Text Technologies

Machine Learning in IR
Victor Lavrenko

Text Classification
- Predict category / label for a string of text:
  - Is this an email spam or not?
  - Will this be a Sina Weibo post be censored?
  - Does this blogger like or hate our product?
  - Will the Apple stock rise or fall after this news story?
- Have +/− training examples, predict label for new test d
- Centroid (Rocchio) classifier:
  - text d = vector over terms
  - Compute centroids: \( c_+, c_− \)
  - If \( d \cdot c_+ > d \cdot c_− \), predict label is positive
  - Decision rule: \( d \cdot c_+ - d \cdot c_− > 0 \)
  - Or: \( d \cdot w > 0 \) where \( w = c_− - c_+ \)

Large Margin Classifiers
- Learn a decision boundary \( w^t \cdot d > 0 \) iff d is positive
- Problem: many such w (assuming examples separable)
- Maximum-margin: "buffer zone" around boundary
  - As far as possible from nearest training examples: \( d^t w > 0 (+) \)
- Support Vector Machine (SVM)
  - Best classification accuracy
  - Can be slow to train (use SMO/QP)
- Passive Aggressive (PA)
  - Test to train, streaming
  - Accuracy can be lower
- What works in practice:
  - Don't use non-linear versions
  - Don't do feature selection / LSI
  - Use if idf weighting, normalize: \( ||w||=1 \)

Support Vector Machine
- Find boundary vector \( w \) that satisfies conditions:
  \[ \sum_{i=1}^{m} x_i y_i w_i = 1 \text{ (positive)} \]
  \[ \sum_{i=1}^{m} x_i y_i w_i = 1 \text{ (negative)} \]
- Maximize margin around \( w \)
- Convex hull: polytope around all positives/negatives
  - Any \( p = \sum a_i x_i \) is a linear combination of some \( x_i \)
  - Find two nearest points \( h^+, h^− \)
  - Margin must be \( \min \{ ||h^− - h^+|| \} \)
  - Boundary must be halfway
  - Perp to: \( a = \sum a_i x_i \)
  - Looked for nearest points:
    \[ \min ||h^− - h^+|| = \min ||w|| \]

Passive Aggressive Algorithm
- On-line algorithm: learn from massive streams of data
  - Get an example, update classifier, throw away example
- SMO algorithm (simplified)
  \[ \min \|h\|^2 = \min \left\{ \sum a_i x_i - \sum d_i \right\} = \min \left\{ \sum a_i x_i \right\} \]
  \[ \text{subject to } a_i = 0, a_i \geq 0 \]
- Sequential Minimal Optimization (SMO/PR):
  - Initialize \( a_i = 0 \)
  - Repeat until convergence:
    - (1) Pick two weights, e.g. \( a_i, a_j \)
      - Feasible to find most promising pair
    - (2) Compute new \( a_i, a_j \)
      - Loop until no fixed
      - Fixed point solution (basal)