Training Strategies

CS 6355: Structured Prediction

So far we saw

- What is structured output prediction?
- Different ways for modeling structured prediction
  - Conditional random fields, factor graphs, constraints
- What we only occasionally touched upon:
  - Algorithms for training and inference
    - Viterbi (inference in sequences)
    - Structured perceptron (training in general)

Rest of the semester

- Strategies for training
  - Structural SVM
  - Stochastic gradient descent
  - More on local vs. global training
- Algorithms for inference
  - Exact inference
  - “Approximate” inference
  - Formulating inference problems in general
- Latent/hidden variables, representations and such

Up next

- Structural Support Vector Machine
  - How it naturally extends multiclass SVM
- Empirical Risk Minimization
  - Or: how structural SVM and CRF are solving very similar problems
- Training Structural SVM via stochastic gradient descent
  - And some tricks

Where are we?

- Structural Support Vector Machine
  - How it naturally extends multiclass SVM
- Empirical Risk Minimization
  - Or: how structural SVM and CRF are solving very similar problems
- Training Structural SVM via stochastic gradient descent
  - And some tricks

Recall: Binary and Multiclass SVM

- Binary SVM
  - Maximize margin
  - Equivalently,

    Minimize norm of weights such that the closest points to the hyperplane have a score §1

- Multiclass SVM
  - Each label has a different weight vector (like one-vs-all)
  - Maximize multiclass margin
  - Equivalently,

    Minimize total norm of the weights such that the true label is scored at least 1 more than the second best one

Multiclass SVM in the separable case

Recall hard binary SVM

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$
$$\text{s.t.} \forall i, \quad y_i\mathbf{w}^T\mathbf{x}_i \geq 1$$

We have a data set D = {<$\mathbf{x}_i$, $\mathbf{y}_i$>}

Multiclass SVM in the separable case

Recall hard binary SVM

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$
$$\text{s.t.} \forall i, \quad y_i \mathbf{w}^T \mathbf{x}_i \geq 1$$

Size of the weights. Effectively, regularizer

We have a data set D = {<$\mathbf{x}_i$, $\mathbf{y}_i$>}

$$\min_{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K} \quad \frac{1}{2}\sum_k \mathbf{w}_k^T \mathbf{w}_k$$
$$\text{s.t.} \quad \mathbf{w}_{\mathbf{y}_i}^T \mathbf{x} - \mathbf{w}_k^T \mathbf{x} \geq 1$$

$$\forall (\mathbf{x}_i, \mathbf{y}_i) \in D,$$
$$k \in \{1, 2, \cdots, K\}, k \neq \mathbf{y}_i,$$

The score for the true label is higher than the score for ***any*** other label by 1

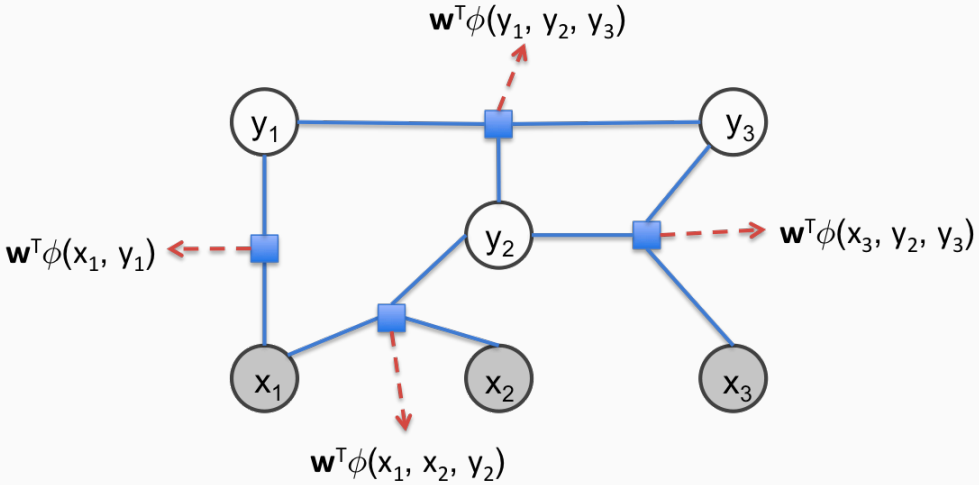
Structural SVM: First attempt

Modeling

Suppose we have some definition of a structure (a factor graph)

- And feature definitions for each factor (i.e. "part") p as $\Phi_p(\mathbf{x}, \mathbf{y}_p)$
- Remember: we can talk about the feature vector for the *entire* structure
  - Features sum over the parts

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{p \in \text{parts}(\mathbf{x})} \Phi_p(\mathbf{x}, \mathbf{y}_p)$$



