

Tutorial: Pattern Recognition in Acoustic Signal Processing

Mark Hasegawa-Johnson

These slides:

<http://www.isle.uiuc.edu/slides/2009/Hasegawa-Johnson09ASA1.pdf>

ASA Spring Meeting, May 21, 2009



Outline

- 1 Why Use Pattern Recognition?
- 2 Algorithm Selection
- 3 Tutorial: Discriminative Methods
 - Hypothesis Space: Universal Approximators
 - Training Criteria: Differentiable Error Metric
 - Training Algorithm: Chain Rule
 - Wrinkle #1: Recognition, Tracking
 - Wrinkle #2: Small Training Corpus
- 4 Tutorial: Bayesian Methods
 - Hypothesis Space: Latent Variables
 - Training Criteria: Maximum Likelihood, MAP, MaxEnt
 - Training and Inference Algorithms: Bayes' Rule
 - Wrinkle #1: HMM Regression, Switching Kalman Smoothers
- 5 Tutorial: Hybrids
- 6 Conclusions

The Scientific Method

$$y = h(x)$$

Hypothesize-Measure-Test

- 1 Based on knowledge of the physical situation, form:
 - 1 a hypothesis
 - 2 a null hypothesis
- 2 Collect data: $(x_i, y_i), 1 \leq i \leq N$.
- 3 Test the hypothesis: measure $P(\text{data}|\text{null hypothesis})$

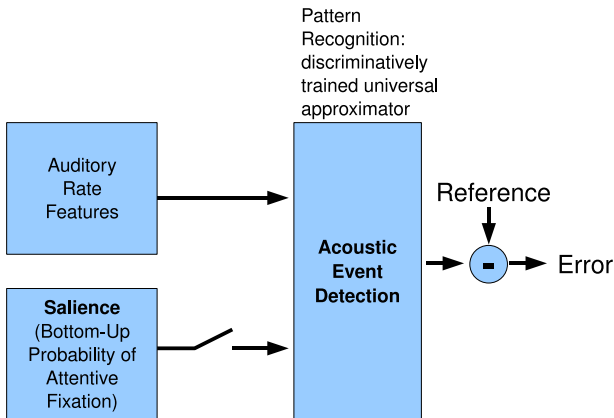
The Pattern Recognition Method

$$y = h(x)$$

Hypothesize-Measure-Learn-Test

- 1 Form an infinite set of hypotheses (called the “hypothesis space”), usually a parameterized universal approximator.
- 2 Collect:
 - 1 training data $((x_i, y_i), 1 \leq i \leq N)$
 - 2 testing data $((x_i, y_i), N + 1 \leq i \leq N + M)$
- 3 Train the hypothesis: maximize $P(\text{hypothesis}|\text{training data})$
- 4 Test the hypothesis: measure $P(\text{testing data}|\text{hypothesis})$

Example: PR in the Scientific Method



Hypothesis 1:
Signal Model,
Saliency of an
Acoustic Event

Hypothesis 2:
Saliency
contributes to
acoustic event

Criteria for Choosing a Pattern Recognizer

① Structure of the Model

- ① **Discriminative Training:** All parameters in the model can be simultaneously adjusted to minimize global error metric
- ② **Bayesian Training:** Components must be separately trained, then combined without blowing up

Criteria for Choosing a Pattern Recognizer

① Structure of the Model

- ① Discriminative
- ② Bayesian

② Size of the Training Database

- ① **Empirical Risk Minimization:** Training database includes 10,000 independent trials; train model to minimize training database error
- ② **Structural Risk Minimization:** Training database smaller than 1000 trials; train model to minimize

$$P(\text{Error}) \leq (\text{Training Corpus Error}) + \lambda \frac{(\text{Model Complexity})}{(\text{Training Corpus Size})}$$

Criteria for Choosing a Pattern Recognizer

① Structure of the Model

- ① Discriminative
- ② Bayesian

② Size of the Training Database

- ① Empirical Risk Minimization
- ② Structural Risk Minimization

③ Dynamic State

- ① $y = h(x)$ has no hidden state (**classification, regression**)
- ② $y = h(x)$ has hidden state (**recognition, tracking**)

