Pegasos: an Online Algorithm for Learning Support Vector Machines

An equivalent problem to SVMs:

$$\min_{\theta} \left( \frac{1}{2} \|\theta\|^2 + C \sum_i f(y_i(\theta \cdot x_i)) \right)$$

where \( f(z) = \max(0, 1 - z) \)

We’ll first rewrite this in a slightly different form:

$$\min_{\theta} \left( \frac{\lambda}{2} \|\theta\|^2 + \frac{1}{n} \sum_i f(y_i(\theta \cdot x_i)) \right)$$

where \( f(z) = \max(0, 1 - z) \)
The Pegasos Algorithm (Shalev-Shwartz et al 2010, 2007)

- Inputs: training set \( \{(x_i, y_i)\}_{i=1}^n \), \( T \)
- Initialization: \( \theta_1 = 0 \)
- For \( t = 1 \ldots T \):
  1. Pick an example \( i \in \{1 \ldots n\} \) uniformly at random
  2. If \( y_i(\theta_t \cdot x_i) < 1 \)
     \[
     \theta_{t+1} = \left(1 - \frac{1}{t}\right) \theta_t + \frac{1}{\lambda t} y_i x_i
     \]
     else if \( y_i(\theta_t \cdot x_i) \geq 1 \)
     \[
     \theta_{t+1} = \left(1 - \frac{1}{t}\right) \theta_t
     \]
- Return \( \theta_{T+1} \)
Guarantees for Pegasos

\[ g(\theta) = \frac{\lambda}{2}||\theta||^2 + \frac{1}{n} \sum_i f(y_i(\theta \cdot x_i)) \]

- Define \( \theta^* = \arg \min_{\theta} g(\theta) \)
- Assume for all examples \( ||x_i|| \leq R \).
- With high probability, after \( T \) iterations of Pegasos we have

\[ g(\theta_{T+1}) \leq g(\theta^*) + \frac{CR^2 \log T}{\lambda T} \]

where \( C > 1 \) is some constant

(Note: a precise statement is a little more involved than this, but this is basically the correct result)
Deriving Pegasos

\[ g(\theta) = \frac{\lambda}{2} ||\theta||^2 + \frac{1}{n} \sum_i f(y_i(\theta \cdot x_i)) \]

- Batch gradient descent:
  1. Set \( \theta_1 = 0 \)
  2. For \( t = 1 \ldots T \)
     - Calculate \( \nabla_t = \frac{d}{d\theta} g(\theta_t) \)
     - Set \( \theta_{t+1} = \theta_t - \eta_t \nabla_t \) where \( \eta_t > 0 \) is a step size
  3. Return \( \theta_{T+1} \)
Deriving Pegasos (continued)

\[ g(\theta) = \frac{\lambda}{2} ||\theta||^2 + \frac{1}{n} \sum_i f(y_i(\theta \cdot x_i)) \]

▶ **Stochastic** gradient descent:

1. Set \( \theta_1 = 0 \)
2. For \( t = 1 \ldots T \)
   - Choose an \( i \in \{1 \ldots n\} \) at random, define
     \[ g_i(\theta) = \frac{\lambda}{2} ||\theta||^2 + f(y_i(\theta \cdot x_i)) \]
     This is an approximation to \( g(\theta) \) based on example \( i \) alone.
   - Calculate \( \nabla_t = \frac{d}{d\theta} g_i(\theta_t) \)
   - Set \( \theta_{t+1} = \theta_t - \eta_t \nabla_t \) where \( \eta_t > 0 \) is a step size
3. Return \( \theta_{T+1} \)
But what is the Gradient of $g_i(\theta)$?

$$g_i(\theta) = \frac{\lambda}{2} \|\theta\|^2 + f(y_i(\theta \cdot x_i))$$

- Clearly

$$\frac{d}{d\theta} \left( \frac{\lambda}{2} \|\theta\|^2 \right) = \lambda \theta$$

- But $f(z) = \max\{0, 1 - z\}$ is not differentiable. However, a sub-gradient of $f(y_i(\theta \cdot x_i))$ is

$$-y_i x_i \quad \text{if} \quad y_i(\theta \cdot x_i) < 1; \quad 0 \quad \text{otherwise}$$

- Hence a subgradient of $g_i(\theta)$ is

$$\lambda \theta - 1\{y_i(\theta \cdot x_i) < 1\} y_i x_i$$

where $1\{\ldots\}$ is 1 if \ldots is true, 0 otherwise
Putting it All Together

- If $\nabla_t = \lambda \theta_t - 1 \{ y_i (\theta_t \cdot x_i) < 1 \} y_i x_i$ is the sub-gradient at iteration $t$
- And we use the update

$$\theta_{t+1} = \theta_t - \eta_t \nabla_t$$

where

$$\eta_t = \frac{1}{\lambda t}$$

then we have precisely the Pegasos updates