Lecture 38
Mixture of Experts Neural Network

Outline

- Committee Classifier
  - Linear Committee Classifiers
  - General Committee Classifiers
- Mixture of Experts Based Approach
  - Gating Network
  - Design Procedure
- Examples

The Concept of A Committee

- Decomposition of a learning task into subtasks and learned by cooperative modules.
- Committee Machine – A committee of expert classifiers to perform classification task jointly.

Potential Benefits

Potential Benefit

- Better overall performance
- Reuse existing pattern classification expertise
- Heterogeneity
  - Expert classifiers need not be of the same type.
  - Different features can be used for different classifiers.
- Anonymity:
  - Black-box, proprietary expert classifiers can be used

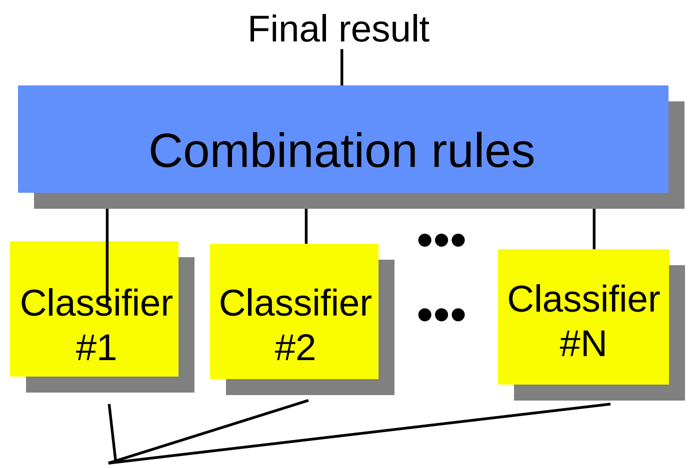
Potential pitfalls

- Higher computation cost

Combination Rules

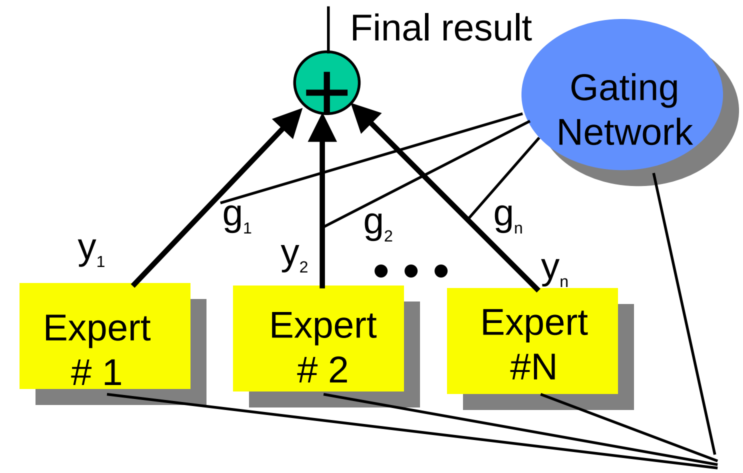
- Committee Classifier

  Unconditional combination rules: <u>independent</u> of individual feature vector

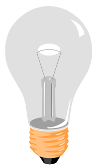


- Mixture of experts Classifier

  Conditional combination rules (gating network): <u>depend</u> on individual feature vector



Committee Classifier



Basic Ideas:

Combination rule = A meta classifier

Output of each expert = meta feature

Combination rules:

- (weighted) linear combination
- (weighted) voting,
- Stack generalization (classifier of classifier).

Linear Committee Classifier:
Minimum Variance Estimate

$$\boxed{y(x) = \sum_i w(x,i)y(x,i)}$$

- Assume each expert gives an UNBIASED estimate of the posterior prob., i.e., $E\{\varepsilon(x,i)\} = 0$; and the co-variance $E\{\varepsilon(x,i)\ \varepsilon(x,k)\ |x\} = \sigma_i^2(x)\delta_{k,i}$ is known.
- Find $\{w(x,i);\ 1 \le i \le n\}$ subject to $\sum_i w(x,i) = 1$ such that $\mathrm{Var}\{y(x)\} = \mathrm{Var}\{\sum_i w(x,i)y(x,i)\}$ is minimized.
- <u>Optimal solution</u>: Let $C = 1/\sum_k [1/\sigma_k^2(x)]$, then $\mathrm{Var}\{y(x)\} \ge C$, and $w(x,i) = C/\sigma_i^2(x)$
  Note that $C \le \sigma_k^2(x)$.

Minimum Variance Solution

$\mathrm{Var}\{y(x)\} = \mathrm{Var}\{\sum_i w(x,i)y(x,i)\} = \sum_i w^2(x,i)\mathrm{Var}\{y(x,i)\}$

Use Lagrange multiplier, solve unconstrained optimization problem:

$$C = \sum_k w^2(x,k)\, \sigma_k^2(x) + \lambda\, (1 - \sum_k w(x,k))$$

Solution: $w(x,i) = P/\, \sigma_i^2(x)$ where

$$P = \sum_k [1/\sigma_k^2(x)] = \text{min. } \mathrm{Var}\{y(x)\}$$

If $\{\varepsilon(x,i)\}$ are correlated, $w(x,i)$ may assume negative values in order to minimize $\mathrm{Var}\{y(x)\}$.

