left\{ -\frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_i \alpha_i \right\}
\]

subject to \( \sum_i y_i \alpha_i = 0 \)

► once this is solved, the vector

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

is the normal to the maximum margin plane

► \( b^* \) can be left a free parameter (false pos. vs misses) or set to

\[ b^* = -\frac{w^T (x^+ + x^-)}{2} \]
Support vectors

from the KKT conditions, a
innactive constraint has zero
Lagrange multiplier $\alpha_i$. That is,

- $\alpha_i > 0$  \text{iif}  $y_i(w^*T x_i + b^*) = 1$
- i.e. only for points
  \[ |w^*T x_i + b^*| = 1 \]
  which lie at a distance equal
to the margin
- these support the plane and are
called support vectors
- the decision rule is
  \[
  f(x) = \text{sgn}\left[ \sum_{i \in SV} y_i \alpha_i^* x_i^T x + b^* \right]
  \]
  - the remaining points are irrelevant!
Kernelization

- note that all equations depend only on $x_i^T x_j$
- the kernel trick is trivial: replace by $K(x_i, x_j)$

1) training

$$\max_{\alpha \geq 0} \left\{ -\frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j k(x_i, x_j) + \sum_i \alpha_i \right\}$$
subject to $\sum_i y_\alpha = 0$

$$b^* = -\frac{1}{2} \sum_{i \in SV} y_\alpha^* (K(x_i, x^+) + K(x_i, x^-))$$

2) decision function:

$$f(x) = \text{sgn} \left( \sum_{i \in SV} y_\alpha^* K(x_i, x) + b^* \right)$$
Input space interpretation

- decision function identical to Bayesian decision rule for
  1) class 1 with likelihood
  \[ \sum_{i | i \in SV, y_i \geq 0} \pi_i^* K(x_i, x) \]
  and prior
  \[ \sum_{i | i \in SV, y_i < 0} \alpha_i^* / \sum_{i} \alpha_i \]

  2) class 2 with likelihood
  \[ \sum_{i | i \in SV, y_i < 0} \beta_i^* K(x_i, x) \]
  and prior
  \[ \sum_{i | i \in SV, y_i \geq 0} \alpha_i^* / \sum_{i} \alpha_i \]

- i.e.
  \[ f(x) = \begin{cases} 
  1, & \text{if } \frac{\sum_{i | i \in SV, y_i \geq 0} \pi_i^* K(x_i, x)}{\sum_{i | i \in SV, y_i < 0} \beta_i^* K(x_i, x)} \geq T \\ 
  -1, & \text{otherwise} 
\end{cases} \]
Input space interpretation

- **peculiar kernel estimates**
  - only place kernels on the support vectors, all other points ignored

- **discriminant density estimation**
  - concentrate modeling power where it matters the most, i.e. near classification boundary
  - smart, since points away from the boundary are always well classified, even if density estimates in their region are poor
  - the SVM is a highly efficient combination of the BDR with kernel estimates, complexity $O(|SV|)$ instead of $O(n)$
Limitations of the SVM

appealing, but also points out the limitations of the SVM:

• major problem of kernel density estimation is the choice of bandwidth

• if too small estimates have too much variance, if too large the estimates have too much bias

• this problem appears again for the SVM

• no generic “optimal” procedure to find the kernel or its parameters

• requires trial and error

• note, however, that this is less of a headache since only a few kernels have to be evaluated
Non-separable problems

- so far we have assumed linearly separable classes
- this is rarely the case in practice
- a separable problem is “easy” most classifiers will do well
- we need to be able to extend the SVM to the non-separable case

- basic idea:
  - with class overlap we cannot enforce a margin
  - but we can enforce a soft margin
  - for most points there is a margin, but then there are a few outliers that cross-over, or are closer to the boundary than the margin
Soft margin optimization

- mathematically this can be done by introducing slack variables

- instead of solving

\[
\min_{w,b} \|w\|^2 \quad \text{subject to } y_i(w^T x_i + b) \geq 1 \quad \forall i
\]

- we solve the problem

\[
\min_{w,\xi,b} \|w\|^2 \quad \text{subject to } y_i(w^T x_i + b) \geq 1 - \xi_i \quad \forall i
\]

\[\xi_i \geq 0, \forall i\]

- the \(\xi_i\) are called slacks

- basically, the same as before but points with \(\xi_i > 0\) are allowed to violate the margin
Soft margin optimization

