Lecture 7

Kernel Machines - Structured Prediction

~~BAACE~~

$\boxed{\text{Brace}}$ → $\boxed{\text{BRACE}}$

Input

Output: must be a word!

→ only allow certain structured outputs

Gene sequences:
GAGTCA
→ known protein structures

$\boxed{\text{The cat sat on the mat}}$ → only allow grammatically correct sentences

Kernel Input can be structured → previous lecture
e.g. weighted degree kernels

Kernel Based Structured Prediction:

$$k : (X \times Y) \times (X \times Y) \to \mathbb{R}$$

PSD + symm.: Feature map $\phi : X \times Y \to \mathbb{R}^k$

Model: $f(y|x) = w^T \phi(x, y)$

Multi-Class Classification

$k((x,y),(x',y')) = \langle \phi(x,y), \phi(x',y') \rangle$ ← General

$X \in \mathbb{R}^d$, $Y = \{1, \dots, C\}$
input (e.g. image), output (e.g. "dog")

$K\Big((x,y),(x',y')\Big) = k(x,x') \cdot \mathbb{1}(y = y')$ ← for multiclass

Feature Map: $\Psi(x,y) = \begin{pmatrix} \phi(x) \cdot \mathbb{1}(y=1) \\ \vdots \\ \phi(x) \ \mathbb{1}(y=c) \end{pmatrix}$

$$y|x = \underset{y}{\operatorname{argmax}} \langle w, \Psi(x,y) \rangle = \underset{c}{\operatorname{argmax}} \, w_c^T \phi(x)$$

How to learn such a structured output model?

Model: $f(y|x) = w^T \phi(x, y)$ Prediction $y|x = \underset{y \in Y}{\arg\max} \, f(y|x)$

Find largest-margin model by solving

$$\min_{W} \frac{1}{2}\|W\|^2 + C \sum_{n=1}^{N} \xi_n$$

Constraints: $\forall_{n=1}^{N}, \forall_{y \in y_n}$: $\underbrace{w^T \phi(x_n, y_n)}_{A} - \underbrace{w^T \phi(x_n, y)}_{B} \geq 1 - \xi_n$

y_n: ground truth y: all possible evaluation

A: Score of ground truth   B: score of all possible evaluation

A - B: difference of scores, $A - B \geq 1$ $(-\xi_n)$

ξ_n: Slack term: accounts for wrong labels, noise, ..

$\min_{w} \frac{1}{2}\|w\|^2$ → maximize margin

larger C → smaller ξ_n → hard margin constraint

Structured Prediction vs. HMM

HMM Pro?: - unsupervised

- once model learned, procedure for prediction determined

Struct. Pred. Pro?

- feature map and loss function give more flexibility for structure of problem
- model parameter $w \in \mathbb{R}^D$ can be actively optimized for best performance on supervised task

Summary

- structured output learning enables prediction of structured outputs (trees, sequences)
- assign matching scores to input-output pairs
- supervised - kernel-based framework
- difficulty: efficiently infer which output $y \in Y$ maximizes the score $f(y|x)$