NIPS Workshop: The Generative and Discriminative Learning Interface

Simon Lacoste-Julien

UNIVERSITY OF CAMBRIDGE

Percy Liang

Berkeley University of California

Guillaume Bouchard

xerox

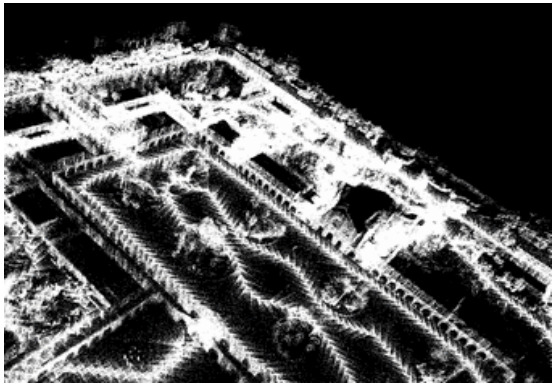
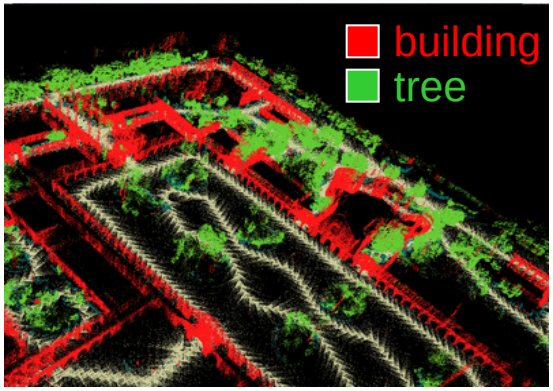
Whistler – Dec 12th 2009

07:30-07:50	**Opening address: themes of the workshop, terminology, open questions** Simon Lacoste-Julien, Percy Liang, Guillaume Bouchard
07:50-08:20	**Invited talk: Generative and Discriminative Models in Statistical Parsing** Michael Collins (MIT)
08:20-08:40	**Generative and Discriminative Latent Variable Grammars** Slav Petrov (Google Research)
08:40-09:00	**Discriminative and Generative Views of Binary Experiments** Mark D. Reid, Robert C. Williamson (Australian National University)
09:00-09:30	Coffee Break
09:30-10:00	**Invited talk: Multi-Task Discriminative Estimation for Generative Models and Probabilities** Tony Jebara (Columbia University)
10:00-	**Poster Session** (see below for abstracts)
	SKI / DISCUSSION BREAK
15:50-16:20	**Invited talk: Generative and Discriminative Image Models** John Winn (Microsoft Research Cambridge)
16:20-16:40	**Learning Feature Hierarchies by Learning Deep Generative Models** Ruslan Salakhutdinov (MIT)
16:40-17:00	**Why does Unsupervised Pre-training Help Deep Discriminant Learning?** Dumitru Erhan, Yoshua Bengio, Aaron Courville Pierre-Antoine Manzagol, Pascal Vincent (Université de Montréal)
17:00-17:30	Coffee Break
17:30-17:50	**Unsupervised Learning by Discriminating Data from Artificial Noise** Michael Gutmann, Aapo Hyvärinen (University of Helsinki)
17:50-18:45	**Panel Discussion** - Panelists:

Overview

- motivation
- terminology
- properties of gen. vs. disc.
- hybrids

Motivation: real-world predictions

	Input $\mathbf{x} \in \mathcal{X}$	(discrete) Output $\mathbf{y} \in \mathcal{Y}$
Machine translation	‘Ce n'est pas un autre problème de classification.’	‘This is not just another classification problem.’
3D object recognition		

Why do we care?