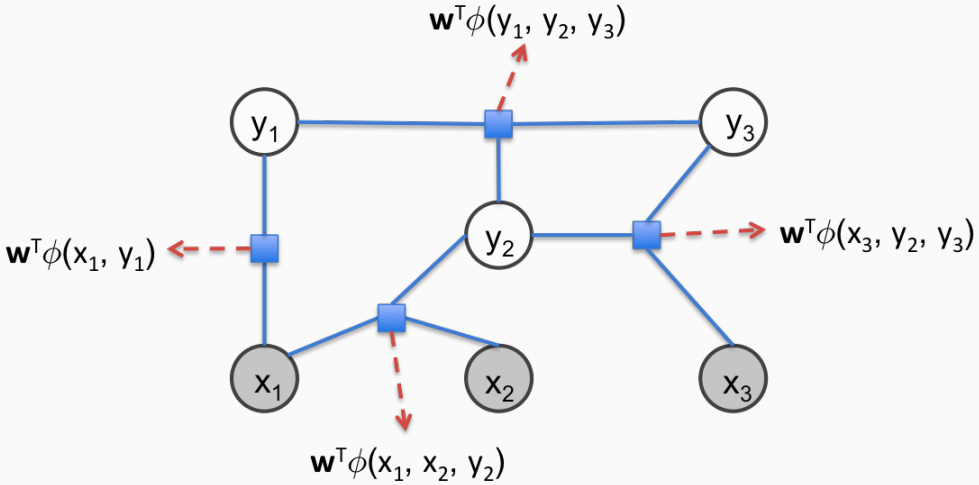
Structural SVM: First attempt

Suppose we have some definition of a structure (a factor graph)

Modeling

- And feature definitions for each factor (i.e. "part") p as $\Phi_p(\mathbf{x}, \mathbf{y}_p)$
- Remember: we can talk about the feature vector for the *entire* structure
  - Features sum over the parts

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{p \in \text{parts}(\mathbf{x})} \Phi_p(\mathbf{x}, \mathbf{y}_p)$$



Data

We also have a data set $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}$

Structural SVM: First attempt

Suppose we have some definition of a structure (a factor graph)

Modeling

- And feature definitions for each factor (i.e. "part") p as $\Phi_p(\mathbf{x}, \mathbf{y}_p)$
- Remember: we can talk about the feature vector for the *entire* structure
  - Features sum over the parts

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{p \in \text{parts}(\mathbf{x})} \Phi_p(\mathbf{x}, \mathbf{y}_p)$$

Data

We also have a data set $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}$

What we want from training (following the multiclass idea)

For each training example $(\mathbf{x}_i, \mathbf{y}_i)$:

- The annotated structure $\mathbf{y}_i$ gets the highest score among all structures

Structural SVM: First attempt

Suppose we have some definition of a structure (a factor graph)

Modeling

- And feature definitions for each factor (i.e. "part") p as $\Phi_p(\mathbf{x}, \mathbf{y}_p)$
- Remember: we can talk about the feature vector for the *entire* structure
  - Features sum over the parts

Data

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{p \in \text{parts}(\mathbf{x})} \Phi_p(\mathbf{x}, \mathbf{y}_p)$$

We also have a data set $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}$

What we want from training (following the multiclass idea)

For each training example $(\mathbf{x}_i, \mathbf{y}_i)$:

- The annotated structure $\mathbf{y}_i$ gets the highest score among all structures
- Or to be safe, $\mathbf{y}_i$ gets a score that is at least one more than all other structures

$$\forall \mathbf{y} \neq \mathbf{y}_i, \qquad \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{y}) + 1$$

Structural SVM: First attempt

Suppose we have some definition of a structure (a factor graph)

Modeling

- And feature definitions for each factor (i.e. "part") p as $\Phi_p(\mathbf{x}, \mathbf{y}_p)$
- Remember: we can talk about the feature vector for the *entire* structure
  - Features sum over the parts

Data

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{p \in \text{parts}(\mathbf{x})} \Phi_p(\mathbf{x}, \mathbf{y}_p)$$

We also have a data set $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}$

What we want from training (following the multiclass idea)

For each training example $(\mathbf{x}_i, \mathbf{y}_i)$:

- The annotated structure $\mathbf{y}_i$ gets the highest score among all structures
- Or to be safe, $\mathbf{y}_i$ gets a score that is at least one more than all other structures

$$\forall \mathbf{y} \neq \mathbf{y}_i, \qquad \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{y}) + 1$$

Structural SVM: First attempt

Suppose we have some definition of a structure (a factor graph)

Modeling

- And feature definitions for each factor (i.e. "part") p as $\Phi_p(\mathbf{x}, \mathbf{y}_p)$
- Remember: we can talk about the feature vector for the *entire* structure
  - Features sum over the parts

Data

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{p \in \text{parts}(\mathbf{x})} \Phi_p(\mathbf{x}, \mathbf{y}_p)$$

