Introduction to Structured SVM

Raviteja Vemulapalli

University of Maryland, College Park

August 1, 2013

Structured Output Learning

- Input: $x \in \mathcal{X}$
- Output: $y \in \mathcal{Y}$
- Learn a prediction function
    - $f : \mathcal{X} \longrightarrow \mathcal{Y}$

Structured Output Learning

- Input: $x \in \mathcal{X}$
- Output: $y \in \mathcal{Y}$
- Learn a prediction function
  - $f : \mathcal{X} \longrightarrow \mathcal{Y}$
- Typical scenarios
  - Binary classification: $\mathcal{Y} = \{\pm 1\}$
  - Regression: $\mathcal{Y} = \mathcal{R}$

Structured Output Learning

- Input: $x \in \mathcal{X}$
- Output: $y \in \mathcal{Y}$
- Learn a prediction function
  - $f : \mathcal{X} \longrightarrow \mathcal{Y}$
- Typical scenarios
  - Binary classification: $\mathcal{Y} = \{\pm 1\}$
  - Regression: $\mathcal{Y} = \mathcal{R}$
- Structured output learning extends this concept to more complex output spaces.

Structured Output Learning

- Input: $x \in \mathcal{X}$
- Output: $y \in \mathcal{Y}$
- Learn a prediction function
  - $f : \mathcal{X} \longrightarrow \mathcal{Y}$
- Typical scenarios
  - Binary classification: $\mathcal{Y} = \{\pm 1\}$
  - Regression: $\mathcal{Y} = \mathcal{R}$
- Structured output learning extends this concept to more complex output spaces.
  - Multi-class classification: $\mathcal{Y}$ = set of class labels

Structured Output Learning

- Input: $x \in \mathcal{X}$
- Output: $y \in \mathcal{Y}$
- Learn a prediction function
  - $f : \mathcal{X} \longrightarrow \mathcal{Y}$
- Typical scenarios
  - Binary classification: $\mathcal{Y} = \{\pm 1\}$
  - Regression: $\mathcal{Y} = \mathcal{R}$
- Structured output learning extends this concept to more complex output spaces.
  - Multi-class classification: $\mathcal{Y}$ = set of class labels
  - Segmentation: $\mathcal{Y}$ = set of segmentation masks

Structured Output Learning

- Input: $x \in \mathcal{X}$
- Output: $y \in \mathcal{Y}$
- Learn a prediction function
  - $f : \mathcal{X} \longrightarrow \mathcal{Y}$
- Typical scenarios
  - Binary classification: $\mathcal{Y} = \{\pm 1\}$
  - Regression: $\mathcal{Y} = \mathcal{R}$
- Structured output learning extends this concept to more complex output spaces.
  - Multi-class classification: $\mathcal{Y}$ = set of class labels
  - Segmentation: $\mathcal{Y}$ = set of segmentation masks
  - Object localization: $\mathcal{Y}$ = set of bounding boxes

Structured Prediction

- Loss function: $\Delta : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathcal{R}_+$ with $\Delta(y, y) = 0$.

Structured Prediction

- Loss function: $\Delta : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathcal{R}_+$ with $\Delta(y, y) = 0$.
- Prediction function

$$f(x) = \underset{y \in \mathcal{Y}}{\text{argmax}}\; g(x, y)$$

  - $g : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathcal{R}$ is an auxiliary function.
  - This can be seen as a generalization of MAP inference in which $g(x, y) = p(y|x)$.

Structured SVM

- $g(x, y, w) = w^\top \phi(x, y)$ where $\phi(x, y)$ is a joint feature map defined on $\mathcal{X} \times \mathcal{Y}$.

Structured SVM

- $g(x, y, w) = w^\top \phi(x, y)$ where $\phi(x, y)$ is a joint feature map defined on $\mathcal{X} \times \mathcal{Y}$.
- $f(x) = \underset{y \in \mathcal{Y}}{\text{argmax}}\ w^\top \phi(x, y)$.