- enlarge toolbox of methods -> leverage advantages of both
- bridge different communities (Bayesian, frequentists, kernel people, neural networks, IEOR, ...)
- improve our understanding of properties of learning

Terminology

- for prediction:

Input	(discrete) Output	loss
$\mathbf{x} \in \mathcal{X}$	$\mathbf{y} \in \mathcal{Y}$	$\ell(\mathbf{y}', \mathbf{y})$

- decision theory goal:
  - given training data $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n\} \sim P$
  - learn decision function $\mathbf{y} = h(\mathbf{x})$

  with low **risk**: $\mathcal{R}(h) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P}\left[\,\ell(\mathbf{y}, h(\mathbf{x}))\,\right]$

Gen. vs. disc. learning

(more discriminative => more tuned towards **risk**)

generative learning

discriminative learning

joint learning

conditional learning

loss-sensitive learning

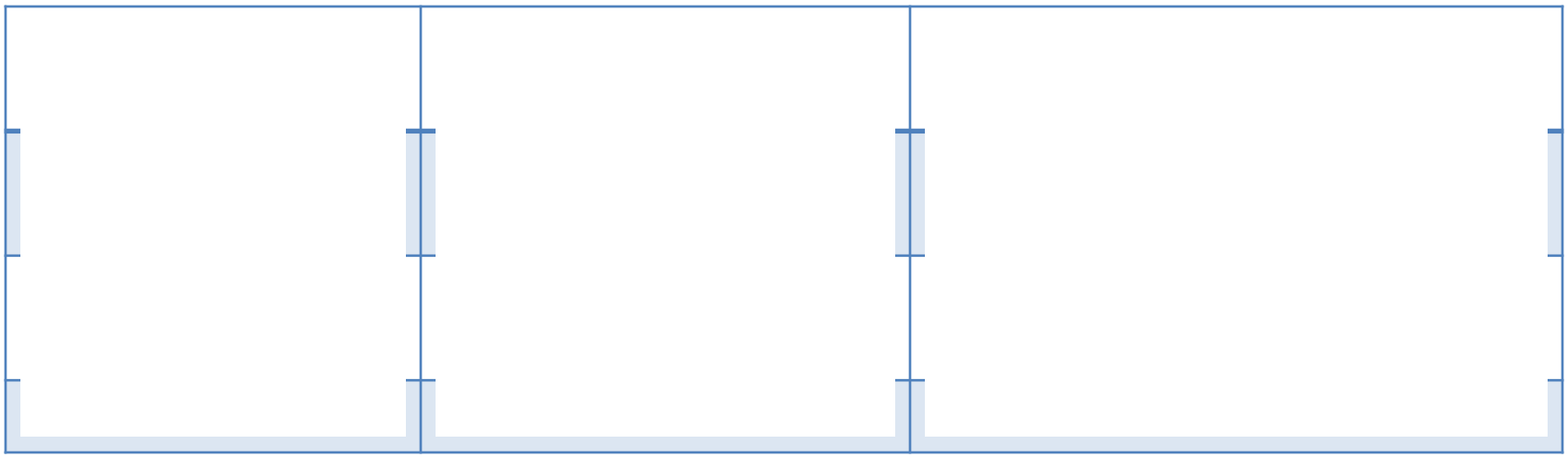
$$\hat{p}(\mathbf{x}, \mathbf{y})$$

$$\hat{p}(\mathbf{y} \mid \mathbf{x})$$

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\mathcal{L}}(\ell, \text{data}, h)$$

$$\hat{h}(\mathbf{x}) = \arg\min_{\mathbf{y} \in \mathcal{Y}} \sum_{\mathbf{y}' \in \mathcal{Y}} \hat{p}(\mathbf{y}' \mid \mathbf{x})\, \ell(\mathbf{y}', \mathbf{y})$$