We also have a data set $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}$

What we want from training (following the multiclass idea)

For each training example $(\mathbf{x}_i, \mathbf{y}_i)$:

- The annotated structure $\mathbf{y}_i$ gets the highest score among all structures
- Or to be safe, $\mathbf{y}_i$ gets a score that is at least one more than all other structures

$$\forall \mathbf{y} \neq \mathbf{y}_i, \qquad \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{y}) + 1$$

Structural SVM: First attempt

Suppose we have some definition of a structure (a factor graph)

Modeling

- And feature definitions for each factor (i.e. "part") p as $\Phi_p(\mathbf{x}, \mathbf{y}_p)$
- Remember: we can talk about the feature vector for the *entire* structure
  - Features sum over the parts

Data

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{p \in \text{parts}(\mathbf{x})} \Phi_p(\mathbf{x}, \mathbf{y}_p)$$

We also have a data set $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}$

What we want from training (following the multiclass idea)

For each training example $(\mathbf{x}_i, \mathbf{y}_i)$:

- The annotated structure $\mathbf{y}_i$ gets the highest score among all structures
- Or to be safe, $\mathbf{y}_i$ gets a score that is at least one more than all other structures

$$\forall \mathbf{y} \neq \mathbf{y}_i, \qquad \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{y}) + 1$$

Structural SVM: First attempt

Maximize margin

$$\text{s.t.} \quad \text{Score for gold structure} \geq \text{Score for other structure} + 1$$

For every training example

Structural SVM: First attempt

Maximize margin

$$\text{s.t.} \quad \mathbf{w}^T \phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T \phi(\mathbf{x}_i, \mathbf{y}) + 1 \quad \forall (\mathbf{x}_i, \mathbf{y}_i) \in D, \forall \mathbf{y} \neq \mathbf{y}_i$$

Score for gold

Score for other

Input with gold structure

Some other structure

Structural SVM: First attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Maximize margin by minimizing norm of **w**

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}) + 1 \quad \forall(\mathbf{x}_i, \mathbf{y}_i) \in D, \forall \mathbf{y} \neq \mathbf{y}_i$$

Score for gold

Score for other

Input with gold structure

Some other structure

Structural SVM: First attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Maximize margin by minimizing norm of **w**

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}) + 1 \quad \forall(\mathbf{x}_i, \mathbf{y}_i) \in D, \forall \mathbf{y} \neq \mathbf{y}_i$$

- Score for gold
- Score for other
- Input with gold structure
- Some other structure

Problem?

Structural SVM: First attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Maximize margin by minimizing norm of **w**

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + 1 \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y} \neq \mathbf{y}_i$$

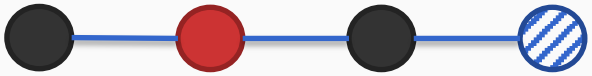
Score for gold

Score for other

Input with gold structure

Some other structure

Problem



Gold structure

Structural SVM: First attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Maximize margin by minimizing norm of **w**

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + 1 \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y} \neq \mathbf{y}_i$$

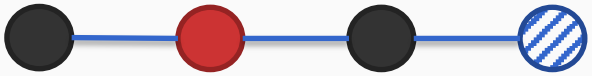
Score for gold

Score for other

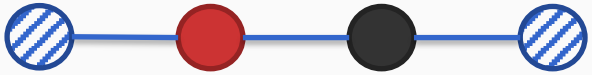
Input with gold structure

Some other structure

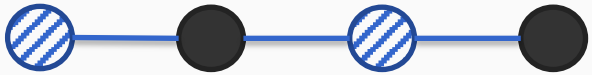
Problem



Gold structure



Other structure A: Only one mistake



Other structure B: Fully incorrect

Structural SVM: First attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Maximize margin by minimizing norm of **w**

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + 1 \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y} \neq \mathbf{y}_i$$

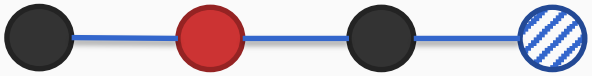
Score for gold

Score for other

Input with gold structure

Some other structure

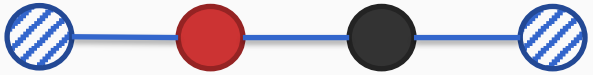
Problem



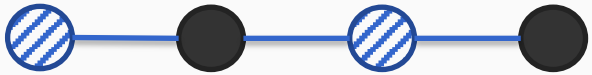
Gold structure

Structure B has is more wrong, but this formulation will be happy if *both* A & B are scored one less than gold!

No partial credit!