Structured SVM

- $g(x, y, w) = w^\top \phi(x, y)$ where $\phi(x, y)$ is a joint feature map defined on $\mathcal{X} \times \mathcal{Y}$.
- $f(x) = \underset{y \in \mathcal{Y}}{\text{argmax}} \; w^\top \phi(x, y)$.
- Structured SVM learning: Learn the parameters w.

Structured SVM

- $g(x, y, w) = w^\top \phi(x, y)$ where $\phi(x, y)$ is a joint feature map defined on $\mathcal{X} \times \mathcal{Y}$.
- $f(x) = \underset{y \in \mathcal{Y}}{\text{argmax}}\ w^\top \phi(x, y)$.
- Structured SVM learning: Learn the parameters w.
- Structured SVM prediction: Use the learned w to find $f(x)$.

Structured SVM Learning

- We learn w such that certain constraints on the training data are satisfied.

Structured SVM Learning

- We learn w such that certain constraints on the training data are satisfied.
- To avoid over-fitting, we use the standard Tikhonov regularizer:

$$\mathcal{R}(w) = \frac{1}{2}\|w\|_2^2$$

Structured SVM Learning - Margin rescaling

- Constraints: Some errors might be worse than others. So, for every training sample (x^i, y^i) we would like to have

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq \Delta(y^i, y),\ \forall y \in \mathcal{Y}.$$

Structured SVM Learning - Margin rescaling

- Constraints: Some errors might be worse than others. So, for every training sample (x^i, y^i) we would like to have

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq \Delta(y^i, y), \ \forall y \in \mathcal{Y}.$$

$$\iff \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\} \leq 0.$$

Structured SVM Learning - Margin rescaling

- Constraints: Some errors might be worse than others. So, for every training sample (x^i, y^i) we would like to have

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq \Delta(y^i, y),\ \forall y \in \mathcal{Y}.$$

$$\iff \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\} \leq 0.$$

- $\ell(x^i, y^i, w) = \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}.$

Structured SVM Learning - Margin rescaling

- Constraints: Some errors might be worse than others. So, for every training sample (x^i, y^i) we would like to have

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq \Delta(y^i, y), \ \forall y \in \mathcal{Y}.$$
$$\iff \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\} \leq 0.$$

- $\ell(x^i, y^i, w) = \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}.$
- $\ell(x^i, y^i, w)$ is called margin-rescaled hinge loss.

Structured SVM Learning - Margin rescaling

- Constraints: Some errors might be worse than others. So, for every training sample (x^i, y^i) we would like to have

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq \Delta(y^i, y), \ \forall y \in \mathcal{Y}.$$

$$\iff \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\} \leq 0.$$

- $\ell(x^i, y^i, w) = \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}.$
- $\ell(x^i, y^i, w)$ is called margin-rescaled hinge loss.
- Note that $\ell(x^i, y^i, w) \geq 0$, since $y = y^i$ makes it zero.

Structured SVM Learning - Margin rescaling

- Constraints: Some errors might be worse than others. So, for every training sample (x^i, y^i) we would like to have

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq \Delta(y^i, y), \ \forall y \in \mathcal{Y}.$$
$$\iff \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\} \leq 0.$$

- $\ell(x^i, y^i, w) = \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}.$
- $\ell(x^i, y^i, w)$ is called margin-rescaled hinge loss.
- Note that $\ell(x^i, y^i, w) \geq 0$, since $y = y^i$ makes it zero.
- $\ell(x^i, y^i, w)$ is a convex function of w.

Structured SVM Learning - Margin rescaling

- Constraints: Some errors might be worse than others. So, for every training sample (x^i, y^i) we would like to have

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq \Delta(y^i, y), \ \forall y \in \mathcal{Y}.$$

$$\iff \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\} \leq 0.$$

- $\ell(x^i, y^i, w) = \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}.$
- $\ell(x^i, y^i, w)$ is called margin-rescaled hinge loss.
- Note that $\ell(x^i, y^i, w) \geq 0$, since $y = y^i$ makes it zero.
- $\ell(x^i, y^i, w)$ is a convex function of w.

$$\underset{w}{\text{minimize}} \ \ \frac{1}{2}\|w\|_2^2 + \frac{C}{N} \sum_{i=1}^{N} \ell(x^i, y^i, w)$$

