Structured Linear Model

Hung-yi Lee
Structured Linear Model

Problem 1: Evaluation
• What does $F(x, y)$ look like? in a specific form

Problem 2: Inference
• How to solve the “arg max” problem

\[ y = \arg \max_{y \in Y} F(x, y) \]

Problem 3: Training
• Given training data, how to find $F(x, y)$
Structured Linear Model: Problem 1

- Evaluation: What does $F(x,y)$ look like?

Characteristics

$$F(x, y) = w_1 \cdot \phi_1(x, y) + w_2 \cdot \phi_2(x, y) + w_3 \cdot \phi_3(x, y) + \ldots$$

Learning from data
Structured Linear Model: Problem 1

• Evaluation: What does $F(x,y)$ look like?
• Example: **Object Detection**

\[
\phi(x, y) = \text{percentage of color red in box } y \\
\text{percentage of color green in box } y \\
\text{percentage of color blue in box } y \\
\text{percentage of color red out of box } y \\
\text{area of box } y \\
\text{number of specific patterns in box } y \\
\text{......}
\]
\[ \phi(x, y) \]

- Convolutional Layer
- Sub-sampling Layer
- Fully-connected Layer
- Output Layer
Structured Linear Model: Problem 1

• Evaluation: What does $F(x,y)$ look like?

• Example: \textit{Summarization}

$\phi_1(x, y)$

Whether the sentence containing the word "important" is in $y$

$\phi_2(x, y)$

Whether the sentence containing the word "definition" is in $y$

$\phi_3(x, y)$

Length of $y$

$\phi_4(x, y)$

How succinct is $y$?

$\phi_5(x, y)$

How representative of $y$?

(a long document)

(Short paragraph)
Structured Linear Model: Problem 1

- **Evaluation**: What does $F(x,y)$ look like?
- **Example**: Retrieval

The degree of relevance with respect to $x$ for the top 1 webpages in $y$.

Is the top 1 webpage more relevant than the top 2 webpage?

How much different information does $y$ cover? *(Diversity)*
Structured Linear Model: Problem 2

• **Inference**: How to solve the “arg max” problem

\[
y = \arg \max_{y \in Y} F(x, y)
\]

\[
F(x, y) = w \cdot \phi(x, y) \quad \Rightarrow \quad y = \arg \max_{y \in Y} w \cdot \phi(x, y)
\]

• Assume we have solved this question.
Structured Linear Model: Problem 3

• Training: Given training data, how to learn $F(x, y)$
  • $F(x, y) = w \cdot \phi(x, y)$, so what we have to learn is $w$

Training data: $\{(x^1, \hat{y}^1), (x^2, \hat{y}^2), \ldots, (x^r, \hat{y}^r), \ldots\}$

We should find $w$ such that

$$\forall r \text{ (All training examples)}$$

$$\forall y \in Y - \{\hat{y}^r\} \text{ (All incorrect label for r-th example)}$$

$$w \cdot \phi(x^r, \hat{y}^r) > w \cdot \phi(x^r, y)$$
Structured Linear Model:
Problem 3

\[ \hat{y}^1, \hat{y}^2 \]

\[ \phi(x^1, \hat{y}^1) \]
\[ \phi(x^1, y) \]

\[ \phi(x^2, \hat{y}^2) \]

\[ \phi(x^2, y) \]
Structured Linear Model: Problem 3

\[
\begin{align*}
\hat{y}^1 & = \phi(x^1, \hat{y}^1) \\
\hat{y}^2 & = \phi(x^2, \hat{y}^2) \\
\phi(x^1, y) & \\
\phi(x^2, y) &
\end{align*}
\]
Structured Linear Model:
Problem 3

\[ \mathbf{w} \mathbf{x} + \mathbf{y} \geq 1 \]

\[ \hat{y}^1 \]

\[ \hat{y}^2 \]

\[ \phi(x^1, \hat{y}^1) \]

\[ \phi(x^1, y) \]

\[ \phi(x^2, \hat{y}^2) \]

\[ \phi(x^2, y) \]

\[ w \cdot \phi(x^1, \hat{y}^1) \geq w \cdot \phi(x^1, y) \]

\[ w \cdot \phi(x^2, \hat{y}^2) \geq w \cdot \phi(x^2, y) \]
Solution of Problem 3

Difficult?