Other structure A: Only one mistake



Other structure B: Fully incorrect

Structural SVM: Second attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

← Maximize margin by minimizing norm of **w**

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + 1 \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y} \neq \mathbf{y}_i$$

- $\mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i)$: Score for gold
- $\mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y})$: Score for other
- $\forall(\mathbf{x}_i,\mathbf{y}_i) \in D$: Input with gold structure
- $\forall \mathbf{y} \neq \mathbf{y}_i$: Some other structure

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

Structural SVM: Second attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Maximize margin by minimizing norm of **w**

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + 1 \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y} \neq \mathbf{y}_i$$

Score for gold

Score for other

Input with gold structure

Some other structure

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

Hamming distance between structures: Counts the number of differences between them

Structural SVM: Second attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

← Maximize margin by minimizing norm of **w**

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + 1 \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y} \neq \mathbf{y}_i$$

Score for gold

Score for other

Input with gold structure

Some other structure

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

because the Hamming distance of **y** and itself is zero

Structural SVM: Second attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}) + \Delta(\mathbf{y}, \mathbf{y}_i) \quad \forall(\mathbf{x}_i, \mathbf{y}_i) \in D, \forall \mathbf{y}$$

Intuition

- It is okay for a structure that is close (in Hamming sense) to the true one to get a score that is close to the true structure
- Structures that are very different from the true structure should get much lower scores

Structural SVM: Second attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

← Maximize margin by minimizing norm of **w**

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}) + \Delta(\mathbf{y}, \mathbf{y}_i) \quad \forall(\mathbf{x}_i, \mathbf{y}_i) \in D, \forall \mathbf{y}$$

Intuition

- It is okay for a structure that is close (in Hamming sense) to the true one to get a score that is close to the true structure
- Structures that are very different from the true structure should get much lower scores

Structural SVM: Second attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

← Maximize margin by minimizing norm of $\mathbf{w}$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}) + \Delta(\mathbf{y}, \mathbf{y}_i) \quad \forall(\mathbf{x}_i, \mathbf{y}_i) \in D, \forall \mathbf{y}$$

↓ Input with gold structure

Intuition

- It is okay for a structure that is close (in Hamming sense) to the true one to get a score that is close to the true structure
- Structures that are very different from the true structure should get much lower scores

Structural SVM: Second attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

← Maximize margin by minimizing norm of $\mathbf{w}$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}) + \Delta(\mathbf{y}, \mathbf{y}_i) \quad \forall(\mathbf{x}_i, \mathbf{y}_i) \in D, \forall \mathbf{y}$$

↓ Score for gold

↓ Input with gold structure

Intuition

- It is okay for a structure that is close (in Hamming sense) to the true one to get a score that is close to the true structure
- Structures that are very different from the true structure should get much lower scores

Structural SVM: Second attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

← Maximize margin by minimizing norm of $\mathbf{w}$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}) + \Delta(\mathbf{y}, \mathbf{y}_i) \quad \forall(\mathbf{x}_i, \mathbf{y}_i) \in D, \forall \mathbf{y}$$

↓ Score for gold

↓ Score for other

↓ Input with gold structure

Intuition

- It is okay for a structure that is close (in Hamming sense) to the true one to get a score that is close to the true structure
- Structures that are very different from the true structure should get much lower scores

Structural SVM: Second attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

← Maximize margin by minimizing norm of $\mathbf{w}$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}) + \Delta(\mathbf{y}, \mathbf{y}_i) \quad \forall(\mathbf{x}_i, \mathbf{y}_i) \in D, \forall \mathbf{y}$$

Score for gold

Score for other

Input with gold structure

Hamming distance between structures.
Defined to be zero if $\mathbf{y} = \mathbf{y}_i$

Intuition

- It is okay for a structure that is close (in Hamming sense) to the true one to get a score that is close to the true structure
- Structures that are very different from the true structure should get much lower scores

Structural SVM: Second attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Maximize margin by minimizing norm of $\mathbf{w}$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}) + \Delta(\mathbf{y}, \mathbf{y}_i) \quad \forall(\mathbf{x}_i, \mathbf{y}_i) \in D, \forall \mathbf{y}$$

Score for gold

Score for other

Hamming distance between structures.
Defined to be zero if $\mathbf{y} = \mathbf{y}_i$

Input with gold structure

Another structure, could be $\mathbf{y}_i$

Intuition

- It is okay for a structure that is close (in Hamming sense) to the true one to get a score that is close to the true structure
- Structures that are very different from the true structure should get much lower scores

Structural SVM: Second attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

← Maximize margin by minimizing norm of $\mathbf{w}$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}) + \Delta(\mathbf{y}, \mathbf{y}_i) \quad \forall(\mathbf{x}_i, \mathbf{y}_i) \in D, \forall \mathbf{y}$$

Score for gold

Score for other

Hamming distance between structures.
Defined to be zero if $\mathbf{y} = \mathbf{y}_i$

Input with gold structure

Another structure, could be $\mathbf{y}_i$

Problem?

