COMBINING CLASSIFIERS:
BAGGING and BOOSTING

Danica Kragic

Pattern Recognition .....
Pattern Recognition

Based on:

- Class-conditional pdfs:
  - Assume a model for pdf and estimate parameters
  - No model assumption: histogram methods, k-nearest neighbours, kernel methods, Bayesian networks
- Discriminant functions: Linear Discriminant Functions (LDA), Support Vector Machines (SVM)
- Similarity measures with stored samples
Discriminant functions

1. Choose class of decision functions in feature space.
2. Estimate the function parameters from the training set.
3. Classify a new pattern on the basis of this decision rule.

$$h(x) = w^T x + w_0 = \sum_i w_i x_i + w_0$$

**Linear Discriminant Functions**

$$h(x) = w^T x + w_0$$

If $h(x) > 0$, $x$ is classified as $k_1$; if $h(x) < 0$, $x$ is classified as $k_2$.

If we define $z = (x_1, x_2, ..., x_p)$, $v = (w_0, w_1, ..., w_p)$, then the function can be written as $h(x) = v^T z$.

Webb writes this as follows. Suppose $y_i = z_i$ for $x_i \in \omega_1$ and $y_i = -z_i$ for $x_i \in \omega_2$, then we seek for a value of $v$ for which $v^T y_i > 0$ for all $y_i$ corresponding to the $x_i$ in the train set.
Classification problem

- Assume a set $S$ of $N$ instances $x_i \in X$ each belonging to one of $M$ classes $\{c^1, \ldots, c^M\}$
- The training set consists of pairs $(x_i, c_i)$
- A classifier $C$ assigns a classification $C(x) \in \{c^1, \ldots, c^M\}$ to an instance $x$
- The classifier learned in trial $i$ is denoted $C^i$ while $C^n$ is the composite bagged or boosted classifier

Bagging and Boosting

- Bagging and Boosting aggregate multiple hypotheses generated by the same learning algorithm invoked over different distributions of training data [Breiman 1996, Freund and Schapire 1996]
- Bagging and Boosting generate a classifier with a smaller error on the training data as it combines multiple hypotheses which individually have a large error

Linear Discriminant Analysis

- Possible classes: $\{\}$
- $LDA: w_1 x_1 + w_2 x_2 + w_3 > 0$
- otherwise

Bagging and Boosting

- Bagging replicates training sets by sampling with replacement from the training instances.
- Boosting uses all instances but weights them and therefore produces different classifiers.
- Classifiers are then combined by voting to create a composite classifier.
- Bagging: classifiers have same votes;
- Boosting: vote dependant on the classifiers’ accuracy
Bagging

- For each trial, generate new training set of size $N$ with replacements (multiple occurrence of instances)
- A classifier $C^t$ is generated for each training set
- Final classifier $C^*$ is formed by aggregating the $T$ classifiers
- An instance $x$ is classified by counting votes for which
  $$C^t(x) = k$$
  and $C^*(x)$ represents the class with most votes
Boosting

- Boosting maintains a weight $w_i$ for each instance $\langle x_i, \ldots, c_i \rangle$ in the training set
- The higher the weight $w_i$, the more the instance $x_i$ influences the next hypotheses learned
- At each trial, the weights are adjusted to reflect the performance of the previously learned hypothesis, with the result that the weight of correctly classified instances is decreased and the weight of incorrectly classified instances is increased

$\alpha_t = \frac{1}{2} \ln((1 - \varepsilon_t)/\varepsilon_t)$

$C_{BO}(x) = \arg\max_{c_j \in C} \sum_{t=1}^{T} \alpha_t \delta(c_j, c_t(x))$

- Each hypothesis vote is a function of its accuracy
Boosting

- The error of the hypothesis $C^t$ is measured with respect to the weights
  \[ \varepsilon_t = \sum_{i \text{ such that } C_t(x_i) \neq c_i} w_t^i / \sum_i w_t^i \]
  \[ \alpha_t = \frac{1}{2} \ln\left(\frac{1 - \varepsilon_t}{\varepsilon_t}\right) \]
- Update the weights $w_t^i$ of correctly and incorrectly classified instances by
  \[ w_{t+1}^i = w_t^i \cdot e^{-\alpha_t} \quad \text{if} \quad C^t(x_i) = c_i \]
  \[ w_{t+1}^i = w_t^i \cdot e^{\alpha_t} \quad \text{if} \quad C^t(x_i) \neq c_i \]
- Afterwards normalize the $w_{t+1}^i$ such that they form a proper distribution $\sum_i w_{t+1}^i = 1$

Bayes MAP Hypothesis

Bayes MAP hypothesis for two classes $x$ and $o$
red: incorrect classified instances
Boosted Bayes MAP Hypothesis

- Boosted Bayes MAP hypothesis has more complex decision surface than individual hypothesis alone

**Toy Example**

weak classifiers = vertical or horizontal half-planes

**Round 1**

- $h_1$
- $D_2$

$\epsilon_1=0.30$
$\alpha_1=0.42$

**Round 2**

- $h_2$
- $D_3$

$\epsilon_2=0.21$
$\alpha_2=0.65$
Round 3

Final Classifier

\[ H_{\text{final}} = \text{sign} \left( \begin{array}{c} 0.42 \\ + 0.65 \\ + 0.92 \end{array} \right) \]

\[ v_3 = 0.14 \]
\[ \alpha_3 = 0.92 \]

Danica Kragic, 2004