Constraints Aware Learning and Inference

Daniel Khashabi
Based on

• Part 1: Constraint Driven Learning (Chang et al, 2008)
• Part 2: Measurements in Exponential Family (Liang, 2009)
Part 1

Constraints Driven Learning
Constrained Conditional Models (aka ILP Inference)

How to solve?

This is an Integer Linear Program

Solving using ILP packages gives an exact solution.

Cutting Planes, Dual Decomposition & other search techniques are possible

Penalty for violating the constraint.

(Soft) constraints component

How far \( y \) is from a “legal” assignment

How to train?

Training is learning the objective function

Decouple? Decompose?

How to exploit the structure to minimize supervision?
Information extraction without **Prior Knowledge**


\[
\operatorname*{argmax}_y \lambda \cdot F(x, y)
\]

**Prediction result of a trained HMM**

- **AUTHOR**: Lars Ole Andersen
- **TITLE**: Program analysis and specialization for the C Programming language
- **TECH-REPORT**: PhD thesis
- **INSTITUTION**: DIKU, University of Copenhagen
- **DATE**: May 1994

Violates lots of **natural** constraints!
Examples of Constraints

• Each field must be a consecutive list of words and can appear at most once in a citation.

• State transitions must occur on punctuation marks.

• The citation can only start with AUTHOR or EDITOR.

• The words pp., pages correspond to PAGE.
• Four digits starting with 20xx and 19xx are DATE.
• Quotations can appear only in TITLE

Easy to express pieces of “knowledge”

Non Propositional; May use Quantifiers
Constraints Driven Learning (CoDL)

Several Training Paradigms

For N iterations do

\( T = \phi \)

For each \( x \) in unlabeled dataset

\[ h \leftarrow \arg\max_y w^T \phi(x, y) - \sum \rho_k d_c(x, y) \]

\( T = T \cup \{(x, h)\} \)

\( (w, \rho) = \gamma (w_0, \rho_0) + (1 - \gamma) \text{learn}(T) \)

[Chang, Ratinov, Roth, ACL’07;ICML’08,ML, to appear]
Generalized by Ganchev et. al [PR work]

Supervised learning algorithm parameterized by \((w, \rho)\). Learning can be justified as an optimization procedure for an objective function.

Inference with constraints: augment the training set

Learn from new training data
Weigh supervised & unsupervised models.

Excellent Experimental Results showing the advantages of using constraints, especially with small amounts on labeled data [Chang et. al, Others]
Constraints Driven Learning (CODL)

Objective function:
\[
f_{\Phi,C}(x, y) = \sum w_i \phi_i(x, y) - \sum \rho_i d_{C_i}(x, y).
\]

- Semi-Supervised Learning Paradigm that makes use of constraints to bootstrap from a small number of examples.

Constraints are used to:
- **Bootstrap** a semi-supervised learner
- **Correct** weak models predictions on unlabeled data, which in turn are used to keep training the model.

[Chang, Ratinov, Roth, ACL’07; ICML’08, MLJ, to appear]
Generalized by Ganchev et. al [PR work]
Part 2

Learning from Measurements in Exponential Families
The big picture

target predictor $p^*$ → human information → learning algorithm → learned predictor $\hat{p}$

Example:

$y$: FEAT FEAT FEAT FEAT FEAT FEAT ...

$x$: View of Los Gatos Foothills ...

AVAIL AVAIL AVAIL ... SIZE SIZE SIZE SIZE ...

Available July 1 ... 2 bedroom 1 bath ...
The big picture

target predictor $p^*$ → human → information → learning algorithm → learned predictor $\hat{p}$

Example:

$y$: FEAT FEAT FEAT FEAT FEAT FEAT ...

$x$: View of Los Gatos Foothills ...

Available July 1 ... 2 bedroom 1 bath ...

Types of information:

Labeled examples (specific) [standard supervised learning]
The big picture

- **Target predictor** $p^*$
- **Human**
- **Information**
- **Learning algorithm**
- **Learned predictor** $\hat{p}$

Example:

$y$: FEAT FEAT FEAT FEAT FEAT FEAT ...

$x$: View of Los Gatos Foothills ...

Available July 1 ... 2 bedroom 1 bath ...

Types of information:

- Labeled examples (specific) [standard supervised learning]
- Constraints (general) [Chang, et al., 2007; Druck, et al., 2008]
The big picture

target predictor $p^*$ → human → information → learning algorithm → learned predictor $\hat{p}$

Example:

$y$: FEAT FEAT FEAT FEAT FEAT FEAT ...

$x$: View of Los Gatos Foothills ...