Structural SVM: Second attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

← Maximize margin by minimizing norm of $\mathbf{w}$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i, \mathbf{y}) + \Delta(\mathbf{y}, \mathbf{y}_i) \quad \forall(\mathbf{x}_i, \mathbf{y}_i) \in D, \forall \mathbf{y}$$

Score for gold

Score for other

Input with gold structure

Another structure, could be $\mathbf{y}_i$

Hamming distance between structures.
Defined to be zero if $\mathbf{y} = \mathbf{y}_i$

Problem? What if the data is not separable?

What if these constraints are not satisfied for any $\mathbf{w}$ for a given dataset?

Structural SVM: Third attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

← Maximize margin by minimizing norm of **w**

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

- $\mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i)$: Score for gold
- $\mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y})$: Score for other
- $\Delta(\mathbf{y},\mathbf{y}_i)$: Hamming distance
- $\forall(\mathbf{x}_i,\mathbf{y}_i) \in D$: Input with gold structure
- $\forall \mathbf{y}$: All structures

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \xi_i \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

Slack variable for each example, **must be positive**

Structural SVM: Third attempt

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

← Maximize margin by minimizing norm of **w**

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

Score for gold

Score for other

Hamming distance

Input with gold structure

All structures

Also minimize total slack

Slack variable for each example, **must be positive**

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i$$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \xi_i \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

Structural SVM: Third attempt

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i$$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \xi_i \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

Structural SVM: Third attempt

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i$$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \xi_i \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

Input with gold structure

Another structure

For every labeled example, and every competing structure

Structural SVM: Third attempt

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i$$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \xi_i \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall\mathbf{y}$$

Score for gold

Score for other

Hamming distance

Input with gold structure

Another structure

For every labeled example, and every competing structure,
the score for the ground truth should be greater than the score for the competing structure by the Hamming distance between them

Structural SVM: Third attempt

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i$$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \xi_i \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

Score for gold

Score for other

Hamming distance

Slack variable for each example

Input with gold structure

Another structure

Slack variables allow some examples to be misclassified.

Structural SVM: Third attempt

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i$$

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \xi_i \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

$$\xi_i \geq 0 \qquad \forall i.$$

Score for gold

Score for other

Hamming distance

Input with gold structure

Another structure

All slacks must be positive

Slack variable for each example

Slack variables allow some examples to be misclassified.

Structural SVM: Third attempt

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i$$

Maximize margin & minimize slack
C: the tradeoff parameter

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \xi_i \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

$$\xi_i \geq 0 \qquad \forall i.$$

Score for gold

Score for other

Hamming distance

Input with gold structure

Another structure

All slacks must be positive

Slack variable for each example

Slack variables allow some examples to be misclassified.

Minimizing the slack forces this to happen as few times as possible

Structural SVM: Third attempt

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i$$

Maximize margin & minimize slack
C: the tradeoff parameter

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \xi_i \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

$$\xi_i \geq 0 \qquad \forall i.$$

Score for gold

Score for other

Hamming distance

Input with gold structure

Another structure

All slacks must be positive

Slack variable for each example

Slack variables allow some examples to be misclassified.

Minimizing the slack forces this to happen as few times as possible

Questions?

Structural SVM

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i$$

Maximize margin, minimize slack
C: the tradeoff parameter

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \xi_i \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

$$\xi_i \geq 0 \qquad \forall i.$$

Score for gold

Score for other

All slacks must be positive

Hamming distance

Slack variable for each example

Input with gold structure

Another structure

Structural SVM

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i$$

Maximize margin, minimize slack
C: the tradeoff parameter

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \xi_i \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

$$\xi_i \geq 0 \qquad \forall i.$$

Score for gold

Score for other

All slacks must be positive

Hamming distance

Slack variable for each example

Input with gold structure

Another structure

Equivalent formulation

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \max_{\mathbf{y}} \left(\mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i)\right)$$

Structural SVM

Questions?

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i$$

Maximize margin, minimize slack
C: the tradeoff parameter

$$\text{s.t.} \quad \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i) \geq \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \xi_i \quad \forall(\mathbf{x}_i,\mathbf{y}_i) \in D, \forall \mathbf{y}$$

$$\xi_i \geq 0 \qquad \forall i.$$

Score for gold

Score for other

Hamming distance

Input with gold structure

Another structure

All slacks must be positive

Slack variable for each example

Equivalent formulation

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \max_{\mathbf{y}} \left(\mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i)\right)$$

Comments

- Other slightly different formulations exist
  - Generally same principle
- Multiclass is a special case of structure
  - Structural SVM strictly generalizes multiclass SVM Exercise: Work it out
- Can be seen as minimizing structured version of hinge loss
  - Remember **empirical risk minimization**?
- Learning as optimization
  - We have framed the optimization problem
  - We haven't seen how it can be solved yet
    - That is, we don't have a learning algorithm yet

Where are we?

