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
MICHAELE 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-Aantoine
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

<table>
<thead>
<tr>
<th>Input</th>
<th>(discrete) Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x \in \mathcal{X}$</td>
<td>$y \in \mathcal{Y}$</td>
</tr>
</tbody>
</table>

- **Machine translation**
  - ‘Ce n'est pas un autre problème de classification.’
  - ‘This is not just another classification problem.’

- **3D object recognition**

![3D object recognition example](image)
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:

\[ \text{Input (discrete) Output loss} \]
\[ x \in \mathcal{X} \quad y \in \mathcal{Y} \quad \ell(y', y) \]

• decision theory goal:
  – given training data \( \mathcal{D} = \{(x_i, y_i)_{i=1}^n\} \sim P \)
  – learn decision function \( y = h(x) \)
  with low risk: \( \mathcal{R}(h) = \mathbb{E}_{(x,y) \sim P} [\ell(y, h(x))] \)
Gen. vs. disc. learning

(more discriminative => more tuned towards risk)

generative learning

joint learning
\[ \hat{p}(x, y) \]

\[ \hat{h}(x) = \arg \min_{y \in \mathcal{Y}} \sum_{y' \in \mathcal{Y}} \hat{p}(y' | x) \ell(y', y) \]

discriminative learning

conditional learning
\[ \hat{p}(y | x) \]

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

loss-sensitive learning

surrogate empirical loss
## Gen. vs. Disc. Properties

<table>
<thead>
<tr>
<th>Property</th>
<th>Generative</th>
<th>Conditional</th>
<th>Discriminative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Modularity</strong></td>
<td>Probabilistic model: coherent, flexible and modular</td>
<td>Not calibrated for cascading</td>
<td></td>
</tr>
<tr>
<td><strong>Distribution robustness</strong></td>
<td>Depends on true $p(x) &amp; p(y</td>
<td>x)$</td>
<td>Depends on true $p(y</td>
</tr>
<tr>
<td><strong>Changing loss</strong></td>
<td>Flexible prediction for different losses</td>
<td></td>
<td>Tailored to the loss</td>
</tr>
<tr>
<td><strong>Unlabelled data</strong></td>
<td>Simple</td>
<td>Difficult</td>
<td></td>
</tr>
<tr>
<td><strong>Computational Efficiency</strong></td>
<td>Sometimes trivial (counts) Sometimes intractable (MRF)</td>
<td>No closed form</td>
<td>No closed form</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Structured: intractable when loopy e.g.</td>
<td>Structured: more tractable (no normalization)</td>
</tr>
</tbody>
</table>
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]

\[
\max_{\theta} \lambda \log p_{\theta}(x, y) + (1 - \lambda) \log p_{\theta}(y | x)
\]

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

\[
\max_{\theta, \theta'} \log \sum_{y} p_{\theta}(x, y) + \log p_{\theta'}(y | x) + \log p(\theta, \theta')
\]
Some hybrid paradigms

Blending

Interpolate [Bouchard & Triggs, 2004; McCallum et al., 2006; Liang et al., 2010]
\[
\max_\theta \lambda \log p_\theta(x, y) + (1 - \lambda) \log p_\theta(y | x)
\]

Couple parameters [Lasserre et al., 2006; Agarwal & Daume, 2009]
\[
\max_{\theta, \theta'} \log \sum_y p_\theta(x, y) + \log p_{\theta'}(y | x) + \log p(\theta, \theta')
\]

Model part of $x$ [Liang & Jordan, 2008]
\[
p_\theta(y, x_1, x_2) \quad p_\theta(y | x_1, x_2)
\]
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]

$$\max_{\theta} \lambda \log p_{\theta}(x, y) + (1 - \lambda) \log p_{\theta}(y | x)$$

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

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

Model part of $x$ [Liang & Jordan, 2008]

$$p_{\theta}(y, x_1, x_2) \quad p_{\theta}(y, x_1 | x_2) \quad p_{\theta}(y | x_1, x_2)$$

**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]

\[
\max_{\theta} \lambda \log p_\theta(x, y) + (1 - \lambda) \log p_\theta(y | x)
\]

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

\[
\max_{\theta, \theta'} \log \sum_y p_\theta(x, y) + \log p_{\theta'}(y | x) + \log p(\theta, \theta')
\]

Model part of \(x\) [Liang & Jordan, 2008]

\[
 p_\theta(y, x_1, x_2) \quad p_\theta(y, x_1 | x_2) \quad p_\theta(y | x_1, x_2)
\]

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

<table>
<thead>
<tr>
<th>Generative</th>
<th>Discriminative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Non-probabilistic</strong></td>
<td><strong>SVM</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Decision trees</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Neural nets</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Boosting</strong></td>
</tr>
<tr>
<td><strong>Probabilistic</strong></td>
<td><strong>Conditional</strong></td>
</tr>
<tr>
<td><strong>Joint</strong></td>
<td><strong>Logistic regression</strong></td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td><strong>Conditional Random Fields</strong></td>
</tr>
<tr>
<td>Hidden Markov Models</td>
<td></td>
</tr>
<tr>
<td>Markov Random Fields</td>
<td></td>
</tr>
<tr>
<td>Bayesian Networks</td>
<td></td>
</tr>
</tbody>
</table>
# Taxonomy

<table>
<thead>
<tr>
<th>Generative</th>
<th>Discriminative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-probabilistic</td>
<td></td>
</tr>
<tr>
<td>K-Nearest Neighbors</td>
<td>SVM</td>
</tr>
<tr>
<td></td>
<td>Decision trees</td>
</tr>
<tr>
<td></td>
<td>Neural nets</td>
</tr>
<tr>
<td></td>
<td>Boosting</td>
</tr>
<tr>
<td>Probabilistic</td>
<td>Conditional</td>
</tr>
<tr>
<td>Joint</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>Conditional Random Fields</td>
</tr>
<tr>
<td>Hidden Markov Models</td>
<td></td>
</tr>
<tr>
<td>Markov Random Fields</td>
<td></td>
</tr>
<tr>
<td>Bayesian Networks</td>
<td></td>
</tr>
</tbody>
</table>