Structured SVM Learning - Slack rescaling

- Constraints: Use fixed margin

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq 1, \ \forall y \in \mathcal{Y}$$

Structured SVM Learning - Slack rescaling

- Constraints: Use fixed margin

$$
\begin{aligned}
& w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq 1, \ \forall y \in \mathcal{Y} \\
\iff & 1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) \leq 0, \ \forall y \in \mathcal{Y}.
\end{aligned}
$$

Structured SVM Learning - Slack rescaling

- Constraints: Use fixed margin

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq 1, \ \forall y \in \mathcal{Y}$$
$$\iff 1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) \leq 0, \ \forall y \in \mathcal{Y}.$$

- Penalize the constraint violation according to $\Delta(y^i, y)$.

$$\ell(x^i, y^i, w) = \max_{y \in \mathcal{Y}} \ \{\Delta(y^i, y)(1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y))\}$$

Structured SVM Learning - Slack rescaling

- Constraints: Use fixed margin

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq 1, \ \forall y \in \mathcal{Y}$$
$$\iff 1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) \leq 0, \ \forall y \in \mathcal{Y}.$$

- Penalize the constraint violation according to $\Delta(y^i, y)$.

$$\ell(x^i, y^i, w) = \max_{y \in \mathcal{Y}} \ \{\Delta(y^i, y)(1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y))\}$$

- $\ell(x^i, y^i, w)$ is called slack-rescaled hinge loss.

Structured SVM Learning - Slack rescaling

- Constraints: Use fixed margin

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq 1, \ \forall y \in \mathcal{Y}$$
$$\iff 1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) \leq 0, \ \forall y \in \mathcal{Y}.$$

- Penalize the constraint violation according to $\Delta(y^i, y)$.

$$\ell(x^i, y^i, w) = \max_{y \in \mathcal{Y}} \ \{\Delta(y^i, y)(1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y))\}$$

- $\ell(x^i, y^i, w)$ is called slack-rescaled hinge loss.
- Note that $\ell(x^i, y^i, w) \geq 0$, since $y = y^i$ makes it zero.

Structured SVM Learning - Slack rescaling

- Constraints: Use fixed margin

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq 1, \ \forall y \in \mathcal{Y}$$
$$\iff 1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) \leq 0, \ \forall y \in \mathcal{Y}.$$

- Penalize the constraint violation according to $\Delta(y^i, y)$.

$$\ell(x^i, y^i, w) = \max_{y \in \mathcal{Y}} \ \{\Delta(y^i, y)(1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y))\}$$

- $\ell(x^i, y^i, w)$ is called slack-rescaled hinge loss.
- Note that $\ell(x^i, y^i, w) \geq 0$, since $y = y^i$ makes it zero.
- $\ell(x^i, y^i, w)$ is a convex function of w.

Structured SVM Learning - Slack rescaling

- Constraints: Use fixed margin

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq 1, \ \forall y \in \mathcal{Y}$$
$$\iff 1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) \leq 0, \ \forall y \in \mathcal{Y}.$$

- Penalize the constraint violation according to $\Delta(y^i, y)$.

$$\ell(x^i, y^i, w) = \max_{y \in \mathcal{Y}} \ \{\Delta(y^i, y)(1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y))\}$$

- $\ell(x^i, y^i, w)$ is called slack-rescaled hinge loss.
- Note that $\ell(x^i, y^i, w) \geq 0$, since $y = y^i$ makes it zero.
- $\ell(x^i, y^i, w)$ is a convex function of w.

$$\underset{w}{\text{minimize}} \ \ \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N} \ell(x^i, y^i, w)$$

Structured SVM Learning - Optimization

$$\underset{w}{\text{minimize}} \;\; \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\ell(x^i, y^i, w)$$

Structured SVM Learning - Optimization

$$\underset{w}{\text{minimize}} \;\; \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\ell(x^i, y^i, w)$$

- $\ell(x^i, y^i, w)$ is convex but not differentiable due to the max operation.

Structured SVM Learning - Optimization

$$\underset{w}{\text{minimize}} \;\; \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\ell(x^i, y^i, w)$$

- $\ell(x^i, y^i, w)$ is convex but not differentiable due to the max operation.
- Gradient descent cannot be used.

