Lecture 7, MIT 6.867 (Machine Learning), Fall 2010

Michael Collins

January 25, 2012

Pegasos: an Online Algorithm for Learning Support Vector Machines

An equivalent problem to SVMs:

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

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

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

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

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

The Pegasos Algorithm (Shalev-Shwartz et al 2010, 2007)

- Inputs: training set $\{(\underline{x}_i, y_i)\}_{i=1}^n$, T
- Initialization: $\underline{\theta}_1 = \underline{0}$
- For $t = 1 \ldots T$:
  1. Pick an example $i \in \{1 \ldots n\}$ uniformly at random
  2. If $y_i(\underline{\theta}_t \cdot \underline{x}_i) < 1$

     $$\underline{\theta}_{t+1} = \left(1 - \frac{1}{t}\right)\underline{\theta}_t + \frac{1}{\lambda t} y_i \underline{x}_i$$

     else if $y_i(\underline{\theta}_t \cdot \underline{x}_i) \geq 1$

     $$\underline{\theta}_{t+1} = \left(1 - \frac{1}{t}\right)\underline{\theta}_t$$

- Return $\underline{\theta}_{T+1}$

Guarantees for Pegasos

$$g(\underline{\theta}) = \frac{\lambda}{2}||\underline{\theta}||^2 + \frac{1}{n}\sum_i f(y_i(\underline{\theta} \cdot \underline{x}_i))$$

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

$$g(\underline{\theta}_{T+1}) \leq g(\underline{\theta}^*) + \frac{CR^2 \log T}{\lambda T}$$

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

$$g(\underline{\theta}) = \frac{\lambda}{2}||\underline{\theta}||^2 + \frac{1}{n}\sum_i f(y_i(\underline{\theta} \cdot \underline{x}_i))$$

- Batch gradient descent:
  1. Set $\underline{\theta}_1 = \underline{0}$
  2. For $t = 1 \ldots T$
     - Calculate
       $$\nabla_t = \frac{d}{d\underline{\theta}} g(\underline{\theta}_t)$$
     - Set $\underline{\theta}_{t+1} = \underline{\theta}_t - \eta_t \nabla_t$ where $\eta_t > 0$ is a step size
  3. Return $\underline{\theta}_{T+1}$

Deriving Pegasos (continued)

$$g(\underline{\theta}) = \frac{\lambda}{2}||\underline{\theta}||^2 + \frac{1}{n}\sum_i f(y_i(\underline{\theta} \cdot \underline{x}_i))$$

- **Stochastic** gradient descent:
  1. Set $\underline{\theta}_1 = \underline{0}$
  2. For $t = 1 \ldots T$
     - Choose an $i \in \{1 \ldots n\}$ at random, define

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

       This is *an approximation to* $g(\underline{\theta})$ *based on example* i *alone*.
     - Calculate $\nabla_t = \frac{d}{d\underline{\theta}} g_i(\underline{\theta}_t)$
     - Set $\underline{\theta}_{t+1} = \underline{\theta}_t - \eta_t \nabla_t$ where $\eta_t > 0$ is a step size
  3. Return $\underline{\theta}_{T+1}$

But what is the Gradient of $g_i(\underline{\theta})$?

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

- Clearly

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

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

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

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

$$\lambda\underline{\theta} - \mathbf{1}\{y_i(\underline{\theta} \cdot \underline{x}_i) < 1\}y_i\underline{x}_i$$

where $\mathbf{1}\{\ldots\}$ is 1 if $\ldots$ is true, 0 otherwise

Putting it All Together

- If $\nabla_t = \lambda \underline{\theta}_t - \mathbf{1}\{y_i(\underline{\theta}_t \cdot \underline{x}_i) < 1\} y_i \underline{x}_i$ is the sub-gradient at iteration t
- And we use the update

$$\underline{\theta}_{t+1} = \underline{\theta}_t - \eta_t \nabla_t$$

where

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

then we have precisely the Pegasos updates