- note that the problem is not really well defined
- by making $\xi_i$ arbitrarily large, any $w$ will do
- we need to penalize large $\xi_i$
- this is done by solving instead

$$\min_{w,\xi,b} \||w\||^2 + Cf(\xi)$$

subject to $y_i(w^T x_i + b) \geq 1 - \xi_i \ \forall i$

$\xi_i \geq 0, \forall i$

- $f(\xi)$ is usually a norm. We consider
  - the 1-norm: $f(\xi) = \sum_i \xi_i$
  - the 2-norm: $f(\xi) = \sum_i \xi_i^2$
2-norm SVM

\[
\min_{w, \xi, b} \|w\|^2 + C \sum_i \xi_i^2
\]

subject to \(y_i (w^T x_i + b) \geq 1 - \xi_i \quad \forall i\), \(\xi_i \geq 0, \forall i\)

• note that
  
  • if \(\xi_i < 0\), and the constraint (**\) is satisfied then
  • (**\) is satisfied by \(\xi_i = 0\) and the cost will be smaller
  • hence \(\xi_i < 0\) is never a solution and the positivity constraints on the \(\xi_i\) are redundant
  • they can therefore be dropped
2-norm SVM

this leads to

\[
\min_{w, \xi, b} \frac{1}{2}\|w\|^2 + \frac{C}{2} \sum_i \xi_i^2 \\
\text{subject to } y_i (w^T x_i + b) \geq 1 - \xi_i \quad \forall i
\]

and

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

from which

\[
\nabla_w L = 0 \iff w - \sum_i \alpha_i y_i x_i = 0 \iff w^* = \sum_i \alpha_i y_i x_i
\]

\[
\nabla_b L = 0 \iff \sum_i y_i \alpha_i = 0
\]

\[
\nabla_{\xi_i} L = 0 \iff C \xi_i - \alpha_i = 0
\]
2-norm dual problem

plugging back \( w^* = \sum_i \alpha_i y_i x_i \), \( \sum_i y_i \alpha_i = 0 \), \( \xi_i = \frac{\alpha_i}{C} \) we get the Lagrangian

\[
L(w^*, b, \xi, \alpha, r) = \frac{1}{2} \|w\|^2 + \frac{C}{2} \sum_i \left( \frac{\alpha_i}{C} \right)^2 + \sum_i \alpha_i \left[ 1 - \frac{\alpha_i}{C} - y_i (w^T x_i + b) \right]
\]

\[
= \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j x^T_i x_j - \frac{1}{2} \sum_i \alpha_i^2 C + \sum_i \alpha_i - \sum_{ij} \alpha_i \alpha_j y_i y_j x^T_i x_j
\]

\[
- \sum_i \alpha_i y_i b = 0
\]

\[
= -\frac{1}{2} \left( \sum_{ij} \alpha_i \alpha_j y_i y_j x^T_i x_j + \frac{1}{C} \sum_i \alpha_i^2 \right) + \sum_i \alpha_i
\]

\[
= -\frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \left( x^T_i x_j + \delta_{ij} \frac{\delta_{ij}}{C} \right) + \sum_i \alpha_i \quad \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}
\]
2-norm dual problem

- the dual problem is

\[
\max_{\alpha \geq 0} \left\{ -\frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \left( x_i^T x_j + \frac{\delta_{ij}}{C} \right) + \sum_i \alpha_i \right\}
\]

subject to \( \sum_i y_i \alpha_i = 0, \quad \alpha_i \geq 0 \)

- same as hard margin, with \( 1/C \times I \) added to kernel matrix

- this:
  - increments the eigenvalues by \( 1/C \) making the problem better conditioned
  - for larger \( C \), the extra term is smaller, and outliers have a larger influence
  - (less penalty on them, more reliance on data term)
Soft dual for 1-norm

\[
\min_{w,\xi,b} \|w\|^2 + C \sum \xi_i \quad \text{subject to } y_i (w^T x_i + b) \geq 1 - \xi_i \quad \forall i \\
\xi_i \geq 0, \forall i
\]

the Lagrangian is

\[
L(w, b, \xi, \alpha, r) = \frac{1}{2} \|w\|^2 + C \sum \xi_i + \sum \alpha_i [1 - \xi_i - y_i (w^T x_i + b)] - \sum r_i \xi_i
\]

and setting derivatives to zero

\[
\nabla_w L = 0 \iff w - \sum \alpha_i y_i x_i = 0 \iff w^* = \sum \alpha_i y_i x_i
\]