Stack Generalization – Nonlinear Committee Machines

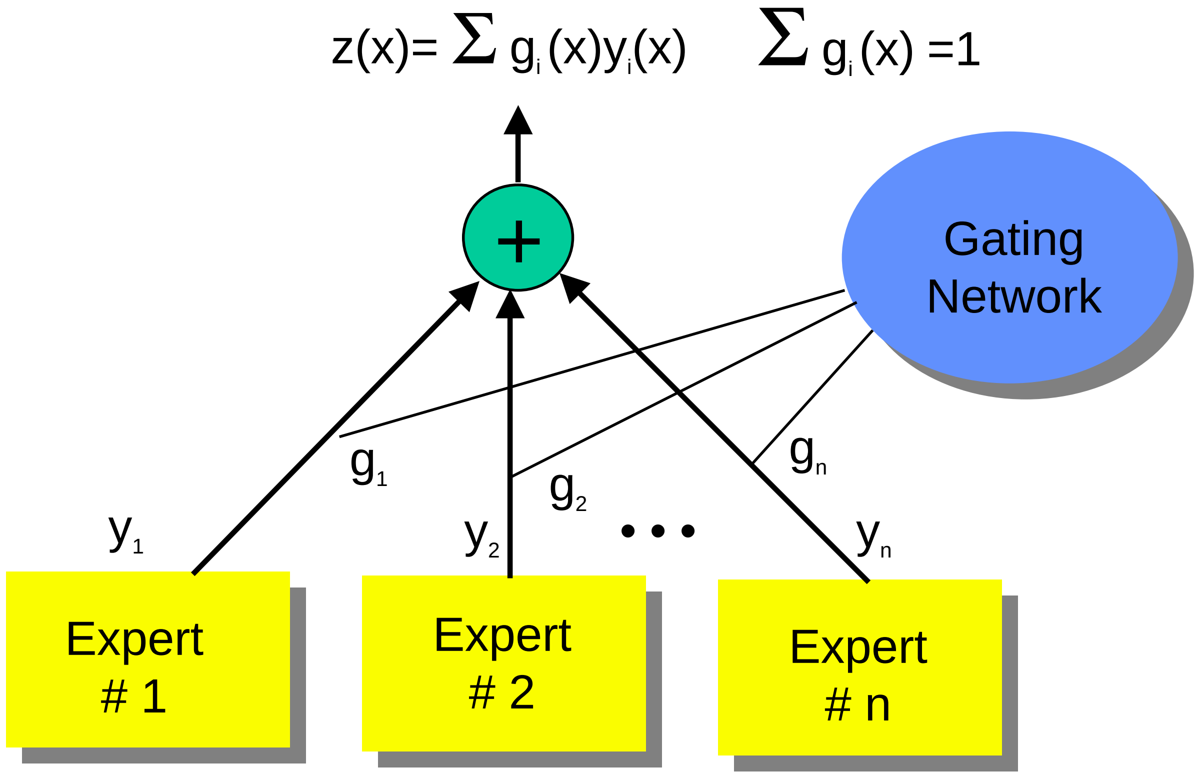
- Treat output of experts as new features
- Perform pattern classification on these new features
- A “classifier of classifier” approach!
- Most general combination rules.
- Results mixed.
- Aliasing Problem:
  – Same feature vector (composed of output of expert classifiers) with different labels.

Examples of Aliasing Problem

x1	x2	T	y1	y2	y3
0	0	0	0	1	0
0	1	1	0	1	1
1	0	1	1	1	0
1	1	0	1	0	0

If only y1 and y2 are used, it is impossible to tell whether y1=0 and y2 = 1 implies T = 0 or T = 1!

Mixture of Expert Network

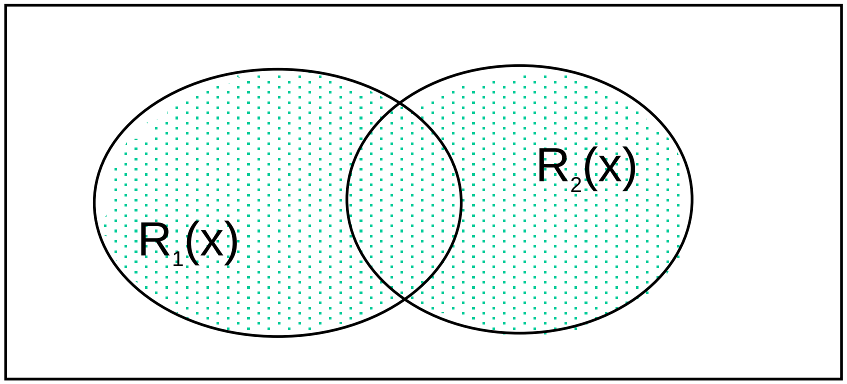


Are Many Experts Better Than One?

