The big picture

target
predictor $p^*$

human
The big picture

Example:

\[ y: \text{FEAT FEAT FEAT FEAT FEAT } \ldots \]
\[ x: \text{View of Los Gatos Foothills } \ldots \]
\[ \text{AVAIL AVAIL AVAIL } \ldots \text{SIZE SIZE SIZE SIZE } \ldots \]
\[ \text{Available July 1 } \ldots \text{2 bedroom 1 bath } \ldots \]
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 $p^*$ $\xrightarrow{}$ human information $\xrightarrow{}$ learning algorithm $\xrightarrow{}$ 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

Example:

*y*: 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

Example:

\[ y: \text{FEAT FEAT FEAT FEAT FEAT FEAT...} \]
\[ x: \text{View of Los Gatos Foothills...} \]
\[ \text{AVAIL AVAIL AVAIL... SIZE SIZE SIZE SIZE...} \]
\[ \text{Available July 1... 2 bedroom 1 bath...} \]

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

Example:

\( y: \text{FEAT FEAT FEAT FEAT FEAT FEAT ...} \)

\( x: \text{View of Los Gatos Foothills ...} \)

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

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

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

$\sigma( X_1, Y_1 )$
$\sigma( X_2, Y_2 )$
$\sigma( X_3, Y_3 )$
...
$\sigma( X_i, Y_i )$
...
$\sigma( X_n, Y_n )$
Measurements

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

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

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

\begin{align*}
\sigma(X_1, Y_1) \\
\sigma(X_2, Y_2) \\
\sigma(X_3, Y_3) \\
\vdots \quad \vdots \\
\sigma(X_i, Y_i) \\
\vdots \quad \vdots \\
\sigma(X_n, Y_n) \\
+ \text{noise}
\end{align*}

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

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

Fully-labeled example:
\[ \sigma_j(x, y) = \mathbb{1}[x = \text{View of Los} ..., y = * * * ...] \]

Partially-labeled example:
\[ \sigma_j(x, y) = \mathbb{1}[x = \text{View of Los} ..., y_1 = *] \]

Labeled predicate:
\[ \sigma_j(x, y) = \sum_i \mathbb{1}[x_i = \text{View}, y_i = *] \]
Examples of measurements

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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:
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:

\[ \theta \]

\[ X_i \]

\[ Y_i \]

\[ n \]

\[ \tau \]

Exponential families:

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

\[ \phi(x, y) \in \mathbb{R}^d: \text{model features} \]
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
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 $p(\theta, Y \mid \tau, X)$
Learning via Bayesian inference

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

Variational formulation:

$$\min_{q \in \mathcal{Q}_{\theta, Y}} \text{KL} (q(\theta, Y) \| p(\theta, Y \mid \tau, X))$$
Learning via Bayesian inference

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

Variational formulation:

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

Approximations:

- $Q_{\theta, Y}$: mean-field factorization of $q(Y)$ and degenerate $\tilde{\theta}$
Learning via Bayesian inference

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

Variational formulation:

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

Approximations:

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

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

Variational formulation:
\[
\min_{q \in Q_{\theta,Y}} \text{KL} \left( q(\theta, Y) \mid\mid p(\theta, Y \mid \tau, X) \right)
\]

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 → saddlepoint problem
Take alternating stochastic gradient steps
Information geometry viewpoint

(assume zero measurement noise)

\[ \mathcal{P} \overset{\text{def}}{=} \{ p_\theta(y \mid x) : \theta \in \mathbb{R}^d \} \]
Information geometry viewpoint

(assume zero measurement noise)

\[ \mathcal{Q} \overset{\text{def}}{=} \{ q(y \mid x) : \mathbb{E}_q[\sigma] = \tau \} \]

\[ \mathcal{P} \overset{\text{def}}{=} \{ p_\theta(y \mid x) : \theta \in \mathbb{R}^d \} \]
Information geometry viewpoint

(assume zero measurement noise)

\( \mathcal{Q} \overset{\text{def}}{=} \{ q(y \mid x) : \mathbb{E}_q[\sigma] = \tau \} \)