Not as difficult as expected
Algorithm

- **Input**: training data set \( \{(x^1, \hat{y}^1), (x^2, \hat{y}^2), \ldots, (x^r, \hat{y}^r), \ldots\} \)
- **Output**: weight vector \( w \)
- **Algorithm**: Initialize \( w = 0 \)
  
  - do
    - For each pair of training example \( (x^r, \hat{y}^r) \)
      - Find the label \( \tilde{y}^r \) maximizing \( w \cdot \phi(x^r, y) \)
        \[
        \tilde{y}^r = \arg \max_{y \in Y} w \cdot \phi(x^r, y) \quad \text{(question 2)}
        \]
      - If \( \tilde{y}^r \neq \hat{y}^r \), update \( w \)
        \[
        w \rightarrow w + \phi(x^r, \hat{y}^r) - \phi(x^r, \tilde{y}^r)
        \]
  - until \( w \) is not updated  ➡️ **We are done!**
Algorithm - Example
Algorithm - Example

Initialize $w = 0$

pick $(x^1, \hat{y}^1)$

$\tilde{y}^1 = \arg \max_{y \in Y} w \cdot \phi(x^1, y)$

If $\tilde{y}^1 \neq \hat{y}^1$, update $w$

$$w \rightarrow w + \phi(x^1, \hat{y}^1) - \phi(x^1, \tilde{y}^1)$$

Because $w=0$ at this time, $\phi(x^1, y)$ always 0

Random pick one point as $\tilde{y}^r$
Algorithm - Example

pick \((x^2, \hat{y}^2)\)

\[\tilde{y}^2 = \arg \max_{y \in Y} w \cdot \phi(x^2, y)\]

If \(\tilde{y}^2 \neq \hat{y}^2\), update \(w\)

\[w \rightarrow w + \phi(x^2, \hat{y}^2) - \phi(x^2, \tilde{y}^2)\]
Algorithm - Example

pick \((x^1, \hat{y}^1)\) again

\[\tilde{y}^1 = \arg \max_{y \in Y} w \cdot \phi(x^1, y)\]

\[\tilde{y}^1 = \hat{y}^1 \quad \Rightarrow \text{do not update } w\]

pick \((x^2, \hat{y}^2)\) again

\[\tilde{y}^2 = \arg \max_{y \in Y} w \cdot \phi(x^2, y)\]

\[\tilde{y}^2 = \hat{y}^2 \quad \Rightarrow \text{do not update } w\]

\[\phi(x^1, \hat{y}^1) \geq \phi(x^1, y)\]

\[w \cdot \phi(x^2, \hat{y}^2) \geq w \cdot \phi(x^2, y)\]

So we are done
Assumption: Separable

- There exists a weight vector $\hat{w}$ such that $\|\hat{w}\| = 1$

$\forall r$ (All training examples)

$\forall y \in Y - \{\hat{y}^r\}$ (All incorrect label for an example)

$\hat{w} \cdot \phi(x^r, \hat{y}^r) \geq \hat{w} \cdot \phi(x^r, y)$ (The target exists)

$\hat{w} \cdot \phi(x^r, \hat{y}^r) \geq \hat{w} \cdot \phi(x^r, y) + \delta$
Assumption: Separable

\[ \hat{w} \cdot \phi(x^r, \hat{y}^r) \geq \hat{w} \cdot \phi(x^r, y) + \delta \]
Proof of Termination

w is updated once it sees a mistake

\[ w^0 = 0 \rightarrow w^1 \rightarrow w^2 \rightarrow \ldots \ldots \rightarrow w^k \rightarrow w^{k+1} \rightarrow \ldots \]

\[ w^k = w^{k-1} + \phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n) \]

(the relation of \( w^k \) and \( w^{k-1} \))

Proof that: The angle \( \rho_k \) between \( \hat{w} \) and \( w_k \) is smaller as \( k \) increases

