Sequential Minimal Optimization

Sean Thomas
Support Vector Machines

- Original problem (linear, hard margin):
  \[
  \min_{\mathbf{w}, \mathbf{b}} \|\mathbf{w}\| \quad \text{subject to} \quad y_i (\mathbf{w} \cdot \mathbf{x}_i - \mathbf{b}) - 1 \geq 0
  \]
  - where \( \mathbf{w} \cdot \mathbf{x}_i - \mathbf{b} \) is the classifier

- Dual problem (non-linear, soft margin):
  \[
  \max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j k_\lambda (\mathbf{x}_i, \mathbf{x}_j) \alpha_i \alpha_j \quad \text{subject to}
  \]
  
  - \( 0 \leq \alpha_i \leq C \) and \( \sum_{i=1}^{n} y_i \alpha_i = 0 \)
How do they relate?

- *Knowledge Discovery with Support Vector Machines by Lutz Hamel*
  - Great explanation of prerequisites
- Substitute for the margin vector after introducing the Lagrangian multipliers

\[
\bar{w} = \sum_{i=1}^{n} \alpha_i y_i \bar{x}_i
\]
Karush Kuhn Tucker Conditions

- safely classified
  - $\alpha_i = 0 \Rightarrow y_i f(\bar{x}_i) \geq 1$

- in the danger zone
  - $0 < \alpha_i < C \Rightarrow y_i f(\bar{x}_i) = 1$

- incorrectly classified
  - $\alpha_i = C \Rightarrow y_i f(\bar{x}_i) \leq 1$
Sequential Minimal Optimization

  - training a SVM is solving a QP problem
    - breaks problem down into many smallest possible QP problems that can be solved analytically
    - avoids larger numerical QP problems that previous algorithms used

- Implementation used in Weka
  - Look for SMO
Preliminary Results

- Prototype version in R
  - Fairly easy to debug
- First C++ iteration
  - Sparse Vector Implementation
  - Mirrors prototype output
SVM vs. ANN

- ~ 4 hours to understand algorithm
- ~ 8 hours for prototype version
- ~ 8 hours for C++ version

- Days to understand and complete
  - Partly due to inexperience
SVM vs. ANN

- To be determined
  - Hopefully, it will be less of a problem

- Final Product needed
  - Multiple layers
  - Pre-training
  - Data manipulation
SVM vs. ANN

- Firm Mathematical Background
  - Convex function
  - Theory of kernels
  - Lagrangian dual
  - Not as susceptible to noise, especially away from margin

- Neat Biological Inspiration
  - Gradient Descent
    - Local minima
    - Dependent on starting values
  - High Dimensionality requires much more training time
  - Susceptible to noise everywhere
Preliminary Conclusions

• SVM
  – Seems more robust
  – Mathematically appealing

• ANN
  – Much more sensitive