- Structural Support Vector Machine
  - How it naturally extends multiclass SVM
- Empirical Risk Minimization
  - Or: how structural SVM and CRF are solving very similar problems
- Training Structural SVM via stochastic gradient descent
  - And some tricks

Broader picture: Learning as loss minimization

- Collect some annotated data. More is generally better
- Pick a hypothesis class (also called model)
  - Decide how the score decomposes over the parts of the output
- Choose a loss function
  - Decide on how to penalize incorrect decisions
- Learning = minimize empirical risk + regularizer
  - Typically an optimization procedure needed here

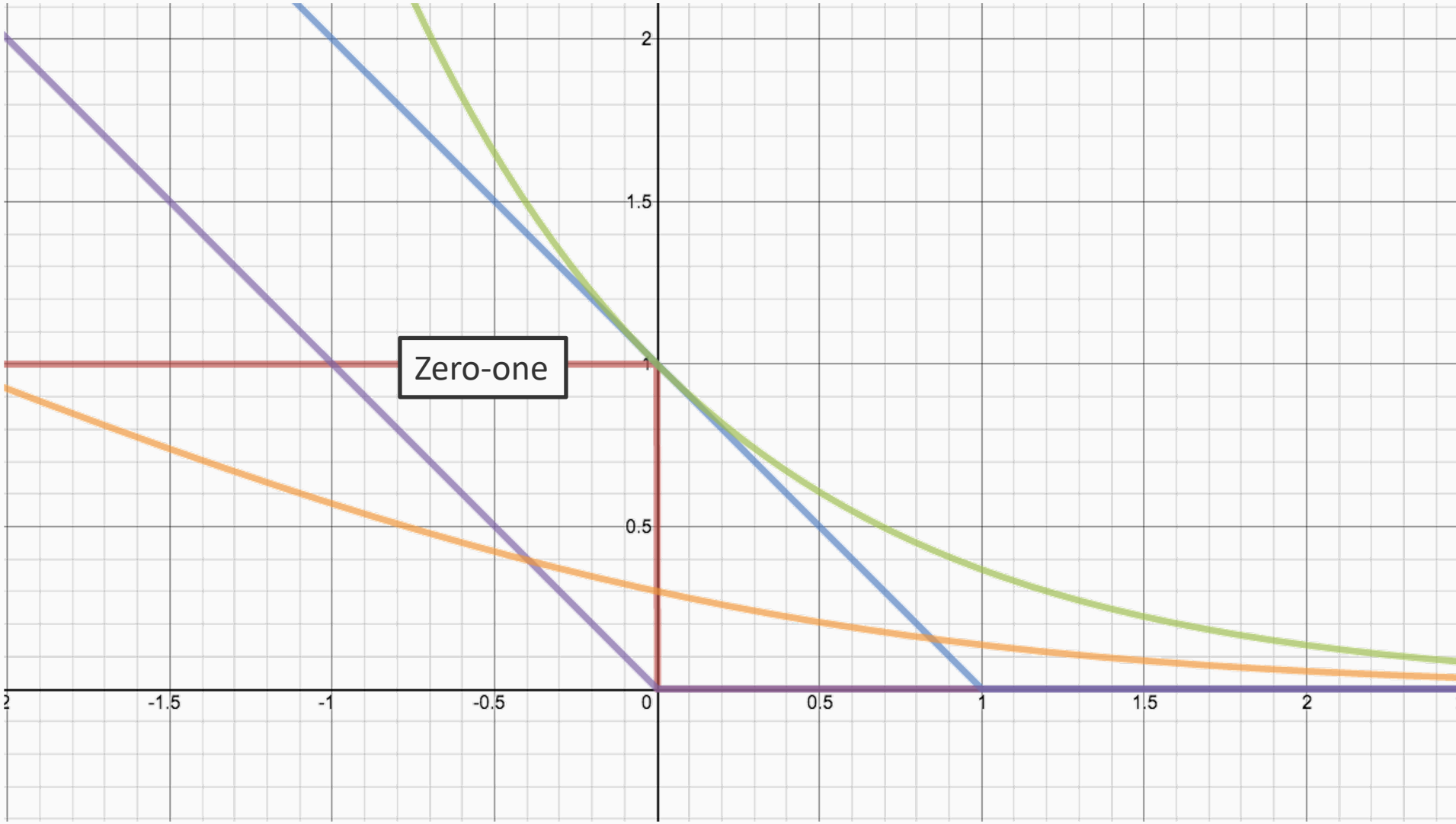
This must look familiar. We have seen this before for binary classification!