Available July 1 ... 2 bedroom 1 bath ...

Types of information:

Labeled examples (specific) [standard supervised learning]
Constraints (general) [Chang, et al., 2007; Druck, et al., 2008]
Measurements: our unifying framework
The big picture

target predictor $p^*$ → human → information → learning algorithm → learned predictor $\hat{p}$

Example:

$y$: FEAT FEAT FEAT FEAT FEAT FEAT ...

$x$: View of Los Gatos Foothills ...

Available July 1 ... 2 bedroom 1 bath ...

Types of information:

- Labeled examples (specific) [standard supervised learning]
- Constraints (general) [Chang, et al., 2007; Druck, et al., 2008]

**Measurements**: our unifying framework

Outline:

1. Coherently learn from diverse measurements
The big picture

target predictor $p^*$ → human → information → learning algorithm → learned predictor $\hat{p}$

Example:

$y$: FEAT FEAT FEAT FEAT FEAT FEAT ...
$x$: View of Los Gatos Foothills ...

Available July 1 ... 2 bedroom 1 bath ...

Types of information:

- Labeled examples (specific) [standard supervised learning]
- Constraints (general) [Chang, et al., 2007; Druck, et al., 2008]

Measurements: our unifying framework

Outline:

1. Coherently learn from diverse measurements
2. Actively select the best measurements
Measurements

\[ X_1, Y_1 \]
\[ X_2, Y_2 \]
\[ X_3, Y_3 \]
\[ \ldots, \ldots \]
\[ X_i, Y_i \]
\[ \ldots, \ldots \]
\[ X_n, Y_n \]
Measurements

Measurement features: $\sigma(x, y) \in \mathbb{R}^k$

$\sigma( X_1, Y_1 )$
$\sigma( X_2, Y_2 )$
$\sigma( X_3, Y_3 )$

\ldots \quad \ldots

$\sigma( X_i, Y_i )$

\ldots \quad \ldots

$\sigma( X_n, Y_n )$
Measurements

Measurement features: \( \sigma(x, y) \in \mathbb{R}^k \)

Measurement values: \( \tau \in \mathbb{R}^k \)

\[
\tau = \sum_{i=1}^{n} \sigma(X_i, Y_i) + \text{noise}
\]
Measurements

Measurement features: $\sigma(x, y) \in \mathbb{R}^k$

Measurement values: $\tau \in \mathbb{R}^k$

$$\tau = \sum_{i=1}^{n} \sigma(X_i, Y_i) + \text{noise}$$
Measurements

Measurement features: \( \sigma(x, y) \in \mathbb{R}^k \)

Measurement values: \( \tau \in \mathbb{R}^k \)

\[
\tau = \sum_{i=1}^{n} \sigma(X_i, Y_i) + \text{noise}
\]

Set \( \sigma \) to reveal various types of information about \( Y \) through \( \tau \)
Examples of measurements

Fully-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \text{View of Los} ..., y = * * * ...]$$
Examples of measurements

Fully-labeled example:

\[ \sigma_j(x, y) = \mathbb{I}[x = \text{View of Los} \ldots, y = \ast \ast \ast \ldots] \]

Partially-labeled example:

\[ \sigma_j(x, y) = \mathbb{I}[x = \text{View of Los} \ldots, y_1 = \ast] \]
Examples of measurements

Fully-labeled example:
\[ \sigma_j(x, y) = \mathbb{I}[x = \text{View of Los} \ldots, y = \ast \ast \ast \ldots] \]

Partially-labeled example:
\[ \sigma_j(x, y) = \mathbb{I}[x = \text{View of Los} \ldots, y_1 = \ast] \]

Labeled predicate:
\[ \sigma_j(x, y) = \sum_i \mathbb{I}[x_i = \text{View}, y_i = \ast] \]
Examples of measurements

Fully-labeled example:
\[ \sigma_j(x, y) = \mathbb{I}[x = \text{View of Los } ..., y = * * * ...] \]

Partially-labeled example:
\[ \sigma_j(x, y) = \mathbb{I}[x = \text{View of Los } ..., y_1 = *] \]

Labeled predicate:
\[ \sigma_j(x, y) = \sum_i \mathbb{I}[x_i = \text{View}, y_i = *] \]

Label proportions:
\[ \sigma_j(x, y) = \sum_i \mathbb{I}[y_i = *] \]
Examples of measurements

Fully-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \text{View of Los} \ldots, y = \ast \ast \ast \ldots]$$

Partially-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \text{View of Los} \ldots, y_1 = \ast]$$

