EE 6885 Statistical Pattern Recognition

Fall 2005
Prof. Shih-Fu Chang
http://www.ee.columbia.edu/~sfchang

Lecture 17 (11/23/05)

Today's lecture
- Application of AdaBoost in Face Detection
  - DHS Chap. 9.5

- Classifier Combination
  - DHS Chap. 9.7

Homework #7 due Nov. 30th

Final Exam
- Dec. 16th Friday 1:10-3 pm, Mudd Rm 644
Bias vs. variance for estimator

Assume \( F \) is a quantity whose value is to be estimated.

- Randomly draw \( n \) samples \( D = \{x_1, x_2, \ldots, x_n\} \)
- Learn \( g_D \) to estimate \( F \)

Expected estimation error:

\[
E_D \left[ \left| g_D - F \right| \right]^2 = \left[ E_D (g_D) - F \right]^2 + E_D \left[ \left| g_D - E_D (g_D) \right| \right]^2
\]

Bias\(^2\)  
Variance

Bias vs. variance for classification

- Ground truth: 2D Gaussian

- Complex models have smaller biases, more variances than simple models
- Increasing training pool size helps reduce the variance
- Occam’s Razor principle
Boosting

- For each component classifier, use the subset of data that is most informative given the current set of component classifiers.

Randomly draw a subset of \( n_1 \) samples \( D_1 \)
Use the most informative subset \( D_2 \) from remaining set

Classifier Fusion

Weak classifier \( C_1 \)
Weak classifier \( C_2 \)

...classifier \( C_k \)

AdaBoost (Freund and Shapire)

- Add weak classifiers until low training error has been achieved
- Each training pattern receives a weight determining its chance of being selected for subsequent learning steps.
- If a pattern is correctly classified, then the weight is decreased.

begin with \( W(i) = 1/n, \quad i = 1, \ldots, n \)

\( k = 1, \ldots, k_{\text{max}} \)

train weak classifier \( C_k \) using \( D \) training patterns with weights \( W(i) \)

\[ E_k \leftarrow \text{training error of } C_k \text{ measured on } D \text{ using weights } W(i) \]

\[ \alpha_k \leftarrow \frac{1}{2} \ln \left[ \frac{(1 - E_k)}{E_k} \right] \]

\[ W_{k+1}(i) = \frac{W_k(i)}{Z_k} \times \begin{cases} e^{-\alpha_k}, & \text{if } x_i \text{ is correctly classified} \\ e^{\alpha_k}, & \text{if } x_i \text{ is incorrectly classified} \end{cases} \]

final classification rule

\[ g(x) = \sum_{k=1}^{k_{\text{max}}} \alpha_k h_k(x) > 0, \quad \text{where } h_k(x) \text{ is the predicted } \{\pm 1\} \text{ from } C_k \]
AdaBoost

It can shown that AdaBoost can maximize “margin” rapidly in iterations and thus has good generalization performance over test data.

AdaBoost Face Detection (Viola and Jones, CVPR 2001)

- Rapid face detection for security and HCI applications
  - 2001 performance:
    - 384x288 images 15 frames per second
    - 2 frames per second on iPaq (200MIPS)
- Main contributions
  - New image representation: integral image
  - Allow rapid computation of Harr like filter responses
  - AdaBoost learning for feature selection
    - In each iteration, choose one weak classifier based on one feature only
  - Combine complex classifiers in a cascade way to discard non-interesting regions quickly
Harr filter like features

- Pros and cons?
- Very simple rectangle difference features
- Sum of the pixels in the white area is subtracted from the sum in the grey area
- Number of rectangles can be increased as needed

Rapid computation

- Compute integral image in one pass
- Rectangle sum can be quickly computed

A very large number of features:

- For each 24x24 detection region, there are more than 180,000 features

Each feature as a weak classifier

\[
h_j(x) = \begin{cases} 
1, & \text{if } f_j(x) > \text{ or } < \theta_j \\
0, & \text{otherwise}
\end{cases}
\]

\(X\) is a 24x24 subimage, \(f_j(x)\) is feature

Image processing

- Each subimage is variance normalized to avoid lighting variation

Training:

- 4916 face images scaled and aligned to 24x24 pixels, plus their vertical mirror images
- 10,000 subimages from 9544 non-face images
- Detect faces at multiple scales, with a factor of 1.25 apart, and multiple overlapped scanning locations
AdaBoost Learning

- The first two features after feature selection

---

Cascade classifier for efficiency

- Break a large classifier into cascade of smaller classifiers
  - E.g., 200 features to \{1, 10, 25, 50, 50\}
- Adjust threshold in early stage so that it rejects unlikely regions quickly
- The latter stages are more difficult. They are trained using only the images passing the early stages.
- The final detector has 38 stages over 6000 features
- On average each sunimage uses 10 features

- Design tradeoffs
  - Number of features in each classifier
  - Threshold uses in each classifier
  - Number of classifiers
- Add stages until objective in P-R is met
Performance over MIT-CMU data set

- Voting by multiple classifiers (learned from the same method) helps slightly

<table>
<thead>
<tr>
<th>Detector</th>
<th>10</th>
<th>31</th>
<th>59</th>
<th>65</th>
<th>78</th>
<th>95</th>
<th>167</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viola-Jones</td>
<td>76.1%</td>
<td>88.4%</td>
<td>91.4%</td>
<td>92.0%</td>
<td>92.1%</td>
<td>92.9%</td>
<td>93.9%</td>
</tr>
<tr>
<td>Viola-Jones (voting)</td>
<td>81.1%</td>
<td>89.7%</td>
<td>91.1%</td>
<td>93.1%</td>
<td>93.2%</td>
<td>93.7%</td>
<td>93.7%</td>
</tr>
<tr>
<td>Rowley-Baluja-Kanade</td>
<td>82.2%</td>
<td>86.0%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>89.2%</td>
<td>90.1%</td>
</tr>
<tr>
<td>Schauderma-Kanade</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>94.4%</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kolh-Ying-Alkaya</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>(94.8%)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Mixture of Experts

- Each component classifier is treated as an expert
- The predictions from each expert are pooled and fused by a gating subsystem

\[ P(y | x, \Theta) = \sum_{r=1}^{k} P(r | x, \theta_0) P(y | x, \theta_r) \]

where \( x \) is the input pattern, \( y \) is the output
\( \theta_0 \) controls the gating system; \( \theta_r \) is parameter for component classifier \( r \)

- How to determine \( P(r | x, \theta_0) \), i.e., mixture priors?