Empirical risk minimization

- Suppose the function *Loss* scores the quality of a prediction with respect to the true structure
  - *Loss*(*f*(**x**), **y**) tells us how good *f* is for this **x** by comparing it against **y**
- Evaluate the quality of the predictor *f* by averaging over the unknown distribution P that generates data
  - *Expected risk:* $\mathbb{E}\left[Loss(f(\mathbf{x}), \mathbf{y})\right]$
- We don't know P, so use the empirical risk $\frac{1}{N}\sum_i Loss\left(f(\mathbf{x}_i), \mathbf{y}_i\right)$ possibly with regularizer

Learning: Minimizing regularized risk; various algorithms exist

The loss function zoo: binary classification



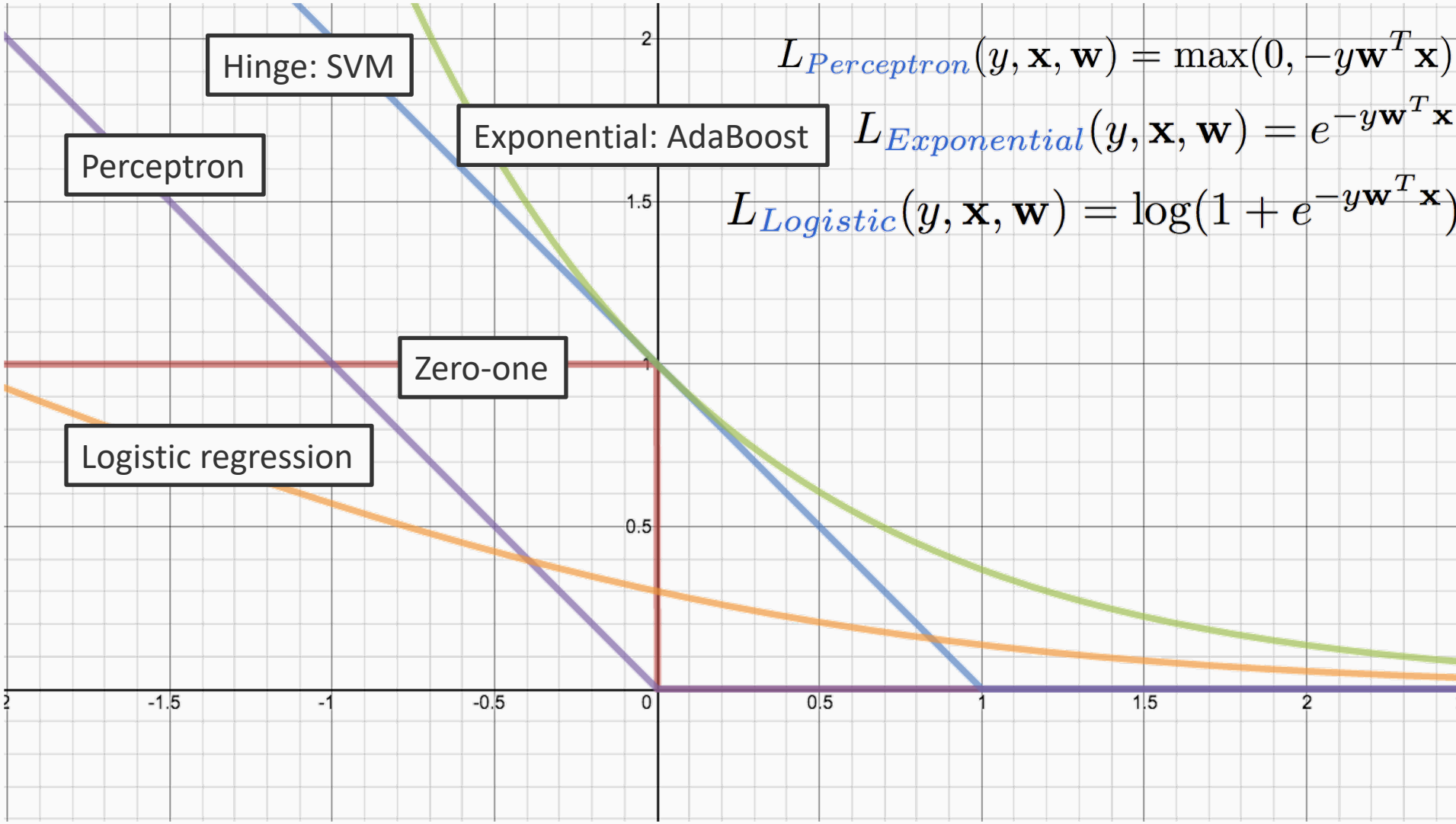
The loss function zoo

$$L_{Hinge}(y, \mathbf{x}, \mathbf{w}) = \max(0, 1 - y\mathbf{w}^T\mathbf{x})$$

$$L_{Perceptron}(y, \mathbf{x}, \mathbf{w}) = \max(0, -y\mathbf{w}^T\mathbf{x})$$

$$L_{Exponential}(y, \mathbf{x}, \mathbf{w}) = e^{-y\mathbf{w}^T\mathbf{x}}$$

$$L_{Logistic}(y, \mathbf{x}, \mathbf{w}) = \log(1 + e^{-y\mathbf{w}^T\mathbf{x}})$$