Analysis \( \cos \rho_k \) (larger and larger?) \[ \cos \rho_k = \frac{\hat{w} \cdot w^k}{\|\hat{w}\| \cdot \|w^k\|} \]

\[ \hat{w} \cdot w^k = \hat{w} \cdot (w^{k-1} + \phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)) \]

\[ = \hat{w} \cdot w^{k-1} + \hat{w} \cdot \phi(x^n, \hat{y}^n) - \hat{w} \cdot \phi(x^n, \tilde{y}^n) \geq \hat{w} \cdot w^{k-1} + \delta \]

\[ \geq \delta \] (Separable)
Proof of Termination

w is updated once it sees a mistake

\[ w^0 = 0 \rightarrow w^1 \rightarrow w^2 \rightarrow \ldots \ldots \rightarrow w^k \rightarrow w^{k+1} \rightarrow \ldots \ldots \]

\[ w^k = w^{k-1} + \phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n) \] (the relation of \( w^k \) and \( w^{k-1} \))

Proof that: The angle \( \rho_k \) between \( \hat{w} \) and \( w_k \) is smaller as \( k \) increases

Analysis \( \cos \rho_k \) (larger and larger?) \[ \cos \rho_k = \frac{\hat{w} \cdot w^k}{\|\hat{w}\| \|w^k\|} \]

\[ \hat{w} \cdot w^k \geq \hat{w} \cdot w^{k-1} + \delta \]

\[ = 0 \geq \delta \]

\[ \hat{w} \cdot w^1 \geq \hat{w} \cdot w^0 + \delta \]

\[ \hat{w} \cdot w^2 \geq \hat{w} \cdot w^1 + \delta \]

\[ \hat{w} \cdot w^1 \geq \delta \]

\[ \hat{w} \cdot w^2 \geq 2\delta \]

\[ \ldots \ldots \]

\[ \hat{w} \cdot w^k \geq k\delta \]

(so what)
Proof of Termination

\[
\cos \rho_k = \frac{\hat{w}}{\|\hat{w}\|} \cdot \frac{w^k}{\|w^k\|}
\]

\[
w^k = w^{k-1} + \phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)
\]

\[
\|w^k\|^2 = \|w^{k-1} + \phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)\|^2
\]

\[
= \|w^{k-1}\|^2 + \|\phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)\|^2 + 2w^{k-1} \cdot (\phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n))
\]

\[
> 0
\]

Assume the distance between any two feature vector is smaller than \(R\)

\[
\leq \|w^{k-1}\| + R^2
\]
Proof of Termination

\[
\cos \rho_k = \frac{\hat{w} \cdot w^k}{\|\hat{w}\| \cdot \|w^k\|} \quad \hat{w} \cdot w^k \geq k\delta \quad \|w^k\|^2 \leq kR^2
\]

\[
\geq \frac{k\delta}{\sqrt{kR^2}} = \sqrt{k} \frac{\delta}{R}
\]

\[
\sqrt{k} \frac{\delta}{R} \leq 1
\]

\[
k \leq \left( \frac{R}{\delta} \right)^2
\]
Proof of Termination

\[ k \leq \left( \frac{R}{\delta} \right)^2 \]

The largest distances between features

Margin: Is it easy to separable red points from the blue ones

Normalization

Larger margin, less update

\[ \phi(x^r, \hat{y}^r) \]
\[ \phi(x^r, y) \]

All feature times 2

\[ \hat{w} \]
\[ \delta \]
\[ \delta \uparrow \]
\[ R \uparrow \]
Structured Linear Model: Reduce 3 Problems to 2

Problem 1: Evaluation
- How to define $F(x,y)$

Problem 2: Inference
- How to find the $y$ with the largest $F(x,y)$

Problem 3: Training
- How to learn $F(x,y)$

Problem A: Feature
- How to define $\phi(x,y)$

$F(x,y) = w \cdot \phi(x,y)$

Problem B: Inference
- How to find the $y$ with the largest $w \cdot \phi(x,y)$