Structured SVM Learning - Optimization

$$\underset{w}{\text{minimize}} \;\; \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\ell(x^i, y^i, w)$$

- $\ell(x^i, y^i, w)$ is convex but not differentiable due to the max operation.
- Gradient descent cannot be used.
- However, subgradient methods can be used.

Structured SVM Learning - Optimization

Margin-rescaled hinge loss:

- $\ell(x^i, y^i, w) = \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}$

Structured SVM Learning - Optimization

Margin-rescaled hinge loss:

- $\ell(x^i, y^i, w) = \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}$
- $\hat{y}^i = \operatorname{argmax}_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}$

  $= \operatorname{argmax}_{y \in \mathcal{Y}} \{\Delta(y^i, y) + w^\top \phi(x^i, y)\}$ (loss-augmented prediction)

Structured SVM Learning - Optimization

Margin-rescaled hinge loss:

- $\ell(x^i, y^i, w) = \max\limits_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}$
- $\hat{y}^i = \operatorname*{argmax}\limits_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}$

  $= \operatorname*{argmax}\limits_{y \in \mathcal{Y}} \{\Delta(y^i, y) + w^\top \phi(x^i, y)\}$ (loss-augmented prediction)
- Subgradient: $\phi(x^i, \hat{y}^i) - \phi(x^i, y^i)$

Structured SVM Learning - Optimization

Margin-rescaled hinge loss:

- $\ell(x^i, y^i, w) = \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}$
- $\hat{y}^i = \operatorname{argmax}_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}$

  $= \operatorname{argmax}_{y \in \mathcal{Y}} \{\Delta(y^i, y) + w^\top \phi(x^i, y)\}$ (loss-augmented prediction)
- Subgradient: $\phi(x^i, \hat{y}^i) - \phi(x^i, y^i)$

Slack-rescaled hinge loss:

- $\ell(x^i, y^i, w) = \max_{y \in \mathcal{Y}} \{\Delta(y^i, y)(1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y))\}$

Structured SVM Learning - Optimization

Margin-rescaled hinge loss:

- $\ell(x^i, y^i, w) = \max\limits_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}$
- $\hat{y}^i = \operatorname*{argmax}\limits_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}$

  $= \operatorname*{argmax}\limits_{y \in \mathcal{Y}} \{\Delta(y^i, y) + w^\top \phi(x^i, y)\}$ (loss-augmented prediction)
- Subgradient: $\phi(x^i, \hat{y}^i) - \phi(x^i, y^i)$

Slack-rescaled hinge loss:

- $\ell(x^i, y^i, w) = \max\limits_{y \in \mathcal{Y}} \{\Delta(y^i, y)(1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y))\}$
- $\hat{y}^i = \operatorname*{argmax}\limits_{y \in \mathcal{Y}} \{\Delta(y^i, y)(1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y))\}$

Structured SVM Learning - Optimization

Margin-rescaled hinge loss:

- $\ell(x^i, y^i, w) = \max\limits_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}$
- $\hat{y}^i = \operatorname*{argmax}\limits_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\}$

  $= \operatorname*{argmax}\limits_{y \in \mathcal{Y}} \{\Delta(y^i, y) + w^\top \phi(x^i, y)\}$ (loss-augmented prediction)
- Subgradient: $\phi(x^i, \hat{y}^i) - \phi(x^i, y^i)$

Slack-rescaled hinge loss:

- $\ell(x^i, y^i, w) = \max\limits_{y \in \mathcal{Y}} \{\Delta(y^i, y)(1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y))\}$
- $\hat{y}^i = \operatorname*{argmax}\limits_{y \in \mathcal{Y}} \{\Delta(y^i, y)(1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y))\}$
- Subgradient: $\Delta(y^i, \hat{y}^i)(\phi(x^i, \hat{y}^i) - \phi(x^i, y^i))$

Structured SVM - Formulation with slack variables

$$\underset{w}{\text{minimize}} \;\; \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\ell(x^i, y^i, w) \tag{1}$$

Structured SVM - Formulation with slack variables

$$\underset{w}{\text{minimize}} \;\; \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N} \ell(x^i, y^i, w) \tag{1}$$

$$\begin{aligned} &\underset{w,\zeta}{\text{minimize}} \;\; \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N} \zeta^i \\ &\text{subject to} \;\; \ell(x^i, y^i, w) \leq \zeta^i, \; \text{for } i = 1, 2, \ldots, N. \end{aligned} \tag{2}$$

Optimization problems (1) and (2) are equivalent.