Criteria for Choosing a Pattern Recognizer

1 Structure of the Model

- 1 Discriminative
- 2 Bayesian

2 Size of the Training Database

- 1 Empirical Risk Minimization
- 2 Structural Risk Minimization

3 Dynamic State

- 1 $y = h(x)$ has no hidden state (**classification, regression**)
- 2 $y = h(x)$ has hidden state (**recognition, tracking**)

4 Function Range

- 1 $y = h(x)$ is an integer (**classification, recognition**)
- 2 $y = h(x)$ is a real-valued vector (**regression, tracking**)

Discriminative Training—Gradient Descent Methods

- 1 Choose a hypothesis space (a universal approximator)
- 2 Choose a differentiable error metric
- 3 Apply the Chain Rule

Universal Approximators

Universal Approximator: Definition

A parameterized function space $h_{\Theta}(x)$, with parameter vector $\Theta \in \mathbb{R}^{\alpha K}$, is called a universal approximator if for any bounded $h(x)$ with finite domain,

$$\lim_{K \rightarrow \infty} \min_{\Theta} \|h_{\Theta}(x) - h(x)\| = 0$$

Example: Sigmoidal Neural Network

- 1 Sigmoidal Neural Network ($\Theta = \{c_1, \dots, c_K, w_1, \dots, w_K\}$)

$$h_{\Theta}(x) = \sum_{k=1}^K c_k \frac{1}{1 + e^{-x^T w_k}}$$

Universal Approximators

More Examples: Universal Approximators

- 1 Sigmoidal Neural Network ($\Theta = \{c_1, \dots, c_K, w_1, \dots, w_K\}$)
- 2 Mixture Gaussian ($\Theta = \{c_1, \dots, c_K, \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K\}$)

$$h_{\Theta}(x) = \sum_{k=1}^K c_k e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$$

- 3 Classification and Regression Tree (CART), K Nearest Neighbors (KNN), etc.
($\Theta = [b_1, \dots, b_K, w_1, \dots, w_K, \mathcal{R}_1, \dots, \mathcal{R}_K]$)

$$h_{\Theta}(x) = x^T w_k + b_k \text{ if } x \in \mathcal{R}_k$$

\mathcal{R}_k is a region with piece-wise linear boundaries.

Differentiable Error Metric

Minkowski Norm Error Metrics

$$\mathcal{E}_L = \frac{1}{N} \sum_{i=1}^N \|h_{\Theta}(x_i) - y_i\|_L^L$$

The “Best” Metric: The Zero Norm

$$\|h_{\Theta}(x_i) - y_i\|_0 \triangleq \begin{cases} 0 & y_i = h_{\Theta}(x_i) \\ 1 & \text{otherwise} \end{cases}$$

Problem: if $\mathcal{E}_0 \neq 0$, what do we do next?

Differentiable Error Metric

Minkowski Norm Error Metrics

$$\mathcal{E}_L = \frac{1}{N} \sum_{i=1}^N \|h_{\Theta}(x_i) - y_i\|_L^L$$

Differentiable Error Metrics: $L = 1$, $L = 2$

- 1 The One Norm (Manhattan Distance, $L = 1$):

$$\frac{\partial \mathcal{E}_1}{\partial h_{\Theta}(x_i)} = \text{sign}(h_{\Theta}(x_i) - y_i)$$

- 2 The Two Norm (Euclidean Distance, $L = 2$):

$$\frac{\partial \mathcal{E}_2}{\partial h_{\Theta}(x_i)} = h_{\Theta}(x_i) - y_i$$

Apply the Chain Rule

The Error Back-Propagation Algorithm

- 1 **Initialize:** Choose some initial parameter set $\Theta^{(0)}$
- 2 **Iterate:** For $t = 1, \dots$ until $\mathcal{E}(\Theta)$ stops changing:

$$\Theta^{(t)} = \Theta^{(t-1)} - \eta \nabla_{\Theta} \mathcal{E}$$

$$\nabla_{\Theta} \mathcal{E} \triangleq \sum_{i=1}^N \left(\frac{\partial \mathcal{E}}{\partial h_{\Theta}(x_i)} \right) (\nabla_{\Theta} h_{\Theta}(x_i))$$

Wrinkle #1: Recognition, Tracking

A **Recursive Neural Net** is a neural net with one or more hidden state variables:

$$h_{\Theta}(x_i) = \sum_{k=1}^K c_k \frac{1}{1 + e^{-[x_i^T, h_{\Theta}(x_{i-1})]w_k}}$$

Training is performed using **Back-Propagation Through Time**:

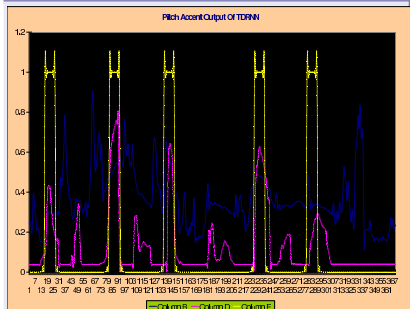
$$\frac{\partial \mathcal{E}_2}{\partial h_{\Theta}(x_i)} = (h_{\Theta}(x_i) - y_i) + (h_{\Theta}(x_{i+1}) - y_{i+1}) \left(\frac{\partial h_{\Theta}(x_{i+1})}{\partial h_{\Theta}(x_i)} \right) + \dots$$

Recursive Neural Nets: Example Application

Task Description

- Blue = x_i (F0=RNN input)
- Yellow = y_i (pitch accent = target RNN output)
- Pink = $f(x_i)$ (RNN-estimated pitch accent probability)

Example Results



Wrinkle #2: Small Training Corpus

Test Corpus Error Bounds based on the Central Limit Theorem

$$P(\text{Error}) \leq \mathcal{E}(X, Y, \Theta) + \mathcal{G}(\Theta)$$

Example Bounds

Minimum Description Length: $K(\Theta)$ describes $h_{\Theta}(x)$ as a binary program, and

$$\mathcal{G}(\Theta) \propto \|K(\Theta)\|_0$$

Support Vector Machines: $\psi(\Theta)$ describes $h_{\Theta}(x)$ as a linear classifier in an augmented feature space, and

$$\mathcal{G}(\Theta) \propto \|\psi(\Theta)\|_2^2$$

Support Vector Machines: Example Application

- Consonant-vowel transitions, and similar manner-change landmarks, are good places to look for information about speech [Stevens, Interspeech 2000]
- Some landmarks occur relatively infrequently—it's hard to learn what they sound like.
- SVMs do the job [Niyogi, Burges, and Ramesh, 1999; Borys and Hasegawa-Johnson, 2005; chance=50%]:

SVM	%ACC	SVM	%ACC
-+Silence	92.1	+ -Silence	91.6
-+Continuant	79.7	+ -Continuant	81.1
-+Sonorant	86.4	+ -Sonorant	91.1
-+Syllabic	88.6	+ -Syllabic	78.5
-+Consonantal	78.1	+ -Consonantal	73.1

Bayesian Methods

- **Bayesian Classification and Recognition:**

$$y^* = \arg \max p_{\theta}(y|x)$$

- **Bayesian Regression and Tracking:**

$$y^* = E \{y|x\}$$

Advantages and Disadvantages

- **Disadvantage:** One must learn $p(x, y)$; this usually requires more data, and is subject to more error, than learning $y = h(x)$ directly
- **Advantage:** Bayesian inference allows modeling of latent variables, state dynamics, and extra sources of information in a principled manner

Hypothesis Space Example: Hidden Markov Model

Let $X = [x_1, \dots, x_N]$ be the observations, let $Y = [y_1, \dots, y_N]$ be the labels. A **Hidden Markov Model** posits the existence of some latent variables $Q = [q_1, \dots, q_N]$ such that

$$h_{\Theta}(X) = \arg \max_Y p_{\Theta}(X, Y)$$

$$p_{\Theta}(X, Y) = \sum_Q \prod_{i=1}^N p_{\Theta}(y_i | y_{i-1}) p_{\Theta}(q_i | q_{i-1}, y_i) p_{\Theta}(x_i | q_i)$$

- $p_{\Theta}(y_i | y_{i-1})$ (the “language model”) is a lookup table
- $p_{\Theta}(q_i | q_{i-1}, y_i)$ (the “pronunciation model”) is a lookup table
- $p_{\Theta}(x_i | q_i, y_i)$ (the “acoustic model”) is a Gaussian, with mean vector μ_q and covariance matrix Σ_q

Learn the Distributions: Maximum Likelihood

Maximum Likelihood Parameter Estimation

$$\Theta = \arg \max \log p_{\Theta}(X, Y)$$

What About Small Training Corpora?

- Maximum Likelihood is a form of **empirical risk minimization**
- Related forms of **structural risk minimization** include
 - **MAP (maximum a posteriori probability)**

$$\Theta = \arg \max (\log p(X, Y|\Theta) + \log p(\Theta))$$

- **MaxEnt(maximum entropy)**

$$\Theta = \arg \max (\log p_{\Theta}(X, Y) + H(p_{\Theta}))$$

Apply Bayes' Rule

Bayes' Rule (a.k.a. the Definition of Conditional Probability)

$$p_{\Theta}(X, Y) = \sum_Q \prod_{i=1}^N p_{\Theta}(y_i | y_{i-1}) p_{\Theta}(q_i | q_{i-1}, y_i) p_{\Theta}(x_i | q_i)$$

Training a Bayesian Classifier: Maximum Likelihood

$$\Theta = \arg \max p_{\Theta}(X, Y)$$

Testing a Bayesian Classifier: Minimum Probability of Error

$$Y = \arg \max p_{\Theta}(X, Y)$$

Bayesian Regression and Tracking

Hidden Markov Regression

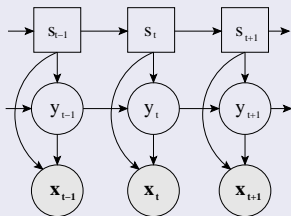
Suppose $y_i \in \mathbb{R}^D$ is a real-valued vector, and (x_i, y_i) are jointly Gaussian:

$$p(x, y|q) \propto \exp \left(-\frac{1}{2} \begin{bmatrix} x_i - \bar{x}_q \\ y_i - \bar{y}_q \end{bmatrix}^T \begin{bmatrix} A_q & B_q \\ B_q^T & C_q \end{bmatrix}^{-1} \begin{bmatrix} x_i - \bar{x}_q \\ y_i - \bar{y}_q \end{bmatrix} \right)$$

then $h(x_i) = \arg \min \mathcal{E}_2$ is

$$E \{y_i | X\} = \sum_{q_i} P(q_i | X) \left(\bar{y}_q + B_q^T A_q^{-1} (x - \bar{x}_q) \right)$$

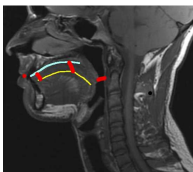
Switching Kalman Smoother



- Setup: exactly like HMM regression, except that (x_i, y_i, y_{i-1}) are jointly Gaussian
- Result: exactly like HMM regression, except that $\bar{y}_{i|q,X}$, $A_{i|q,X}$, and $B_{i|q,X}$ must be updated using interacting multiple Kalman filters:

$$E\{y_i | X\} = \sum_{q_i} P(q_i | X) \left(\bar{y}_{i|q,X} + B_{i|q,X}^T A_{i|q,X}^{-1} (x - \bar{x}_q) \right)$$

Switching Kalman Smoother: Example



Task Description

- x_i is an acoustic spectrum; y_i is the corresponding vector of speech articulator positions
- Results (unpublished):

$$\mathcal{E}_2(\text{Switching Kalman Smoother}) < \mathcal{E}_2(\text{HMM Regression})$$

- Difference is consistent but very small

Hybrid Discriminative-Bayesian Systems

Task Scenario

- y_i is very difficult to classify without dynamic information, (e.g. speech recognition: y_i = words, x_i = short-time spectrum)
- An auxiliary variable f_i can be inferred very accurately using a local classifier (e.g., f_i = phonological distinctive features):

$$\hat{f}_i = h_{\Theta}(x_i)$$

- Training database includes $X = [x_1, \dots, x_N]$,
 $Y = [y_1, \dots, y_N]$, and $F = [f_1, \dots, f_N]$
- Testing database includes only $\tilde{X} = [x_{N+1}, \dots, x_{N+M}]$

Hybrid Training Methods

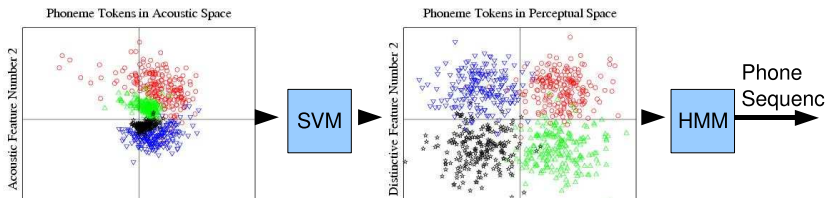
Training Algorithm

- Train $h_{\Theta}(x)$ using discriminative methods
 - Minimize

$$\mathcal{E}_L = \frac{1}{N} \sum_{i=1}^N \|f_i - h_{\Theta}(x_i)\|_L^L$$

- $h_{\Theta}(x_i)$ is a *real-valued vector* that approximates f_i
- Train a probability model $p_{\Theta}(F, X, Y)$ using Bayesian methods
 - Using $p_{\Theta}(F, X, Y)$, Bayesian inference can model dynamics of hidden states, incorporate multiple knowledge sources, etc.

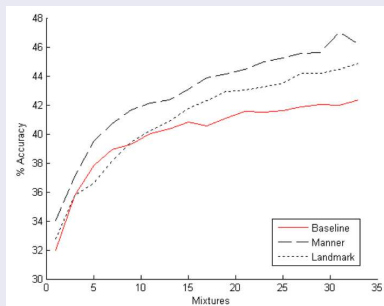
Example: Landmark-Based Speech Recognizer



SVM-HMM Hybrid Landmark-Based Speech Recognizer

- SVM computes a *real-valued distinctive feature* that optimally discriminates between the case $f_i = 1$ (landmark of a specified type is present) and $f_i = -1$ (landmark absent)
- HMM computes $p(\text{phoneme sequence} | \text{landmark sequence})$

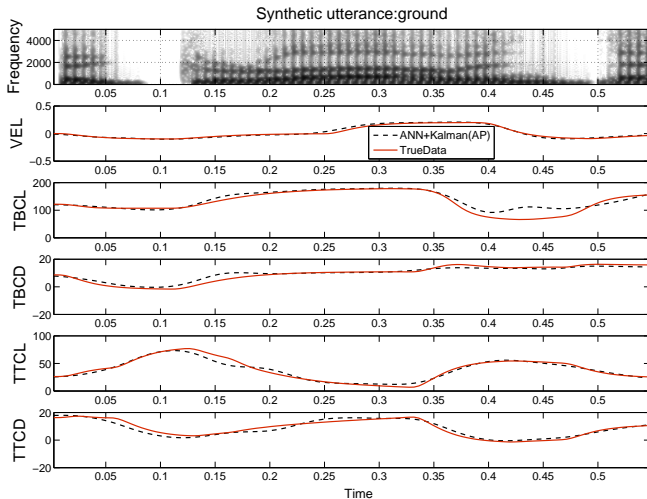
Phone Recognition Accuracy vs. Mixture Size, Telephone Speech



- MFCC = mel frequency cepstral coefficients
- Landmark = detect manner-to-manner landmarks, e.g., obstruent-to-sonorant
- Manner = detect manner-onset landmarks, e.g., onset of sonorant region

Example: RNN with Kalman Smoothing

Mitra et al., in review



Conclusions

- Pattern recognition (especially discriminative training)—it's easy!
 - Choose a hypothesis space (a family of universal approximators)
 - Gradient descent to minimize error
- Bayesian learning simplifies the use of structured models
 - Hidden-state dynamics
 - External sources of information
- Hybrid discriminative-Bayesian methods sometimes give the best of both worlds
 - Discriminative training = minimum error locally
 - Bayesian inference = principled integration of disparate information sources

Thank You!

<http://www.isle.uiuc.edu/slides/2009/Hasegawa-Johnson09ASA1.pdf>