Labeled predicate:

$$\sigma_j(x, y) = \sum_i \mathbb{I}[x_i = \text{View}, y_i = \ast]$$

Label proportions:

$$\sigma_j(x, y) = \sum_i \mathbb{I}[y_i = \ast]$$

Label preference:

$$\sigma_j(x, y) = \sum_i \mathbb{I}[y_i = \text{FEAT}] - \mathbb{I}[y_i = \text{AVAIL}]$$
Examples of measurements

Fully-labeled example:
\[ \pi_j(x, y) = \mathbb{I}[x = \text{View of Los} \ldots, y = * * * \ldots] \]

Partially-labeled example:
\[ \pi_j(x, y) = \mathbb{I}[x = \text{View of Los} \ldots, y_1 = *] \]

Labeled predicate:
\[ \pi_j(x, y) = \sum_i \mathbb{I}[x_i = \text{View}, y_i = *] \]

Label proportions:
\[ \pi_j(x, y) = \sum_i \mathbb{I}[y_i = *] \]

Label preference:
\[ \pi_j(x, y) = \sum_i \mathbb{I}[y_i = \text{FEAT}] - \mathbb{I}[y_i = \text{AVAIL}] \]

Can get measurement values \( \tau \) without looking at all examples

Next: How to combine these diverse measurements coherently?
Prediction model

Bayesian framework:
Prediction model

Bayesian framework:

Exponential families:

\[ p_\theta(y \mid x) = \exp\{ \langle \phi(x, y), \theta \rangle - A(\theta; x) \} \]
Prediction model

Bayesian framework:

Exponential families:

\[ p_\theta(y \mid x) = \exp\{ \langle \phi(x, y), \theta \rangle - A(\theta; x) \} \]

\( \phi(x, y) \in \mathbb{R}^d \): model features
\( \theta \in \mathbb{R}^d \): model parameters
\( A(\theta; x) = \log \int \exp\{ \langle \phi(x, y), \theta \rangle \} dy \): log-partition function
Learning via Bayesian inference

Goal: compute $p(\theta, Y \mid \tau, X)$

Variational formulation:

$$\min_{q \in Q_{\theta, Y}} \text{KL} (q(\theta, Y) \parallel p(\theta, Y \mid \tau, X))$$

Approximations:

- $Q_{\theta, Y}$: mean-field factorization of $q(Y)$ and degenerate $\bar{\theta}$
- $\text{KL}$: measurements only hold in expectation (w.r.t. $q(Y)$)
Learning via Bayesian inference

Goal: compute $p(\theta, Y \mid \tau, X)$

Variational formulation:

$$\min_{q \in Q_{\theta,Y}} \text{KL} (q(\theta, Y) \parallel p(\theta, Y \mid \tau, X))$$

Approximations:

- $Q_{\theta,Y}$: mean-field factorization of $q(Y)$ and degenerate $\tilde{\theta}$
- KL: measurements only hold in expectation (w.r.t. $q(Y)$)

Algorithm:

Apply Fenchel duality $\rightarrow$ saddlepoint problem
Take alternating stochastic gradient steps
Information geometry viewpoint

(assume zero measurement noise)

\[ Q \overset{\text{def}}{=} \{ q(y \mid x) : \mathbb{E}_q[\sigma] = \tau \} \]
\[ \mathcal{P} \overset{\text{def}}{=} \{ p_\theta(y \mid x) : \theta \in \mathbb{R}^d \} \]

\[
\min_{q \in Q, p \in \mathcal{P}} \text{KL} (q \parallel p)
\]
Information geometry viewpoint

(assume zero measurement noise)

\[ Q \overset{\text{def}}{=} \{ q(y \mid x) : \mathbb{E}_q[\sigma] = \tau \} \]
\[ P \overset{\text{def}}{=} \{ p_\theta(y \mid x) : \theta \in \mathbb{R}^d \} \]

\[ \min_{q \in Q, p \in P} \text{KL} \left( q \mid\mid p \right) \]

Interpretation:
- Measurements shape \( Q \)
- Find model in \( P \) with best fit
Model features $\phi$ versus measurement features $\sigma$

Guidelines:

To set $\sigma$, consider human (e.g., full labels)
To set $\phi$, consider statistical generalization (e.g., word suffixes)
Model features $\phi$ versus measurement features $\sigma$

Guidelines:

To set $\sigma$, consider human (e.g., full labels)

To set $\phi$, consider statistical generalization (e.g., word suffixes)

Intuition: consider feature $f(x, y) = \mathbb{I}[x \in A, y = 1]$

If $f$ is a measurement feature (direct):

“inputs in $A$ should be labeled according to $\tau$”

If $f$ is a model feature (indirect):

