Perceptron Learning Algorithm

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

1) $w(0) =$ arbitrary, $j=1$, $k=0$
2) Pick point $(x(j),d(j))$.
3) If $w(k)^T x(k)d(k) > 0$ 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.
PLA comments

- Energy function
  \[ J(w) = - (\text{sum of synaptic strengths of misclassified points}) \]
  \[ w(k+1) = w(k) - \mu(k) \nabla J(w(k)) \] (gradient descent)
- Margins
- Proof (Novikoff, requires margins)
- Homogeneously linearly separable
- Version space (weight space where feasible solutions lie)
Energy Function

- \( J(w) = - \sum_{x \in \mathcal{M}} (w^T x_d) \) where \( \mathcal{M} \) is the set of misclassified points by \( w \).
- Can find optimal solution using iterative gradient descent algorithm with
  \[ w(k+1) = w(k) - \mu(k) \nabla J(w(k)) \] (gradient descent)
- Gradient given by \( \nabla J(w(k)) = - \sum_{x \in \mathcal{M}} x_d \) (use all misclassified points).
- Perceptron algorithm given uses estimate of gradient (one misclassified training input).
Assume $||w||=1$ and let margin be $\gamma(i) = d(i) x(i)^Tw$, then $\gamma(i)$ is distance from $x(i)$ point to hyperplane.

If minimum margin $> 0$, then training points are linearly separable.
Novikoff’s Theorem

Let $S$ be a nontrivial training set and assume zero threshold with

- $w^*$ a solution
- $||w^*|| = 1$
- $w(0) = 0$

Let $\max ||x(k)|| = \beta$ be finite and $\min d(k)x(k)^Tw^* = \gamma > 0$.

The number of updates will be upper bounded by $k \leq (\beta/\gamma)^2$. 
Proof of Novikoff’s Theorem

\[ <w(k), w^*> = <w(k-1) + \mu x(k-1)d(k-1), w^*> \geq <w(k-1), w^*> + \mu \gamma \geq \mu k \gamma. \]

\[ ||w(k)||^2 \leq ||w(k-1)||^2 + ||\mu x(k-1)||^2 \leq ||w(k-1)||^2 + \mu^2 \beta^2 \leq k \mu^2 \beta^2. \]

\[ ||w^*||^2 k \mu^2 \beta^2 \geq ||w^*||^2 ||w(k)||^2 \geq <w(k), w^*>^2 \geq (k \mu \gamma)^2. \]

Since \( ||w^*|| = 1 \), then \( k \leq (\beta / \gamma)^2. \)
Version space is a set $\mathcal{W}$, such that for any $w \in \mathcal{W}$, all training points are correctly labeled.
Weight characteristics

- If $w(0)=0$, then $w^* = \sum \alpha(k)x(k)$. 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.
- 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.