An iterative learning algorithm that can find linear threshold function to partition linearly separable set of points. Assume zero threshold value.

1) $w(0) = \text{arbitrary, } j=1, k=0$
2) Pick point $(x(j), d(j))$.
3) If $w(k)^\top x(k)d(k) > 0$ go to 5)
4) $w(k+1) = w(k) + \mu x(k)d(k)$, $k=k+1$
5) Increment $j$, if at end of data, set $j=1$, check if cycled through data without changing $w$, if not go to 2
6) Otherwise stop.
Weight characteristics

- If \( w(0)=0 \), then \( w^*= \Sigma \alpha(k)x(k) \). Weight vector is a weighted sum of inputs. The largest magnitude \( \alpha(k) \) is associated with the input that is updated the most and has the smallest margin.
- Nonzero threshold handled with input set to 1. Transformation from input space to feature space.
- Many possible solutions for weight vector as observed from version space.
- PLA is an ill posed problem.
Optimum Margin Classifiers

Consider methods based on optimum margin classifiers or Support Vector Machines (SVM)

Here we consider a different algorithm that is well posed with a unique optimal solution. The solution is based on finding the largest minimal margin of all training points.
Optimal Marginal Classifiers

Given a set of points that are linearly separable:

which hyperplane should you choose to separate points?

Choose hyperplane that maximizes distance between two sets of points.
Finding Optimal Hyperplane

- Draw convex hull around each set of points.
- Find shortest line segment connecting two convex hulls.
- Find midpoint of line segment.
- Optimal hyperplane is perpendicular to segment at midpoint of line segment.
Alternative Characterization of Optimal Margin Classifiers

Maximizing margins equivalent to minimizing magnitude of weight vector.

\[
\mathbf{W}^T (\mathbf{u} - \mathbf{v}) = 2
\]

\[
\mathbf{W}^T (\mathbf{u} - \mathbf{v}) / \|\mathbf{W}\| = 2 / \|\mathbf{W}\| = 2m
\]

\[
\mathbf{W}^T \mathbf{u} + b = 1
\]

\[
\mathbf{W}^T \mathbf{v} + b = -1
\]
Quadratic Programming Problem

Problem statement:
$$\max_{(w, b)} \min (||x - x(i)||) : x \in \mathbb{R}^n, \ w^T x + b = 0, \ 1 \leq i \leq l$$

which is equivalent to solving (QP) problem:
$$\begin{align*}
\text{Min} & \quad \frac{1}{2} ||w||^2 \\
\text{subject to} & \quad d_i (w^T x(i) + b) \geq 1, \ \forall i
\end{align*}$$
Lagrange Multipliers

Constrained optimization can be dealt with by introducing Lagrange multipliers $\alpha(i) \geq 0$ and augmenting objective function

$$L(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum \alpha(i) (d(i) (w^T x(i) + b) - 1)$$
Support Vectors

QP program is convex and note that at solution
\[ \frac{\partial L(w,b,\alpha)}{\partial w} = 0 \] and \[ \frac{\partial L(w,b,\alpha)}{\partial b} = 0 \]
leading to
\[ \sum \alpha(i) d(i) = 0 \] and \[ w = \sum \alpha(i) d(i) x(i) \]
Weight vector expressed in terms of subset of training vectors \( x_i \) and \( \alpha_i \) that are nonzero. These are support vectors. By KKT conditions we have that
\[ \alpha(i) (d(i) (w^T x(i) + b) - 1) = 0 \]
indicating that support vectors lie on margin.
Support Vectors

Points in red are support vectors.
Dual optimization representation

Solving the primal QP problem is equivalent to solving the dual QP problem. Primal variables $w, b$ are eliminated.

$$\max W(\alpha) = \sum \alpha(i) - \frac{1}{2} \sum \alpha(i) \alpha(j) d(i) d(j)(x(i)^T x(j))$$
subject to $\alpha(i) \geq 0$, and $\sum \alpha(i) d(i) = 0$.

The hyperplane decision function can be written as
$$f(x) = \text{sgn} \left( \sum \alpha(i) d(i) x^T x(i) + b \right)$$
Comments about SVM solution

- Solution can be found in primal or dual spaces. When we discuss kernels later there are advantages of finding solution in dual space.
- Threshold, $b$ is determined after weight $w$ is found.
- In version space, SVM solution is center of largest hypersphere that fits in version space.
- Solution can be modified to
  - Handle training data that is not linearly separable via slack variables
  - Implement nonlinear discriminant function via kernels
- Solving quadratic programming problems
Handling data that not linearly separable

- Data is often noisy or inconsistent resulting in training data being not linearly separable.
- PLA algorithm can be modified (pocket algorithm), however there are problems with algorithm termination.
- SVM can easily handle this case by adding slack variables to handle pattern recognition.
Gaussian data drawn from two classes
Training examples below margin

margins

Optimal hyperplane
Formulation with slack variables

Optimal margin classifier with slack variables and kernel functions described by Support Vector Machine (SVM).

\[
\min_{(w, \xi)} \frac{1}{2}||w||^2 + \gamma \sum \xi(i) \\
\text{subject to } \xi(i) \geq 0 \ \forall i, \ d(i) (w^T x(i) + b) \geq 1 - \xi(i), \ \forall i, \ \text{and } \gamma > 0.
\]

In dual space

\[
\max \ W(\alpha) = \sum \alpha(i) - \frac{1}{2} \sum \alpha(i)\alpha(j) d(i) d(j) x(i)^T x(j) \\
\text{subject to } \gamma \geq \alpha(i) \geq 0, \ \text{and } \sum \alpha(i) d(i) = 0.
\]

Weights can be found by \( w = \sum \alpha(i) d(i) x(i). \)
Solving QP Problem

- Quadratic programming problem with linear inequality constraints.
- Optimization problem involves searching space of feasible solutions (points where inequality constraints satisfied).
- Can solve problem in primal or dual space.
References

- Kernel website: http://www.kernel-machines.org/