\[
\nabla_b L = 0 \iff \sum y_i \alpha_i = 0
\]

\[
\nabla_{\xi_i} L = 0 \iff C - \alpha_i - r_i = 0
\]
The dual problem

- plugging back \( w^* = \sum_i \alpha_i y_i x_i, \sum_i y_i \alpha_i = 0, \quad r_i = C - \alpha_i \) we get the Lagrangian

\[
L(w^*, b, \xi, \alpha, r) = \frac{1}{2} \|w\|^2 + C \sum_i \xi_i + \sum_i \alpha_i \left[1 - \xi_i - y_i \left(w^T x_i + b\right)\right] - \sum_i r_i \xi_i
\]

\[
= \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j x_i^T x_j + C \sum_i \xi_i + \sum_i \alpha_i \left(1 - \xi_i\right) - \sum_{ij} \alpha_i \alpha_j y_i y_j x_i^T x_j \\
- \sum_i \alpha_i y_i b - \sum_i (C - \alpha_i) \xi_i = 0
\]

\[
= -\frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_i \alpha_i
\]

- this is exactly like the hard-margin case with the extra constraint \( \alpha_i = C - r_i \) for all \( i \)
The dual problem

**in summary:**

- \( w^* = \sum \alpha_i y_i x_i \), \( \sum_i y_i \alpha_i = 0 \), \( r_i = C - \alpha_i \) (*)

- \( L(w^*, b, \xi, \alpha, r) = -\frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_i \alpha_i \)

- since \( r_i \) are Lagrange multipliers, \( r_i \geq 0 \), (*) means that \( \alpha_i \leq C \)

- also, from the KKT conditions
  - i) \( r_i > 0 \iff \xi_i = 0 \), \( r_i = 0 \iff \xi_i > 0 \) (\( r_i \xi_i = 0 \))
  - ii) \( \alpha_i \left[ 1 - \xi_i - y_i (w^T x_i + b) \right] = 0 \)

- since \( \alpha_i = C - r_i \), i) implies that
  - a) \( \xi_i = 0 \) and \( \alpha_i < C \) or b) \( \xi_i > 0 \) and \( \alpha_i = C \)

- from ii), in case a), we have \( y_i (w^T x_i + b) = 1 \), i.e. \( x_i \) is on the margin

- in case b), \( y_i (w^T x_i + b) = 1 - \xi_i \), i.e. \( x_i \) is an outlier

- finally, as before, correctly classified points when \( \alpha_i = 0 \) (\( r_i = C \iff \xi_i = 0 \))
The dual problem

- overall, dual problem is

\[
\max_{\alpha \geq 0} \left\{ -\frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_i \alpha_i \right\}
\]

subject to

\[
\sum_i y_i \alpha_i = 0,
\]

\[
0 \leq \alpha_i \leq C
\]

- the only difference with respect to the hard margin case is the box constraint on the \( \alpha_i \).

- geometrically we have this
Support vectors

- are the points with $\alpha_i > 0$
- as before, the decision rule is

$$f(x) = \text{sgn} \left[ \sum_{i \in SV} y_i \alpha_i^* x_i^T x + b^* \right]$$

where $SV = \{i \mid \alpha_i^* > 0\}$

- and $b^*$ chosen s.t.
  - $y_i g(x_i) = 1$, for all $x_i$ s.t. $0 < \alpha_i < C$
- the box constraint on Lagrange multipliers
  - makes intuitive sense
  - prevents a single outlier from having large impact
Soft margin SVM

- note that $C$ controls the importance of outliers
  - larger $C$ implies that more emphasis is given to minimizing the number of outliers
  - more of them will be within the margin

- 1-norm vs 2-norm
  - as usual, the 1-norm tends to limit more drastically the outlier contributions
  - this makes it a bit more robust, and it tends to be used more frequently in practice

- common problem:
  - not really intuitive how to set up $C$
  - usually cross-validation, there is a need to cross-validate with respect to both $C$ and kernel parameters
\( \nu \)-SVM

▶ a more recent formulation has been introduced to try to overcome this

\[
\min_{w,\xi,\rho,b} \|w\|^2 - \nu \rho + \frac{1}{n} \sum_i \xi_i
\]

subject to \( y_i (w^T x_i + b) \geq \rho - \xi_i \), \( \forall i \),

\( \xi_i \geq 0, \forall i \), \( \rho \geq 0 \)

▶ advantages:

- \( \nu \) has intuitive interpretation:
  - 1) \( \nu \) is an upper bound on the proportion of training vectors that are margin errors, i.e. for which
    \[
    y_i g(x_i) \leq \rho
    \]
  - 2) \( \nu \) is a lower bound on total number of support vectors

- more discussion on the homework
Connections to regularization

we talked about penalizing functions that are too complicated, to improve generalization

instead of the empirical risk, we should minimize the regularized risk

\[ R_{reg}[f] = R_{emp}[f] + \lambda \Omega[f] \]

the SVM seems to be doing this in some sense:

- it is designed to have as few errors as possible on training set (this is controlled by the soft margin weight C)
- we maximize the margin, by minimizing \( ||w||^2 \) (which is a form of complexity penalty)
- hence, minimizing the margin must be connected to enforcing some form of regularizer
Connections to regularization

- the connection can be made explicit
- consider the 1-norm SVM

\[
\begin{align*}
\min_{w, \xi, b} & \quad \|w\|^2 + C \sum_i \xi_i, \\
\text{subject to} & \quad y_i g(x_i) \geq 1 - \xi_i, \quad \forall i \\
\text{and} & \quad \xi_i \geq 0, \quad \forall i
\end{align*}
\]

- the constraints can be rewritten as
  - i) \( \xi_i \geq 0 \) and ii) \( \xi_i \geq 1 - y_i g(x_i) \)
- which is equivalent to

\[
\xi_i \geq \max\left[0, \ 1 - y_i g(x_i)\right] = \left[1 - y_i g(x_i)\right]_+
\]
Connections to regularization

- note that the cost $\|\mathbf{w}\|^2 + C \sum_i \xi_i$ can only increase with larger $\xi_i$

- hence, at the optimal solution

$$\xi^*_i = [1 - y_i g(x_i)]_+$$

- and the problem is

$$\min_{w, b} \|\mathbf{w}\|^2 + C \sum_i [1 - y_i g(x_i)]_+$$

- which is equivalent to

$$\min_{w, b} \sum_i [1 - y_i g(x_i)]_+ + \lambda \|\mathbf{w}\|^2$$

(by making $\lambda = 1/C$)
Connections to regularization

- this

$$\min_{w,b} \sum_i \left[ 1 - y_i g(x_i) \right]_+ + \lambda \| w \|^2$$

- can be seen as a regularized risk

$$R_{reg}[f] = \sum_i L[x_i, y_i, f] + \lambda \Omega[f]$$

with

- 1) loss function

$$L[x, y, g] = \left[ 1 - yg(x) \right]_+$$

- 2) standard regularizer

$$\Omega[w] = \| w \|^2$$
The SVM loss

(it is interesting to compare the SVM loss)

\[ L[x, y, g] = [1 - yg(x)]_+ \]

with the “0-1” loss:

- the SVM loss penalizes large negative margins
- assigns some penalty to anything with margin less than 1
- for the “0-1” loss the errors are all the same

the regularizer

- penalizes planes of large w.
- standard measure of complexity in regularization theory
Regularization

 mê the regularization connection could be used to derive the SVM from classical results, e.g. the representer theorem

Theorem: Let

- \( \Omega: [0, \infty) \rightarrow \mathcal{H} \) be a strictly monotonically increasing function,
- \( \mathcal{H} \) the RKHS associated with a kernel \( k(x,y) \)
- \( L[y, f(x)] \) a loss function

then, if

\[
    f^* = \arg\min_f \left[ \sum_{i=1}^{n} L[y_i, f(x_i)] + \lambda \Omega \left( \|f\|^2 \right) \right]
\]

\( f^* \) admits a representation of the form

\[
    f^* = \sum_{i=1}^{n} \alpha_i k(., x_i)
\]
Regularization

- in the SVM case, this immediately means that
  \[ w^* = \sum_i \alpha_i y_i k(., x_i) \]
  and we could have derived the SVM from it.

- note that we have seen that it is this theorem which makes
  the problem one of optimization on a finite dim. space
  \[ \alpha^* = \arg \min_{\alpha} \left[ \sum_{i=1}^n L[Y, K\alpha] + \lambda \Omega(\alpha^T K\alpha) \right] \]
  even though \( \mathcal{H} \) is infinite dimensional

- in this sense the SVM is really nothing new
  - regularization has been used since the beginning of the century
  - it has just shown that, under appropriate loss, it provides explicit guarantees on generalization error (more on future classes)
Any Questions?