Structured SVM - Formulation with slack variables

Margin-rescaled structured SVM:

$$\ell(x^i, y^i, w) \leq \zeta^i$$

Structured SVM - Formulation with slack variables

Margin-rescaled structured SVM:

$$\ell(x^i, y^i, w) \le \zeta^i$$

$$\Longleftrightarrow \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\} \le \zeta^i$$

Structured SVM - Formulation with slack variables

Margin-rescaled structured SVM:

$$\begin{aligned}
& \ell(x^i, y^i, w) \leq \zeta^i \\
\iff & \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\} \leq \zeta^i \\
\iff & \Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) - \zeta^i \leq 0, \; \forall y \in \mathcal{Y}
\end{aligned}$$

Structured SVM - Formulation with slack variables

Margin-rescaled structured SVM:

$$
\begin{aligned}
& \ell(x^i, y^i, w) \leq \zeta^i \\
\iff & \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\} \leq \zeta^i \\
\iff & \Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) - \zeta^i \leq 0, \ \forall y \in \mathcal{Y}
\end{aligned}
$$

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N} \zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) - \zeta^i \leq 0, \ \forall y \in \mathcal{Y}$$

$$\zeta^i \geq 0$$

Structured SVM - Formulation with slack variables

Margin-rescaled structured SVM:

$$
\begin{aligned}
& \ell(x^i, y^i, w) \leq \zeta^i \\
\iff & \max_{y \in \mathcal{Y}} \{\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y)\} \leq \zeta^i \\
\iff & \Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) - \zeta^i \leq 0, \ \forall y \in \mathcal{Y}
\end{aligned}
$$

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N} \zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) - \zeta^i \leq 0, \ \forall y \in \mathcal{Y}$$

$$\zeta^i \geq 0$$

- QP with number of constraints proportional to $|\mathcal{Y}|$.

Structured SVM - Formulation with slack variables

Slack-rescaled structured SVM:

$$\ell(x^i, y^i, w) \leq \zeta^i$$

Structured SVM - Formulation with slack variables

Slack-rescaled structured SVM:

$$
\begin{aligned}
& \ell(x^i, y^i, w) \leq \zeta^i \\
\iff & \max_{y \in \mathcal{Y}} \{\Delta(y^i, y)(1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y))\} \leq \zeta^i
\end{aligned}
$$

Structured SVM - Formulation with slack variables

Slack-rescaled structured SVM:

$$
\begin{aligned}
& \ell(x^i, y^i, w) \le \zeta^i \\
\iff & \max_{y \in \mathcal{Y}} \{\Delta(y^i, y)(1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y))\} \le \zeta^i \\
\iff & 1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) - \frac{\zeta^i}{\Delta(y^i, y)} \le 0, \ \forall y \in \mathcal{Y} \setminus y^i
\end{aligned}
$$

Structured SVM - Formulation with slack variables

Slack-rescaled structured SVM:

$$\ell(x^i, y^i, w) \le \zeta^i$$

$$\iff \max_{y \in \mathcal{Y}} \{\Delta(y^i, y)(1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y))\} \le \zeta^i$$

$$\iff 1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) - \frac{\zeta^i}{\Delta(y^i, y)} \le 0, \ \forall y \in \mathcal{Y} \setminus y^i$$

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N} \zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) - \frac{\zeta^i}{\Delta(y^i, y)} \le 0, \ \forall y \in \mathcal{Y} \setminus y^i$$

$$\zeta^i \ge 0$$

Structured SVM - Formulation with slack variables

Slack-rescaled structured SVM:

$$
\begin{aligned}
& \ell(x^i, y^i, w) \leq \zeta^i \\
\iff & \max_{y \in \mathcal{Y}} \{\Delta(y^i, y)(1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y))\} \leq \zeta^i \\
\iff & 1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) - \frac{\zeta^i}{\Delta(y^i, y)} \leq 0, \ \forall y \in \mathcal{Y} \setminus y^i
\end{aligned}
$$

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N} \zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) - \frac{\zeta^i}{\Delta(y^i, y)} \leq 0, \ \forall y \in \mathcal{Y} \setminus y^i$$

$$\zeta^i \geq 0$$

- QP with number of constraints proportional to $|\mathcal{Y}|$.