Structured classifiers: Different learning objectives

- Structural SVM

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathrm{C}\sum_i \max_{\mathbf{y}}\left(\mathbf{w}^T\Phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i)\right)$$

- Conditional Random Field (via the maximum a posteriori criterion)

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathrm{C}\sum_i -\log P(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{w})$$

Where P is defined as

$$P(\mathrm{y_i} \mid x_i, w) = \frac{\exp(\mathbf{w}^T\Phi(\mathbf{x}_i,\mathbf{y}_i)}{Z(\mathbf{x}_i,\mathbf{w})}$$

Structured classifiers: Different learning objectives

- Structural SVM

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathrm{C}\sum_i \max_{\mathbf{y}} \left(\mathbf{w}^T\Phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i)\right)$$

Regularizer

- CRF

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathrm{C}\sum_i -\log P(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{w})$$

Where P is defined as

$$P(\mathrm{y_i} \mid x_i, w) = \frac{\exp(\mathbf{w}^T\Phi(\mathbf{x}_i,\mathbf{y}_i)}{Z(\mathbf{x}_i,\mathbf{w})}$$

Structured classifiers: Different learning objectives

- Structural SVM

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathrm{C}\sum_i \max_{\mathbf{y}}\left(\mathbf{w}^T\Phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i)\right)$$

Regularizer

How badly does **w** do on the training data

- CRF

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathrm{C}\sum_i -\log P(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{w})$$

Where P is defined as

$$P(\mathrm{y_i} \mid x_i, w) = \frac{\exp(\mathbf{w}^T\Phi(\mathbf{x}_i,\mathbf{y}_i)}{Z(\mathbf{x}_i,\mathbf{w})}$$

Structured classifiers: Different learning objectives

- Structural SVM

Structured hinge loss

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathrm{C}\sum_i \max_{\mathbf{y}}\left(\mathbf{w}^T\Phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i)\right)$$

Regularizer

How badly does **w** do on the training data

- CRF

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathrm{C}\sum_i -\log P(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{w})$$

Where P is defined as

$$P(\mathrm{y_i} \mid x_i, w) = \frac{\exp(\mathbf{w}^T\Phi(\mathbf{x}_i,\mathbf{y}_i)}{Z(\mathbf{x}_i,\mathbf{w})}$$

Structured classifiers: Different learning objectives

- Structural SVM

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathrm{C}\sum_i \max_{\mathbf{y}}\left(\mathbf{w}^T\Phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i)\right)$$

Regularizer

How badly does **w** do on the training data

- CRF

Log loss

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathrm{C}\sum_i -\log P(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{w})$$

Where P is defined as

$$P(\mathrm{y_i} \mid x_i, w) = \frac{\exp(\mathbf{w}^T\Phi(\mathbf{x}_i,\mathbf{y}_i)}{Z(\mathbf{x}_i,\mathbf{w})}$$

Structured classifiers: Different learning objectives

- Structural SVM

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathrm{C}\sum_i \max_{\mathbf{y}}\left(\mathbf{w}^T\Phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i)\right)$$

Regularizer

How badly does **w** do on the training data

- CRF

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathrm{C}\sum_i -\log P(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{w})$$

- Structured Perceptron

$$\min_{\mathbf{w}} \sum_i \max_{\mathbf{y}}\left(\mathbf{w}^T\Phi(\mathbf{x}_i,\mathbf{y}) - \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i)\right)$$

Structured classifiers: Different learning objectives

- Structural SVM

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathrm{C}\sum_i \max_{\mathbf{y}}\left(\mathbf{w}^T\Phi(\mathbf{x}_i,\mathbf{y}) + \Delta(\mathbf{y},\mathbf{y}_i) - \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i)\right)$$

Regularizer

How badly does **w** do on the training data

- CRF

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathrm{C}\sum_i -\log P(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{w})$$

- Structured Perceptron

Structured Perceptron loss

$$\min_{\mathbf{w}} \sum_i \max_{\mathbf{y}}\left(\mathbf{w}^T\Phi(\mathbf{x}_i,\mathbf{y}) - \mathbf{w}^T\phi(\mathbf{x}_i,\mathbf{y}_i)\right)$$

Summary

- Different structured training objectives are really different loss functions

- The structured versions of hinge, log and Perceptron losses all involve inference
  - Hinge, Perceptron: Solve a maximization problem
  - Log: Solve an expectation problem

- Learning as stochastic optimization, even for structures
  - But, computing the loss (and the gradient) can be expensive