Machine Learning Paradigms for Speech Recognition

Dong Wang

Borrowed from Deng's TASLP 2013 paper

ML and ASR

- ML introduces interesting ideas to ASR
- ASR is a large test bed for ML
- Some techniques are from ASR to ML

Background

- II.A Fundamentals

TABLE I
DEFINITIONS OF A SUBSET OF COMMONLY USED SYMBOLS AND NOTATIONS IN THIS ARTICLE

Symbol	Meaning
$\mathcal{X}$	Space of input vectors
$\mathcal{Y}$	Set of output labels
$p(\mathbf{x}, y)$	Joint distribution $p(\mathbf{X} = \mathbf{x}, Y = y)$
$\mathcal{F}$	Space of decision functions $f : \mathcal{X} \rightarrow \mathcal{Y}$
$f(\mathbf{x}; \lambda)$	Decision function
$d_y(\mathbf{x}; \lambda)$	Discriminant function
λ	Model or decision function parameters
$L(f(\mathbf{x}), y)$	Loss function
$\mathrm{E}_{p(\mathbf{x},y)}[\cdot]$	Expectation $\mathrm{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y)}[\cdot]$
$\mathcal{D}$	Training data $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})_{i=1}^{m}$

Background

- II.A Fundamentals

$$f(\mathbf{x}) = \arg\max_{y} d_y(\mathbf{x}); \qquad (1)$$

$$R_p(f) = \mathrm{E}_{p(\mathbf{x},y)}\left[L\left(f(\mathbf{x}), y\right)\right] \qquad (3)$$

$$R_{\mathrm{emp}}(f) = \frac{1}{m}\sum_{i=1}^{m} L\left(f(\mathbf{x}^{(i)}), y^{(i)}\right) \qquad (4)$$

$$J(f) = R_{\mathrm{emp}}(f) + \gamma C(f) \qquad (5)$$

Background

- II.A Fundamentals

$$q(f) = p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f)p(f)}{p(\mathcal{D})}, \qquad (6)$$

$$f_{Bayes}(\mathbf{x}) \triangleq \mathrm{E}_{q(f)}\left[f(\mathbf{x})\right] \qquad (7)$$

$$q^*(f) = \arg\min_q \left(\mathrm{E}_{q(f)}\left[R_{\mathrm{emp}}(f)\right] + \lambda D\left(q(f)\|p(f)\right)\right) \qquad (8)$$

Background

- II.B Speech recognition: a structured sequence classification problem in machine learning
- II.C A high level summary of machine learning paradigms

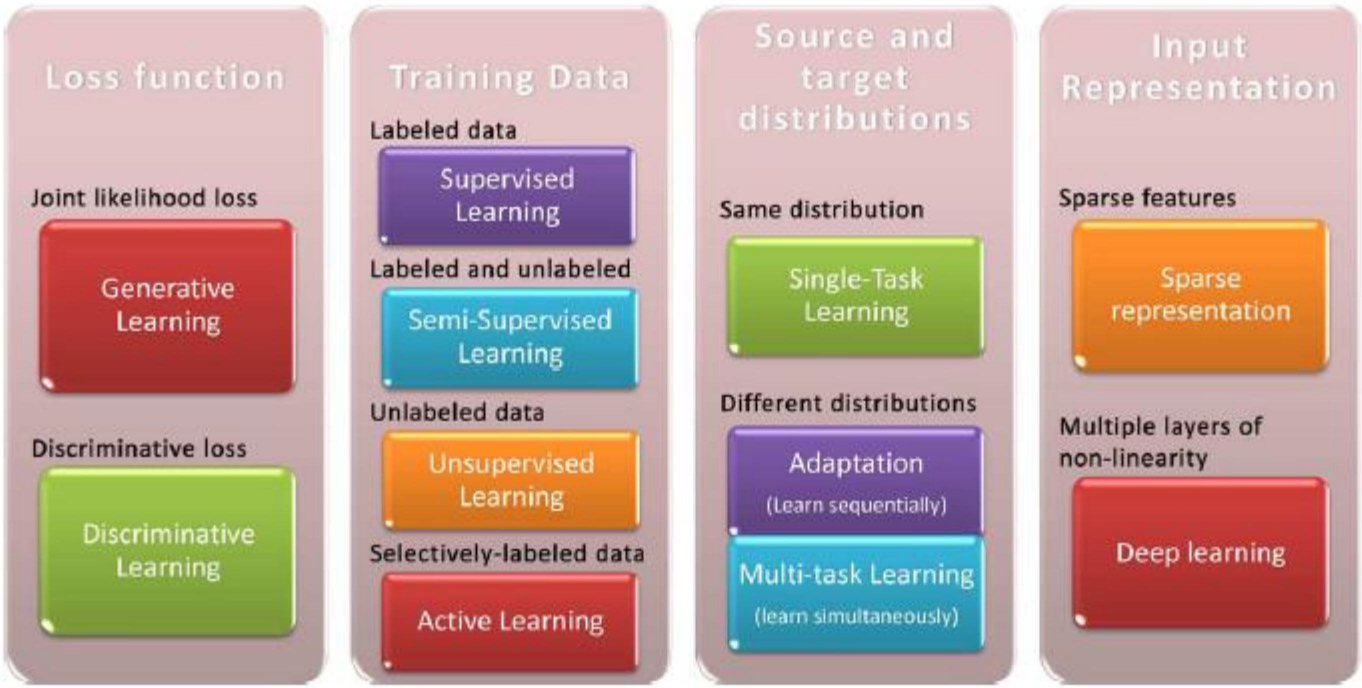


Fig. 1. An overview of ML paradigms and their distinct characteristics.

III. Generative Learning

- Generative learning:
  - Use a generative model and
  - Objective function is based on joint likelihood loss defined on the generative model
- Discriminative learning
  - Using a discriminative model or
  - Applying a discriminative training objective function to a generative model

III. Generative Learning

- III.A Models

$$d_y(\mathbf{x}; \lambda) = \ln p(\mathbf{x}, y; \lambda) = \ln p(\mathbf{x}|y; \lambda)p(y; \lambda) \qquad (9)$$

  - A simple form of generative model leads to simple decision boundary, e.g., LDA
  - Naïve bayes
  - Latent variables can model more complex distributions, pLSA, LDA, GMM
  - Graphical model: directed (HMM) and undirected models (MRF).

III. Generative Learning

- III.B Loss function

$$L\left(f(\mathbf{x}), y\right) = -\ln p(\mathbf{x}, y; \lambda) \qquad (10)$$

  - Factorization
  - MLE training (a) structure correct (b)training data from the true distribution (c) training data is infinite

III. Generative Learning

- III. C generative learning in speech recognition
  - HMM/GMM
  - Baum-welch learning

$$R_{\mathrm{emp}}(f) = -\sum_i \ln p(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}; \pi, A, B) \qquad (11)$$

  - State tying

$$C(f) = \prod_{(m,n)\in\mathcal{T}} \delta(b_m = b_n) \qquad (12)$$

  - PMC, VTS

III. Generative Learning

- III.D Trajectory/Segment models
  - Capture dynamic properties of speech in the temporal dimension more faithfully than HMM
  - Stochastic segmentations, trajectory segmental model, trjactory HMM, hidden dynamic models
  - Some temporal trajectory structure built into the mdoels

$$\mathbf{z}(k+1) = \mathbf{g}_k\left[\mathbf{z}(k), \boldsymbol{\Lambda}_s\right] + \mathbf{w}_s(k) \tag{13}$$

$$\mathbf{o}(k') = \mathbf{h}_{k'}\left[\mathbf{z}(k'), \boldsymbol{\Omega}_{s'}\right] + \mathbf{v}_{s'}(k'). \tag{14}$$

III. Generative Learning

- III.D Trajectory/Segment models

$$\mathbf{z}(k+1) = \mathbf{A}_s \mathbf{z}(k) + \mathbf{B}_s \mathbf{w}_s(k) \quad (15)$$

$$\mathbf{o}(k) = \mathbf{C}_s \mathbf{z}(k) + \mathbf{v}_s(k). \quad (16)$$

  – Difficulties
    - No much science on articulatory mechanism
    - Just generative models
    - No-parametric Bayesian not well studied
    - Limited model assumptions. Isolated dynamic. More Bayesian approach is required

III. Generative Learning

- Dynamic graphical models

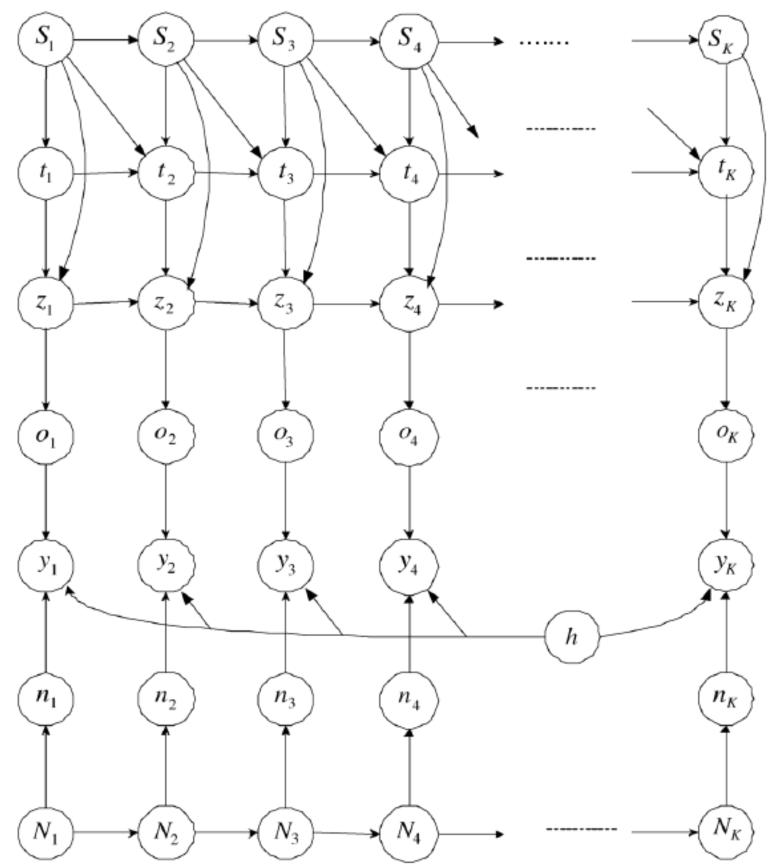
$$p\left[\mathrm{t}(k)|s_k, s_{k-1}, \mathrm{t}(k-1)\right] = \begin{cases} \delta\left[\mathrm{t}(k) - \mathrm{t}(k-1)\right] & if\ s_k = s_{k-1}, \\ \mathcal{N}\left(\mathrm{t}(k); \mathbf{m}(s_k), \mathbf{\Sigma}(s_k)\right) & otherwise. \end{cases} \quad (17)$$

$$p_{\mathbf{z}}\left[\mathbf{z}(k+1)|\mathbf{z}(k), \mathrm{t}(k), \mathbf{s}_k\right] = p_{\mathbf{w}}\left[\mathbf{z}(k+1) - \mathbf{\Phi}_{s_k}\mathbf{z}(k) - (\mathbf{I} - \mathbf{\Phi}_{s_k})\,\mathrm{t}(k)\right], \quad (18)$$

$$\mathbf{z}(k+1) = \mathbf{\Phi}_s\mathbf{z}(k) + (\mathbf{I} - \mathbf{\Phi}_s)\mathrm{t}_s + \mathbf{w}(k). \quad (19)$$

$$\mathbf{o}(k) = \mathbf{h}\left[\mathbf{z}(k)\right] + \mathbf{w}_0(k), \quad (20)$$

$$p_{\mathbf{v}}\left(\mathbf{v}(k)|\mathbf{o}(k), \mathbf{h}, \mathbf{n}(k)\right) = p_{\mathbf{r}}\left[\mathbf{v}(k) - \mathbf{o}(k) + \mathbf{h} + \mathbf{C}\log \times \left[\mathbf{I} + \exp\left[\mathbf{C}^{-1}\left(\mathbf{n}(k) - \mathbf{o}(k) - \mathbf{h}\right)\right]\right]\right]. \quad (21)$$



IV. Discriminative learning

- IV.A Models

$$f(\mathbf{x};\lambda) = -\arg\min_{y'} \sum_{y} \Delta(y', y) p(y|\mathbf{x};\lambda) \qquad (22)$$

- MLP or log linear

$$f(\mathbf{x};\lambda) = \arg\min_{y} p(y|\mathbf{x};\lambda) \qquad (23)$$

$$d(\mathbf{x};\lambda) = \ln p(y|\mathbf{x};\lambda) \qquad (24)$$

t:

- Margin

$$d_y(\mathbf{x};\lambda) = \lambda \cdot \phi(\mathbf{x}, y) \qquad (25)$$

IV. Discriminative learning

- IV.B. Loss functions
  - Probability-based Loss

$$L\left(f(\mathbf{x}), y\right) = -\ln p(y|\mathbf{x}; \lambda). \qquad (26)$$

$$p(\mathbf{y}|\mathbf{x}; \lambda) = \frac{1}{Z_\lambda(\mathbf{x})} \exp \lambda \cdot f(\mathbf{y}, \mathbf{x}). \qquad (27)$$

$$p(\mathbf{y}|\mathbf{x}; \lambda) = \frac{1}{Z_\lambda(\mathbf{x})} \sum_{\mathbf{z}} \exp \lambda \cdot f(\mathbf{y}, \mathbf{z}, \mathbf{x}). \qquad (28)$$

$$L\left(f(\mathbf{x}), \mathbf{y}\right) = -\ln \sum_{\mathbf{y}} \Delta(\mathbf{y}', \mathbf{y}) p(\mathbf{y}|\mathbf{x}; \lambda) \qquad (29)$$

  - Margin-based Loss

IV. Discriminative learning

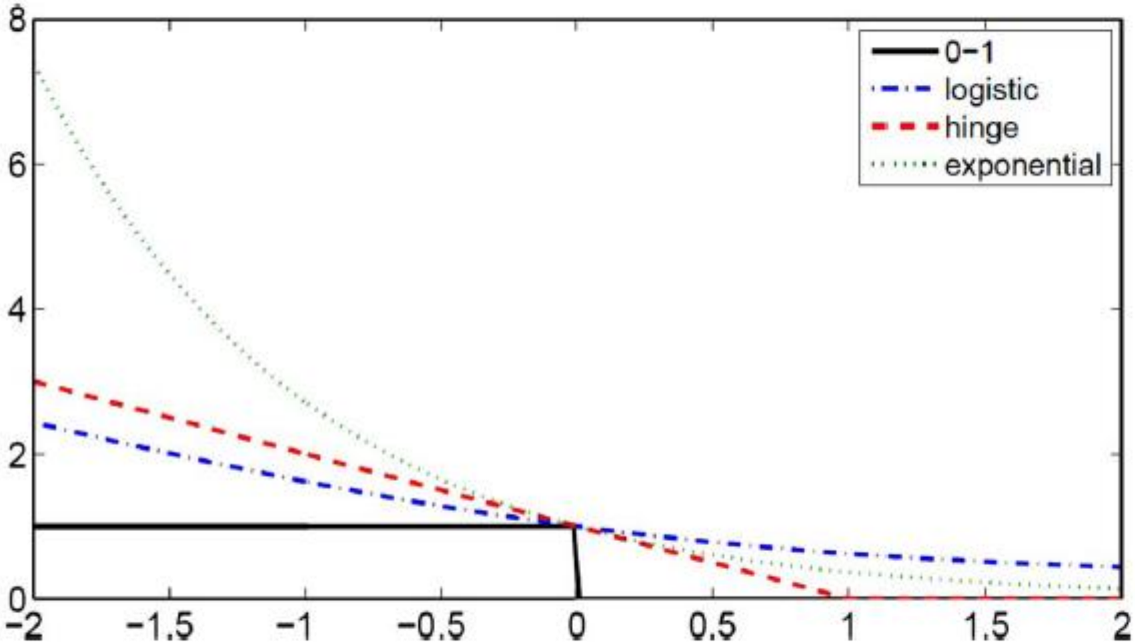


Fig. 3. Convex surrogates of 0–1 loss as discussed and analyzed in [6].

IV. Discriminative learning

- MCE and multi-class hinge

$$L\left(f(\mathbf{x}), y\right) = \sigma\left(-d_y(\mathbf{x};\lambda) + \ln\left[\frac{1}{|\mathcal{Y}|-1}\sum_{y' \neq y} \exp\{d_y(\mathbf{x};\lambda)\eta\}\right]^{\frac{1}{\eta}}\right) \quad (30)$$

$$L\left(f(\mathbf{x}), y\right) = \sum_{y' \neq y} \left|1 - d_y(\mathbf{x};\lambda) + d_{y'}(\mathbf{x};\lambda)\right|_+ \quad (31)$$

$$L\left(f(\mathbf{x}), \mathbf{y}\right) = \sum_{\mathbf{y}' \neq \mathbf{y}} \left|\Delta(\mathbf{y}, \mathbf{y}') - d_{\mathbf{y}}(\mathbf{x};\lambda) + d_{\mathbf{y}'}(\mathbf{x};\lambda)\right|_+ \quad (32)$$

IV. Discriminative learning

- IV.C discriminative learning in speech recognition
  - Models: MEMM, CRF, hidden CRF, MLP (generative models), decision boundary, SVM-HMM
  - Conditional likelihood

$$R_{\mathrm{emp}}(\lambda) = -\sum_i \ln \frac{p(\mathbf{x}^{(i)}, y^{(i)}; \lambda)}{p(\mathbf{x}^{(i)}; \lambda)} \qquad (33)$$

  - Bayesian minimum Risk
    - MCW, MPE, MWE

IV. Discriminative learning

- Large Margin

$$L\left(f(\mathbf{x}),\mathbf{y}\right)=\sum_{\mathbf{y}'\neq\mathbf{y}}\left|\Delta(\mathbf{y},\mathbf{y}')-\ln\frac{p(\mathbf{x},\mathbf{y};\lambda)}{p(\mathbf{x},\mathbf{y};\lambda)}\lambda)\right|_{+} \qquad (36)$$

$$R_{\mathrm{emp}}(f)=\min_{i}\left(d_{y}(\mathbf{x}_i;\lambda)-\max_{y'\neq y}d_{y'}(\mathbf{x}_i;\lambda)\right), \qquad (41)$$

IV. Discriminative learning

- IV.D Discriminative learning for HMM and related generative model
  - MMI, MCE, MWE, MPE
  - fMPE
- IV.E Hybrid generative-discriminative learning
  - Generative model for feature extraction, discriminative model for classification
  - Fisher kernel

V. semi-supervised learning

- V.C. semi-supervised learning
  - Inductive approaches

$$R_{\text{emp}}(f) + \alpha R_{\mathcal{U}}(f) + \gamma C(\lambda) \qquad (44)$$

$$R_{\mathcal{U}}(f) = -\sum_{i=m+1}^{m+n} \ln p(\mathbf{x}^{(i)}; \lambda) \qquad (45)$$

$$R_{\mathcal{U}}(f) = H(y|\mathbf{x}; \lambda) \qquad (46)$$

$$R_{\mathcal{U}}(f) = D(\hat{p} \| \tilde{p}_{\lambda}) \qquad (47)$$

V. semi-supervised learning

- V.C: transductive approaches

$$\min_F L(F, Y) + \gamma C(F, W) \quad (50)$$

V. semi-supervised learning

- V.D. semi-supervised learning in speech recognition
- V.E. Active learning
  - Uncertainty sampling
  - Query-by-committee
  - Exploiting structure in data
  - Submodular active selection: diminishing return

VI. Transfer Learning

- VI.A. Homogeneous transfer
  - 1) data combination
  - 2) model adaptation

$$J(f) = R^{T}_{\mathrm{emp}}(f) + \gamma C(f; f^{S}) \qquad (55)$$

VI. Transfer learning

- VI.B. homogeneous transfer in speech recognition
  - MAP, MLLR, SAT
- VI.C heterogeneous transfer
  - Map directly
  - Map to latent space
- VI.D multi-task learning

$$\min_{\theta,\mathbf{f}} \frac{1}{K} \sum_{k} R^{k}_{\mathrm{emp}}(f^{k}) + \gamma C(\mathbf{f};\theta) \quad (64)$$

VI. Transfer learning

- VI.E. heterogeneous and multi-task learning in ASR
  - Audio-visual recognition
  - Talking head
  - Articulatory learning
  - EEG
  - Cross lingual

VII. Emerging methods

- Deep learning
- Sparse representation
  - Sparse representation and signal recovery
  - Relevance vector machine and relevance detection

VIII. Conclusions

- A lost need to be learned from ML for ASR
- Care should be taken when learning from ML
- ASR and ML combination foster new ideas