- Theorem:

  Denote $R_i(x)$ to be the region in feature space x that classifier #i classifies correctly. Then, the region of correct classification of the MoE classifier

$$R(x) \subseteq \cup R_i(x)$$



Mixture of Expert Model

$z(x) = P\{y|x\} = \sum_i P\{y ,Ei|x\}$

$= \sum_i P\{y|Ei,x\}P\{Ei|x\}$

$= \sum_i [P\{y,Ei,x\}/P\{Ei,x\}][P\{Ei,x\}/P\{x\}]$

$= \sum_i y(x,i)\, g(x,i)$

P{y|x, Ei}: Ei's estimate of posterior Pr. given x.

P{Ei|x}: (Conditional) prior Pr. that p{y|x} is contributed by expert classifier Ei

Gaussian Mixture Model

Assume $y(x,i) \sim \exp\{-0.5[x-m(i)]^T\Sigma^{-1}(i)\,[x-m(i)]\}$, and $g(x,i) = w(i)$ is indep. of x, then

$$z(x) = \sum_i y(x,i)\, w(i)$$

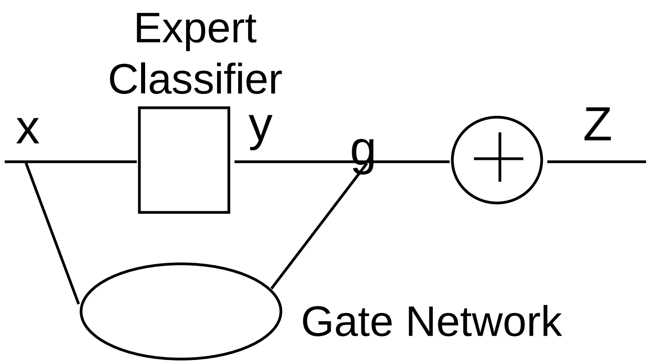
is a Gaussian mixture.

Given $\{m(i),\Sigma(i);\ 1 \le i \le n\}$, and t(x) for $x \in$ training set, w(i) can be found via least square solution:

$$Y\,W = T$$

$\{m(i),\Sigma(i);\ 1 \le i \le n\}$ are often found via clustering of training set data using unsupervised learning.

MoE Design:
An Over-parameterized Problem



Z = g y

Given Z, find g and y?

-> Many possible answers!

1. Assume Expert Classifiers (y) are fixed.
   Find g such that z achieves highest performance.
2. Assume Gating network (g) is fixed.
   Find y to maximize performance of z.
3. Fine tune g and y simultaneously.

Approach I. Fixed Experts

- Assume Experts classifiers' output ($0 \le y(x,i) \le 1$) are given and not to be changed.
- Task: For each x in training set, choose $g_i(x)$ to minimize

$$\| T(x) - z(x) \| = \| T(x) - \Sigma_i y(x,i)\, g(x,i) \|$$

subject to: $\Sigma_i g(x,i) = 1$, and $0 \le g(x,i) \le 1$

Solution: $g(x,i^*) = 1$, If $\| T(x) - y(x,i^*) \| \le \| T(x) - y(x,i) \|$ for $i \ne i^*$, $= 0$ otherwise

- When all experts' opinions (y(x,i)) are fixed for each x, the best strategy (under the contraint) is to pick the winner!

SoftMax Synaptic Weights

- Generalized linear model:

$$g_i(x) = \exp[v_i^t x] \,/\, \Sigma_k \exp[v_k^t x]$$

- Radial basis network model:

$$g_i(x) = \exp[-\|x-v_i\|^2] \,/\, \Sigma_k \exp[-\|x-v_i\|^2]$$

- Both satisfy $\sum_i g(x,i) = 1$, and $0 \le g(x,i) \le 1$.
- $\{v_k\}$: parameters to be estimated from data.
- Gradient descent algorithm can be used to find v_i.

Approach II. Fixed Gating Network

- Assume: g(x,i) are specified for each x in training set.

  If g(x,i)= 0, no restriction on y(x,i).

  If g(x,i) > 0, y(x,i) = T(x).
- Separate training: For expert i, derive a (hopefully simpler) training set = {(x,T(x))| g(x) > 0}, and train each expert independently, may be in parallel!
- Joint training: Let z(x) = Σ_iy(W(i),x,i) g(x,i) then

  W(i,t+1) = W(i,t) + η e(i,t)g(x,i){dy(W(i,t),x,i)/dW(i)}

  gradient descent with add'l g(x,i)!

Approach III. Iterative Approach

- Initiation: Cluster training data into clusters.

  Initialize each expert classifier by training it with data from a single cluster.

  Initialize each corresponding gating network by training it so that g(x,i) = 1 for that cluster, = 0 otherwise.
- Fix gating network, refine individual classifier using approach II.
- Fix expert classifiers, refine gating network using approach I.
- Repeat until convergence criteria met.