surrogate empirical loss

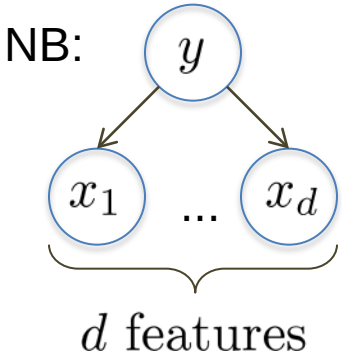


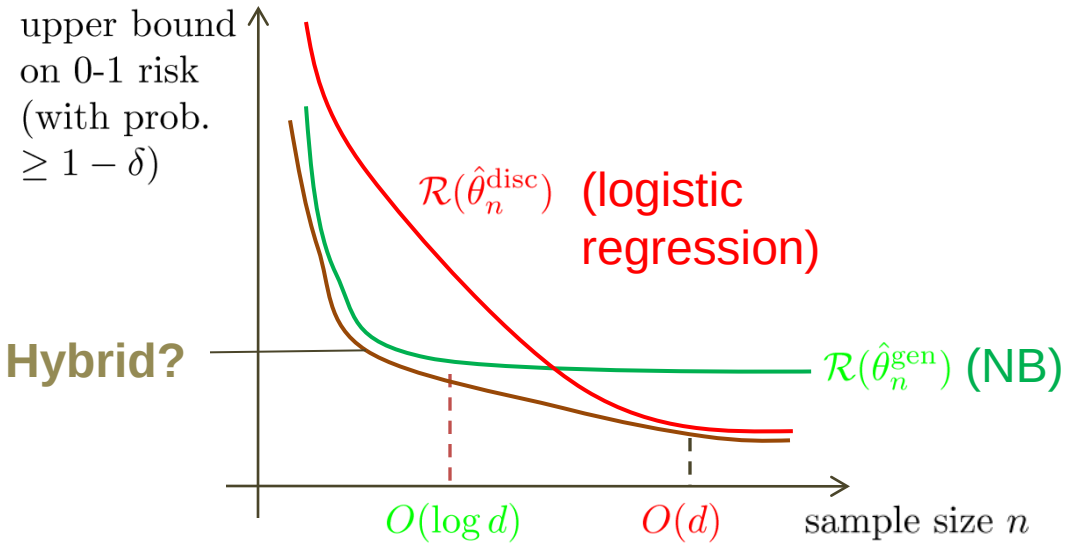
Gen. vs. Disc. Properties

Property	Generative	Conditional	Discriminative
Modularity	Probabilistic model: coherent, flexible and modular		Not calibrated for cascading
Distribution robustness	Depends on true p(x) & p(y|x)	Depends on true p(y|x)	More robust
Changing loss	Flexible prediction for different losses		Tailored to the loss
Unlabelled data	Simple	Difficult	
Computational Efficiency	Sometimes trivial (counts) Sometimes intractable (MRF)	No closed form Structured: intractable when loopy e.g.	No closed form Structured: more tractable (no normalization)

Motivation for interface

- [Ng & Jordan NIPS 02]: Naive Bayes vs. logistic regression for binary classification with linear classifiers





Some hybrid paradigms

Blending

Some hybrid paradigms

Blending

Interpolate [Bouchard & Triggs, 2004; McCallum et al., 2006; Liang et al., 2010]

$$\max_\theta \ \lambda \log p_\theta(x, y) + (1 - \lambda) \log p_\theta(y \mid x)$$

Some hybrid paradigms

Blending

Interpolate [Bouchard & Triggs, 2004; McCallum et al., 2006; Liang et al., 2010]

$$\max_\theta \ \lambda \log p_\theta(x, y) + (1 - \lambda) \log p_\theta(y \mid x)$$

Couple parameters [Lasserre et al., 2006; Agarwal & Daume, 2009]

$$\max_{\theta,\theta'} \ \log \sum_y p_\theta(x, y) + \log p_{\theta'}(y \mid x) + \log p(\theta, \theta')$$

Some hybrid paradigms

Blending

Interpolate [Bouchard & Triggs, 2004; McCallum et al., 2006; Liang et al., 2010]

$$\max_\theta \ \lambda \log p_\theta(x, y) + (1 - \lambda) \log p_\theta(y \mid x)$$

Couple parameters [Lasserre et al., 2006; Agarwal & Daume, 2009]

$$\max_{\theta,\theta'} \ \log \sum_y p_\theta(x, y) + \log p_{\theta'}(y \mid x) + \log p(\theta, \theta')$$