“inputs in $A$ should be labeled similarly”
Results on the Craigslist task

$n = 1000$ total examples (ads), 11 possible labels

Model:

Conditional random field with standard NLP features

Measurements:

- fully-labeled examples
- 33 labeled predicates (e.g., $\sum_i \mathbb{I}[x_i = \text{View}, y_i = \text{FEAT}]$)
Results on the Craigslist task

$n = 1000$ total examples (ads), 11 possible labels

Model:

  Conditional random field with standard NLP features

Measurements:

  • fully-labeled examples
  • 33 labeled predicates (e.g., $\sum_i \mathbb{I}[x_i = \text{View}, y_i = \text{FEAT}]$)

Per-position test accuracy (on 100 examples):

<table>
<thead>
<tr>
<th># labeled examples</th>
<th>10</th>
<th>25</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Expectation Criteria</td>
<td>74.6</td>
<td>77.2</td>
<td>80.5</td>
</tr>
<tr>
<td>Constraint-Driven Learning</td>
<td><strong>74.7</strong></td>
<td><strong>78.5</strong></td>
<td>81.7</td>
</tr>
<tr>
<td>Measurements</td>
<td>71.4</td>
<td>76.5</td>
<td><strong>82.5</strong></td>
</tr>
</tbody>
</table>

Able to integrate labeled examples and predicates gracefully
So far: given measurements, how to learn

Next: how to choose measurements?
Experimental design (active learning)

Utility of measurement $(\sigma, \tau)$:

$$U(\sigma, \tau) = \underbrace{R(\sigma, \tau)}_{\text{reward}} - \underbrace{C(\sigma)}_{\text{cost}}$$
Utility of measurement \((\sigma, \tau)\):

\[
U(\sigma, \tau) = R(\sigma, \tau) - C(\sigma)
\]

When considering \(\sigma\), don’t know \(\tau\), so integrate out:

\[
U(\sigma) = \mathbb{E}_{p(\tau|X)}[U(\sigma, \tau)]
\]
Utility of measurement $(\sigma, \tau)$:

$$U(\sigma, \tau) = \underbrace{R(\sigma, \tau)}_{\text{reward}} - \underbrace{C(\sigma)}_{\text{cost}}$$

When considering $\sigma$, don’t know $\tau$, so integrate out:

$$U(\sigma) = E_{p(\tau|X)}[U(\sigma, \tau)]$$
Experimental design (active learning)

Utility of measurement \((\sigma, \tau)\):

\[
U(\sigma, \tau) = R(\sigma, \tau) - C(\sigma)
\]

reward \quad cost

When considering \(\sigma\), don't know \(\tau\), so integrate out:

\[
U(\sigma) = \mathbb{E}_{p(\tau|X)}[U(\sigma, \tau)]
\]

Choose best measurement feature \(\sigma\):

\[
\sigma^* = \arg \max_{\sigma} U(\sigma)
\]
Part-of-speech tagging results

\( n = 1000 \) total examples (sentences), 45 possible labels

Model: Indep. logistic regression with standard NLP features
Part-of-speech tagging results

$n = 1000$ total examples (sentences), 45 possible labels

Model: Indep. logistic regression with standard NLP features

Measurements:
- fully-labeled examples
- labeled predicates (e.g., $\sum_i \mathbb{I}[x_i = \text{the}, y_i = \text{DT}]$)

Use label entropy as surrogate for assessing measurements
Part-of-speech tagging results

\( n = 1000 \) total examples (sentences), 45 possible labels

Model: Indep. logistic regression with standard NLP features

Measurements:
- fully-labeled examples
- labeled predicates (e.g., \( \sum_i \mathbb{I}[x_i = \text{the}, y_i = \text{DT}] \))

Use label entropy as surrogate for assessing measurements

Test accuracy (on 100 examples):

(a) Labeling examples
Part-of-speech tagging results

$n = 1000$ total examples (sentences), 45 possible labels

Model: Indep. logistic regression with standard NLP features

Measurements:
- fully-labeled examples
- labeled predicates \( (e.g., \sum_i \mathbb{I}[x_i = \text{the}, y_i = \text{DT}]) \)

Use label entropy as surrogate for assessing measurements

Test accuracy (on 100 examples):

(a) Labeling examples

(b) Labeling word types
Summary

target predictor $p^*$  human  →  measurements  →  learning algorithm  →  learned predictor $\hat{p}$
Summary

target predictor $p^*$ → human → measurements → learning algorithm → learned predictor $\hat{p}$

Measurements

Bayesian model
Summary

target predictor $p^*$ -> human -> measurements -> learning algorithm -> learned predictor $\hat{p}$