Formulation with slack variables - Optimization

- Number of constraints is very large (proportional to $|\mathcal{Y}|$).

Formulation with slack variables - Optimization

- Number of constraints is very large (proportional to $|\mathcal{Y}|$).
- Difficult to solve the entire QP.

Formulation with slack variables - Optimization

- Number of constraints is very large (proportional to $|\mathcal{Y}|$).
- Difficult to solve the entire QP.
- Cutting plane method can be used.

Formulation with slack variables - Optimization

- Number of constraints is very large (proportional to $|\mathcal{Y}|$).
- Difficult to solve the entire QP.
- Cutting plane method can be used.
- Start with out any constraints and iteratively add the most violated constraint for each sample.

Formulation with slack variables - Optimization

- Number of constraints is very large (proportional to $|\mathcal{Y}|$).
- Difficult to solve the entire QP.
- Cutting plane method can be used.
- Start with out any constraints and iteratively add the most violated constraint for each sample.
- A smaller QP solved in each iteration.

Formulation with slack variables - Optimization

- Number of constraints is very large (proportional to $|\mathcal{Y}|$).
- Difficult to solve the entire QP.
- Cutting plane method can be used.
- Start with out any constraints and iteratively add the most violated constraint for each sample.
- A smaller QP solved in each iteration.
- Objective function will be ϵ-close to the global minimum after $O(\frac{1}{\epsilon^2})$ iterations.

Formulation with single slack variable

Margin-rescaled (formulation with multiple slack variables):

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) - \zeta^i \leq 0, \ \forall y \in \mathcal{Y}$$

$$\zeta^i \geq 0$$

Formulation with single slack variable

Margin-rescaled (formulation with multiple slack variables):

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$\Delta(y^i, y) - w^\top\phi(x^i, y^i) + w^\top\phi(x^i, y) - \zeta^i \leq 0, \ \forall y \in \mathcal{Y}$$

$$\zeta^i \geq 0$$

Margin-rescaled (formulation with single slack variable):

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + C\zeta$$

subject to, for all $(y^{(1)}, \ldots, y^{(n)}) \in \mathcal{Y} \times \ldots \times \mathcal{Y}$,

$$\frac{1}{N}\sum_{i=1}^{N}\Big(\Delta(y^i, y^{(i)}) - w^\top\phi(x^i, y^i) + w^\top\phi(x^i, y^{(i)})\Big) - \zeta \leq 0$$

$$\zeta \geq 0$$

Formulation with single slack variable

Slack-rescaled (formulation with multiple slack variables):

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) - \frac{\zeta^i}{\Delta(y^i, y)} \leq 0, \ \forall y \in \mathcal{Y} \setminus y^i$$

$$\zeta^i \geq 0$$

Formulation with single slack variable

Slack-rescaled (formulation with multiple slack variables):

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) - \frac{\zeta^i}{\Delta(y^i, y)} \leq 0, \ \forall y \in \mathcal{Y} \setminus y^i$$

$$\zeta^i \geq 0$$

Slack-rescaled (formulation with single slack variable):

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + C\zeta$$

subject to, for all $(y^{(1)}, \ldots, y^{(n)}) \in \mathcal{Y} \times \ldots \times \mathcal{Y}$,

$$\frac{1}{N}\sum_{i=1}^{N}\left(\Delta(y^i, y^{(i)})\left(1 - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y^{(i)})\right)\right) - \zeta \leq 0$$

$$\zeta \geq 0$$

Formulation with single slack variable - Optimization

- Number of constraints is huge (proportional to $|\mathcal{Y}|^N$).

Formulation with single slack variable - Optimization

- Number of constraints is huge (proportional to $|\mathcal{Y}|^N$).
- Difficult to solve the entire QP.

Formulation with single slack variable - Optimization

- Number of constraints is huge (proportional to $|\mathcal{Y}|^N$).
- Difficult to solve the entire QP.
- Cutting plane method can be used.

Formulation with single slack variable - Optimization

- Number of constraints is huge (proportional to $|\mathcal{Y}|^N$).
- Difficult to solve the entire QP.
- Cutting plane method can be used.
- Start with out any constraints and iteratively add the most violated constraint.

Formulation with single slack variable - Optimization

- Number of constraints is huge (proportional to $|\mathcal{Y}|^N$).
- Difficult to solve the entire QP.
- Cutting plane method can be used.
- Start with out any constraints and iteratively add the most violated constraint.
- A smaller QP solved in each iteration.

Formulation with single slack variable - Optimization

- Number of constraints is huge (proportional to $|\mathcal{Y}|^N$).
- Difficult to solve the entire QP.
- Cutting plane method can be used.
- Start with out any constraints and iteratively add the most violated constraint.
- A smaller QP solved in each iteration.
- Objective function will be ϵ-close to the global minimum after $O(\frac{1}{\epsilon})$ iterations.

Structured SVM

Training:

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) - \zeta^i \leq 0, \ \forall y \in \mathcal{Y}$$

$$\zeta^i \geq 0$$

Structured SVM

Training:

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$\Delta(y^i, y) - w^\top\phi(x^i, y^i) + w^\top\phi(x^i, y) - \zeta^i \leq 0, \ \forall y \in \mathcal{Y}$$

$$\zeta^i \geq 0$$

Prediction:

$$f(x) = \underset{y\in\mathcal{Y}}{\text{argmax}} \ w^\top\phi(x, y)$$

Structured SVM

Training:

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$\Delta(y^i, y) - w^\top \phi(x^i, y^i) + w^\top \phi(x^i, y) - \zeta^i \leq 0, \ \forall y \in \mathcal{Y}$$

$$\zeta^i \geq 0$$

Prediction:

$$f(x) = \underset{y \in \mathcal{Y}}{\text{argmax}} \ w^\top \phi(x, y)$$

- Output space $\mathcal{Y}$
- Feature map $\phi(x, y)$
- Loss function $\Delta(y, y^{'})$

Structured SVM - Multi-Class Classification

- $\mathcal{Y}$ = set of class labels

Structured SVM - Multi-Class Classification

- $\mathcal{Y}$ = set of class labels
- Let $\phi_x(x^i)$ be a feature map defined over $\mathcal{X}$.

Structured SVM - Multi-Class Classification

- $\mathcal{Y}$ = set of class labels
- Let $\phi_x(x^i)$ be a feature map defined over $\mathcal{X}$.
- Let $\phi_y(y^i)$ be defined as the vector with 1 in the place of current class and 0 elsewhere.

$$\phi_y(y^i) = [0, \ldots, 1, \ldots, 0]^\top$$

Structured SVM - Multi-Class Classification

- $\mathcal{Y}$ = set of class labels
- Let $\phi_x(x^i)$ be a feature map defined over $\mathcal{X}$.
- Let $\phi_y(y^i)$ be defined as the vector with 1 in the place of current class and 0 elsewhere.

$$\phi_y(y^i) = [0, \ldots, 1, \ldots, 0]^\top$$

- Let $\phi(x^i, y^i) = \phi_y(y^i) \otimes \phi_x(x^i)$
$$= [0, \ldots, 0 \mid \ldots \mid \phi_x(x^i) \mid \ldots \mid 0, \ldots, 0 \mid]$$

Structured SVM - Multi-Class Classification

- $\mathcal{Y}$ = set of class labels
- Let $\phi_x(x^i)$ be a feature map defined over $\mathcal{X}$.
- Let $\phi_y(y^i)$ be defined as the vector with 1 in the place of current class and 0 elsewhere.

$$\phi_y(y^i) = [0, \ldots, 1, \ldots, 0]^\top$$

- Let $\phi(x^i, y^i) = \phi_y(y^i) \otimes \phi_x(x^i)$
  $= [0, \ldots, 0 \mid \ldots \mid \phi_x(x^i) \mid \ldots \mid 0, \ldots, 0 \,|]$
- Let $w = [w_1^\top, w_2^\top, \ldots, w_k^\top]^\top$.

Structured SVM - Multi-Class Classification

- $\mathcal{Y}$ = set of class labels
- Let $\phi_x(x^i)$ be a feature map defined over $\mathcal{X}$.
- Let $\phi_y(y^i)$ be defined as the vector with 1 in the place of current class and 0 elsewhere.

$$\phi_y(y^i) = [0, \ldots, 1, \ldots, 0]^\top$$

- Let $\phi(x^i, y^i) = \phi_y(y^i) \otimes \phi_x(x^i)$
$$= [0, \ldots, 0 \mid \ \ldots \ \mid \phi_x(x^i) \mid \ \ldots \ \mid 0, \ldots, 0 \mid]$$
- Let $w = [w_1^\top, w_2^\top, \ldots, w_k^\top]^\top$.
- Zero-one loss: $\Delta(y, y^{'}) = \mathbb{1}[y \neq y^{'}]$

Structured SVM - Multi-Class Classification

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq \Delta(y^i, y) - \zeta^i, \ \forall y \in \mathcal{Y}$$

$$\zeta^i \geq 0$$

Structured SVM - Multi-Class Classification

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq \Delta(y^i, y) - \zeta^i, \ \forall y \in \mathcal{Y}$$

$$\zeta^i \geq 0$$

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$w_{y^i}^\top \phi_x(x^i) - w_y^\top \phi_x(x^i) \geq 1 - \zeta^i, \ \forall y \in \mathcal{Y} \setminus y^i$$

$$\zeta^i \geq 0$$

- Similar to the Cramer and Signer multiclass formulation.

Structured SVM - Object Localization

- $\mathcal{X}$ = images

Structured SVM - Object Localization

- $\mathcal{X}$ = images
- $\mathcal{Y}$ = set of bounding boxes

Structured SVM - Object Localization

- $\mathcal{X}$ = images
- $\mathcal{Y}$ = set of bounding boxes
- Let ϕ_x denote an image feature map.

Structured SVM - Object Localization

- $\mathcal{X}$ = images
- $\mathcal{Y}$ = set of bounding boxes
- Let ϕ_x denote an image feature map.
- Let $x|_y$ denote the region in image x given by the bounding box y

Structured SVM - Object Localization

- $\mathcal{X}$ = images
- $\mathcal{Y}$ = set of bounding boxes
- Let ϕ_x denote an image feature map.
- Let $x|_y$ denote the region in image x given by the bounding box y
- Let $\phi(x, y) = \phi_x(x|_y)$

Structured SVM - Object Localization

- $\mathcal{X}$ = images
- $\mathcal{Y}$ = set of bounding boxes
- Let ϕ_x denote an image feature map.
- Let $x|_y$ denote the region in image x given by the bounding box y
- Let $\phi(x, y) = \phi_x(x|_y)$
- Area overlap loss:

$$\Delta(y, y^{'}) = 1 - \frac{area(y \cap y^{'})}{area(y \cup y^{'})}$$

Structured SVM - Object Localization

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq \Delta(y^i, y) - \zeta^i, \; \forall y \in \mathcal{Y}$$

$$\zeta^i \geq 0$$

Structured SVM - Object Localization

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$w^\top \phi(x^i, y^i) - w^\top \phi(x^i, y) \geq \Delta(y^i, y) - \zeta^i, \ \forall y \in \mathcal{Y}$$

$$\zeta^i \geq 0$$

$$\underset{w,\zeta}{\text{minimize}} \quad \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\zeta^i$$

subject to, for $i = 1, 2, \ldots, N$

$$w^\top \phi_x(x^i|_{y^i}) - w^\top \phi_x(x^i|_y) \geq 1 - \frac{area(y \cap y')}{area(y \cup y')} - \zeta^i, \ \forall y \in \mathcal{Y}$$

$$\zeta^i \geq 0$$

Questions ??