- Maximize data likelihood
  - gradient decent or EM

\[ l(D, \Theta) = \sum_i \ln \left( \sum_{r=1}^{k} P(r | x', \theta_0) P(y' | x', \theta_r) \right) \]
Converting output from component classifiers

- Convert various output formats to Prob(detecton or relevance)

- **Rank order** \( g_r = 1 - \text{rank} / N \)

- **Binary label \{1,0\}**
  \[
  g_r = \begin{cases} 
  1 - \varepsilon & \text{if label =1} \\
  \varepsilon & \text{if label =0}
  \end{cases}
  \]

- **Multi-category discriminant values to detect a specific category**
  \[
  \text{softmax } g_r = \frac{e^{g_{r,j}}}{\sum_{j=1}^{C} e^{g_{r,j}}}
  \]

---

Mixture of Experts for Video Retrieval

(Yan, Yang, & Hauptmann 2004)

- Need to fuse retrieval results from tools using different modalities (text, image, concept etc)
**A two-level mixture model**
- Text-based search dominates most of times
- Use audio-visual tools to refine text-based results
- Group non-text tools under one level to avoid performance deterioration

---

**Training data:**
- a set of queries \{q\}, and the relevance labels of each document \{s_i\}

- **First level likelihood**
  \[ P(R|q, s_i) = \lambda_1^n(c_q)P_1^n(R|q, s_i) + \lambda_2^n(c_q)P_2^n(R|q, s_i) \]

- **Second level likelihood**
  \[ P_2^n(R|q, s_i) = f \left( \sum_{k=1}^m \lambda_k^n(c_q)P_k(R|q, s_i) \right) \]

- **Total data likelihood**
  \[ l(\lambda^n; X) = \sum_i \log \sum_{t=1,2} \lambda_t^n P_t^n(R|q, s_i) \]

- **Use E-M to estimate** \( \lambda^n \) and \( \lambda^t \)
  \[ Q(\lambda^n; \lambda^t) = \sum_{t=1,2} \sum_i h_{it} \left( \log \lambda^n + \log P_t^n(R|q, s_i) \right) \]

  \( h_{it} \) is the posterior prob. that document \( s_i \) is generated by classifier \( t \)
E-M for Mixture of Expert for retrieval

**Input:** $P^t(R|q, s_i); t=1, 2, \text{ and } y_t \in \{-1, +1\}$

**Output:** $\sum_{t=1}^{T} \lambda_t P^t(R|q, s_i)$ which optimizes $l(\lambda; X)$.

**Algorithm:**
- Initialize $\lambda^{(0)}$ such that $\forall i, 0 < \lambda_i^{(0)} < 1, \sum \lambda_i^{(0)} = 1$
- For $j = 1, 2, ...$
  1. **E-step:** Compute expectation
     
     $h_{it}^{(j)} = \frac{\lambda_i^{(j)} P(R|q, s_i)}{\sum_i \lambda_i^{(j)} P(R|q, s_i)}$

  2. **M-step:** Update parameter $\lambda_i^{(j+1)} = \frac{1}{T} \sum_j h_{it}^{(j)}$
  3. **M-step:** Maximize the weighted log-likelihood in (7)
  4. Check convergence if $||l(\lambda^{(j+1)}; X) - l(\lambda^{(j)}; X) < \epsilon||$

<table>
<thead>
<tr>
<th></th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>0.3</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0.2</td>
<td>0.4</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>0.6</td>
<td>0.5</td>
<td>2</td>
</tr>
<tr>
<td>$s_4$</td>
<td>0.5</td>
<td>0.8</td>
<td>2</td>
</tr>
</tbody>
</table>

- **E-step:** compute $h$
- **M-step:** find $\lambda$

$h$ is the hidden variable indicating the responsible expert

Example of EM learning of weights

<table>
<thead>
<tr>
<th>Video Retrieval with the Query Independent Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
</tr>
<tr>
<td>Incorrect</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Video Retrieval with the Query-Class Dependent Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
</tr>
<tr>
<td>Incorrect</td>
</tr>
</tbody>
</table>

EE6887-Chang

17-19
Active Learning (Learning with Queries)

- Actively select the next training data that are most informative
  - one that is closest to the decision boundary
  - (or) one that has the most ambiguous confidence scores (e.g., similar discriminant values from two classes)

### Bayesian boundary

Active selection of training data
Nearest neighbor classifiers
Start with the far points in the space

Applications (Active SVM)

- Space for weight $w$
  \[ w^T x_j + b = 0, \quad x_j \text{ support vector} \]

- Constraint added by the new data
  \[ w^T x_j + b = 0 \]

- In image retrieval
  - first train a SVM from labeled data
  - now in interactive retrieval
  - select a new sample and present it to user
  - user label the new data
  - use the new label to re-train the weight $w$
  - which sample to choose?

  Choose the un-labeled sample that is closest to the current separation plane. Why?