\( \mathcal{P} \overset{\text{def}}{=} \{ p_\theta(y \mid x) : \theta \in \mathbb{R}^d \} \)

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

(assume zero measurement noise)

\[ Q \overset{\text{def}}{=} \{ q(y \mid x) : \mathbb{E}_q[\sigma] = \tau \} \]

\[ P \overset{\text{def}}{=} \{ p_{\theta}(y \mid x) : \theta \in \mathbb{R}^d \} \]

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

Interpretation:

Measurements shape \( Q \)  
Find model in \( P \) with best fit
Information geometry viewpoint

(assume zero measurement noise)

\[ Q \overset{\text{def}}{=} \{ q(y \mid x) : \mathbb{E}_q[\sigma] = \tau \} \]

\[ P \overset{\text{def}}{=} \{ p_{\theta}(y \mid x) : \theta \in \mathbb{R}^d \} \]

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

Interpretation:
Measurements shape \( Q \) \quad Find model in \( P \) with best fit

Two ways to recover supervised learning:
1. Measure \( \sigma = \phi \): \( P \cap Q \) is the unique solution
Information geometry viewpoint

(assume zero measurement noise)

\[ Q \overset{\text{def}}{=} \{ q(y \mid x) : \mathbb{E}_q[\sigma] = \tau \} \]

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

\[
\min_{q \in Q, p \in \mathcal{P}} \text{KL} (q \| p)
\]

Interpretation:

Measurements shape \( Q \) \hspace{1cm} \text{Find model in} \ \mathcal{P} \ \text{with best fit}

Two ways to recover supervised learning:

1. Measure \( \sigma = \phi \): \( \mathcal{P} \cap Q \) is the unique solution
2. Measure \( \sigma = \{ \mathbb{I}[x = a, y = b] \} \):
   \( Q = \{ \text{empirical distribution} \} \), project onto \( \mathcal{P} \)
Model features $\phi$ versus measurement features $\sigma$
Model features $\phi$ versus measurement features $\sigma$

Guidelines:

To set $\sigma$, consider human (e.g., full labels)
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 $f(x, y) = \mathbb{I}[x \in A, y = 1]$
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 $\tau$”
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
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>
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., \( \sum_i I[x_i = \text{View}, y_i = \text{FEAT}] \))

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>74.7</td>
<td>78.5</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?
Bayesian decision theory

What do we do with an (approximate) posterior \( p(Y, \theta \mid X, \tau) \)?
What do we do with an (approximate) posterior $p(Y, \theta \mid X, \tau)$?

Bayes-optimal predictor:

average over $X'$, max over $\hat{Y}'$, average over $Y'$ of reward
Bayesian decision theory

What do we do with an (approximate) posterior $p(Y, \theta | X, \tau)$?

Bayes-optimal predictor:

average over $X'$, max over $\hat{Y}'$, average over $Y'$ of reward

$R(\sigma, \tau) =$ expected reward of Bayes-optimal predictor
(i.e., how happy we are with the given situation)
Utility of measurement \((\sigma, \tau)\):

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

\(U\) and \(\theta\) are inputs, and \(X_i\), \(Y_i\), and \(\tau\) are intermediate variables. 

\(\sigma\) is the output.
Utility of measurement \((\sigma, \tau)\):

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

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

\[
U(\sigma) = E_{p(\tau|X)}[U(\sigma, \tau)]
\]
Utility of measurement \((\sigma, \tau)\):

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

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

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

$$U(\sigma, \tau) = R(\sigma, \tau) - C(\sigma)$$

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

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

Choose best measurement feature $\sigma$:

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

\[ n = 1000 \text{ total examples (sentences), 45 possible labels} \]

**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., \( \sum_i \mathbb{I}[x_i = the, y_i = DT] \))

Use label entropy as surrogate for assessing measurements

**Test accuracy (on 100 examples):**

![Graph showing test accuracy over number of measurements]

(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
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 $\hat{p}$

Measurements
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