Model part of x [Liang & Jordan, 2008]

$$p_\theta(y, x_1, x_2) \qquad\qquad p_\theta(y \mid x_1, x_2)$$

Some hybrid paradigms

Blending

Interpolate [Bouchard & Triggs, 2004; McCallum et al., 2006; Liang et al., 2010]

$$\max_\theta \ \lambda \log p_\theta(x, y) + (1 - \lambda) \log p_\theta(y \mid x)$$

Couple parameters [Lasserre et al., 2006; Agarwal & Daume, 2009]

$$\max_{\theta,\theta'} \ \log \sum_y p_\theta(x, y) + \log p_{\theta'}(y \mid x) + \log p(\theta, \theta')$$

Model part of x [Liang & Jordan, 2008]

$$p_\theta(y, x_1, x_2) \quad p_\theta(y, x_1 \mid x_2) \quad p_\theta(y \mid x_1, x_2)$$

Some hybrid paradigms

Blending

Interpolate [Bouchard & Triggs, 2004; McCallum et al., 2006; Liang et al., 2010]

$$\max_\theta \ \lambda \log p_\theta(x, y) + (1 - \lambda) \log p_\theta(y \mid x)$$

Couple parameters [Lasserre et al., 2006; Agarwal & Daume, 2009]

$$\max_{\theta,\theta'} \ \log \sum_y p_\theta(x, y) + \log p_{\theta'}(y \mid x) + \log p(\theta, \theta')$$

Model part of x [Liang & Jordan, 2008]

$$p_\theta(y, x_1, x_2) \quad p_\theta(y, x_1 \mid x_2) \quad p_\theta(y \mid x_1, x_2)$$

Staged

Some hybrid paradigms

Blending

Interpolate [Bouchard & Triggs, 2004; McCallum et al., 2006; Liang et al., 2010]

$$\max_\theta \ \lambda \log p_\theta(x, y) + (1 - \lambda) \log p_\theta(y \mid x)$$

Couple parameters [Lasserre et al., 2006; Agarwal & Daume, 2009]

$$\max_{\theta,\theta'} \ \log \sum_y p_\theta(x, y) + \log p_{\theta'}(y \mid x) + \log p(\theta, \theta')$$

Model part of x [Liang & Jordan, 2008]

$$p_\theta(y, x_1, x_2) \quad p_\theta(y, x_1 \mid x_2) \quad p_\theta(y \mid x_1, x_2)$$

Staged

Use generative model as new features in discriminative model

- Very successful in NLP and vision

Some hybrid paradigms

Blending

Interpolate [Bouchard & Triggs, 2004; McCallum et al., 2006; Liang et al., 2010]

$$\max_\theta \ \lambda \log p_\theta(x, y) + (1 - \lambda) \log p_\theta(y \mid x)$$

Couple parameters [Lasserre et al., 2006; Agarwal & Daume, 2009]

$$\max_{\theta,\theta'} \ \log \sum_y p_\theta(x, y) + \log p_{\theta'}(y \mid x) + \log p(\theta, \theta')$$

Model part of x [Liang & Jordan, 2008]

$$p_\theta(y, x_1, x_2) \quad p_\theta(y, x_1 \mid x_2) \quad p_\theta(y \mid x_1, x_2)$$

Staged

Use generative model as new features in discriminative model

- Very successful in NLP and vision

Init. discrim. training with gen.-trained parameters [Hinton et al., 2006]

- Crucial for deep belief networks

Taxonomy

	Generative	**Discriminative**
Non-probabilistic		SVM Decision trees Neural nets Boosting
Probabilistic	**Joint** Naïve Bayes Hidden Markov Models Markov Random Fields Bayesian Networks	**Conditional** Logistic regression Conditional Random Fields

Taxonomy

	Generative	**Discriminative**
Non-probabilistic	K-Nearest Neighbors	SVM Decision trees Neural nets Boosting
Probabilistic	**Joint** Naïve Bayes Hidden Markov Models Markov Random Fields Bayesian Networks	**Conditional** Logistic regression Conditional Random Fields