Measurements

variational approx. —— Bayesian model
Summary

target predictor $p^*$ → human → measurements → learning algorithm → learned predictor $\hat{p}$

Measurements

variational approx. —— Bayesian model

information geometry
Summary

target predictor $p^*$ → human → measurements → learning algorithm → learned predictor $\hat{p}$

Measurements

variational approx. --- Bayesian model --- decision theory

information geometry
Summary

target predictor $p^*$ → human → measurements → learning algorithm → learned predictor $\hat{p}$

Measurements

variational approx. — Bayesian model — decision theory

information geometry

active learning
Approximate Inference

\[ \min_{q \in \mathcal{Q}} \text{KL} (q(Y, \theta) \| p(Y, \theta \mid \tau, X, \sigma)) . \]

\[ \mathcal{Q} \overset{\text{def}}{=} \{ q(Y, \theta) : q(Y, \theta) = q(Y) \delta_{\theta}(\theta) \} . \]

The probabilistic model

\[ p(\theta, Y, \tau \mid X, \sigma) \overset{\text{def}}{=} p(\theta) \prod_{i=1}^{n} p_{\theta}(Y_i \mid X_i) p(\tau \mid X, Y, \sigma) . \]

Log-concave prior:

\[ \log p(\theta) = -h_\phi(\theta) + \text{constant} \]
\[ \log p(\tau \mid X, Y, \sigma) = -h_\sigma(\tau - \sigma^X(Y)) + \text{constant} \]
Approximate Inference

• The probabilistic model

\[ p(\theta, Y, \tau \mid X, \sigma) \overset{\text{def}}{=} p(\theta) \prod_{i=1}^{n} p_{\theta}(Y_i \mid X_i)p(\tau \mid X, Y, \sigma). \]

Log-concave prior:

\[
\log p(\theta) = -h_\phi(\theta) + \text{constant} \\
\log p(\tau \mid X, Y, \sigma) = -h_\sigma(\tau - \sigma^X(Y)) + \text{constant}
\]

For example:

• Gaussian:

\[ h_\phi(\theta) = \frac{\lambda}{2} ||\theta||^2 \]

• Box:

\[ h_\sigma(u) = \mathbb{W}[\forall j, |u_j| \leq \epsilon_j] \]
Approximate Inference

\[
\min_{q \in \mathcal{Q}} \text{KL} (q(Y, \theta) \| p(Y, \theta \mid \tau, X, \sigma)).
\]

\[
\min_{q(Y), \theta} -H(q(Y)) + E_{q(Y)}[h_\sigma(\tau - \sigma^X(Y))] \\
- \sum_{i=1}^{n} E_{q(Y)} \log p_\theta(Y_i \mid X_i) + h_\phi(\theta).
\]

\[
q(Y) = \prod_{i=1}^{n} q_{\beta, \theta}(Y_i \mid X_i)
\]

\[
q_{\beta, \theta}(y \mid x) = \exp\{\langle \sigma(x, y), \beta \rangle + \langle \phi(x, y), \theta \rangle - B(\beta, \theta; x)\},
\]
Fenchel’s Duality Theorem

• Let $f$ be convex and $g$ be concave function

$$\min_{x} (f(x) - g(x)) = \max_{p} (g_{*}(p) - f^{*}(p)).$$

Then:

$$f^{*}(x^{*}) := \sup \{ \langle x^{*}, x \rangle - f(x) | x \in \mathbb{R}^n \}$$

$$g_{*}(x^{*}) := \inf \{ \langle x^{*}, x \rangle - g(x) | x \in \mathbb{R}^n \}$$
Finding a convex lower bound on posterior

\[
\min_{\theta \in \mathbb{R}^d} \max_{\beta \in \mathbb{R}^k} L(\beta, \theta),
\]

\[
L(\beta, \theta) = \langle \tau, \beta \rangle - \sum_{i=1}^{n} B(\beta, \theta; X_i) + \sum_{i=1}^{n} A(\theta; X_i) - h_{\sigma}^*(\beta) + h_{\phi}(\theta),
\]

\[
h_{\sigma}^*(\beta) = \sup_{u \in \mathbb{R}^k} \{ \langle u, \beta \rangle - h_{\sigma}(u) \}
\]

\[
\frac{\partial L(\beta, \theta)}{\partial \beta} \quad \frac{\partial L(\beta, \theta)}{\partial \theta}
\]
References

• Slides: Dan Roth:
  http://l2r.cs.uiuc.edu/~danr/Talks/IndirectSup-MSR-06011.ppt

• Slides: Percy Liang: