Forecasting and Simulation **Summer Term 2024**

`http://ml3.leuphana.de/lectures/summer24/FS`
Machine Learning Group, Leuphana University of Lüneburg

Soham Majumder (soham.majumder@leuphana.de)

Exercise 6: Hidden Markov Support Vector Machines

1. Hidden Markov Perceptron Learning

You have been reading [Altun et al., 2003] and are currently discussing it in the lectures.

1. In Section 4 (Hidden Markov Perceptron Learning) of the paper it is stated how

$$F(\mathbf{x}, \mathbf{y}) = F_1(\mathbf{x}, \mathbf{y}) + F_2(\mathbf{x}, \mathbf{y}),$$

   where we refer to the paper for notation. Please proof the equality explicitly.

2. Please show, why we have

$$\langle \Phi(\mathbf{x}, \mathbf{y}), \Phi(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \rangle = \sum_{s,t} [\![y^{s-1} = \bar{y}^{t-1} \wedge y^s = \bar{y}]\!] + \sum_{s,t} [\![y^s = \bar{y}^t]\!] k(x^s, \bar{x}^t)$$

   in Equation (7) in [Altun et al., 2003]. Note: while more general features are discussed in Section 2 of the paper, Altun et al. restrict the features to “non-overlapping” label-observation features, meaning features of the form $\phi^{tt}_{r\sigma}$. They further state that they restrict themselves to first-order label-label features, i.e. to features of the form $\bar{\phi}^{t(t+1)}_{\sigma\tau}$.

3. Please explain the shape of dual parameters $\alpha_i(\bar{\mathbf{y}})$ in [Altun et al., 2003]. How would you store them?

Solution

1. refer to Ex 6 solutions.pdf
2. refer to Ex 6 solutions.pdf
3. Discussion in class:
   - $\alpha_i(\bar{\mathbf{y}})$ is a scalar
   - We will need an α_i for each training instance $\mathbf{x}_i$ keeping track of observed $\bar{\mathbf{y}} \in \mathcal{Y}(\mathbf{x}_i)$, where $\mathcal{Y}(\mathbf{x}_i)$ is the set of all possible label-sequences for $\mathbf{x}_i$
   - Idea: for each $\mathbf{x}_i$ keep updating a dict: `dict_i = {y_ij: alpha_ij, ...}`

2. HM-SVM

i) Please be able to describe the working set optimization for HM-SVMs (Algorithm 2 in [Altun et al., 2003]) in your own words. Make sure that you understand the notation.

ii) Derive the dual of the soft margin HM-SVM with L_2 penalties in Equation 18 of [Altun et al., 2003].

iii) Show why the penalty term (Equation 20 of [Altun et al., 2003]) can be absorbed into the kernel (Equation 21).

Solution

1. **Initialization and the Role of the Working Set**

   The working set $\mathcal{W}$ initially contains only the correct output sequences for each training example. This means that, at the very start, the algorithm knows only about the correct labels and uses them to start training the model.

   Example: Part-of-Speech Tagging

   Consider a simple example where we have a training dataset with two sentences:

   (a) "The cat sits"
      - Correct tags: (DET, NOUN, VERB)

   (b) "On the mat"
      - Correct tags: (PREP, DET, NOUN)

   Step-by-Step Explanation

   Initialization

   (a) **Weight Vector w Initialization**:
      - Initialize the weight vector $\mathbf{w} = 0$.

   (b) **Working Set $\mathcal{W}$ Initialization**:
      - Initialize the working set $\mathcal{W}$ to contain the correct output sequences.
      - For our example, $\mathcal{W} = \{(\text{DET, NOUN, VERB}), (\text{PREP, DET, NOUN})\}$.

   Iteration 1

   Step 1: Solve the Restricted Optimization Problem

   - Since $\mathbf{w} = 0$ and $\mathcal{W}$ contains only the correct sequences, solving the optimization problem for the current working set means finding a weight vector that maximizes the margin between these correct sequences and any potential incorrect sequences.

   Step 2: Find the Most Violated Constraint

   - For each sentence in the training data, the algorithm computes the scores for all possible sequences (not just the correct ones).
   - The sequence that most violates the margin constraint (i.e., the sequence that has the highest score minus its loss) is identified.

Consider the first sentence "The cat sits":

- Possible sequences and their initial scores (since $\mathbf{w} = 0$, all scores are initially zero):
  (a) (DET, NOUN, VERB)
  (b) (NOUN, VERB, DET)
  (c) (DET, VERB, NOUN)
  (d) (VERB, NOUN, DET)
  (e) etc.
- The correct sequence is (DET, NOUN, VERB). Let's assume the sequence (NOUN, VERB, DET) has the highest score after the initial model update (though initially, the scores are the same, future updates will differentiate them).

Step 3: Update the Working Set

- If the most violated sequence (e.g., (NOUN, VERB, DET)) is sufficiently different from the correct sequence and violates the margin constraint, it is added to the working set $\mathcal{W}$.
- $\mathcal{W} = \{(\text{DET, NOUN, VERB}), (\text{NOUN, VERB, DET}), (\text{PREP, DET, NOUN})\}$.

Iteration 2 and Beyond

(a) **Solve the Restricted Optimization Problem**:
  - With the updated working set, solve the optimization problem again to adjust the weight vector $\mathbf{w}$ so that the margin is maximized considering the new set of constraints.

(b) **Find the Most Violated Constraint**:
  - For each training example, identify the sequence that most violates the margin constraint.
  - This process will likely identify new incorrect sequences as the model starts to differentiate more finely between correct and incorrect sequences.

(c) **Update the Working Set**:
  - Add any newly identified most violated sequences to the working set.
  - For example, for the sentence "On the mat", an incorrect sequence (e.g., (DET, NOUN, VERB)) might be added if it significantly violates the margin constraint.

Key Points and Intuition

- **Initial Working Set**: Starting with only the correct sequences means the model begins by knowing what the correct answers are. However, it doesn't yet know what mistakes to avoid because the working set is too limited.
- **Finding Violations**: By identifying the most violated constraints, the model learns from its mistakes. Each iteration exposes the model to new incorrect sequences that are close to being classified as correct.
- **Iterative Refinement**: The model iteratively refines its weight vector by focusing on sequences that it struggles with the most, progressively improving its ability to differentiate between correct and incorrect sequences.

Why Incorrect Sequences Are Added

If the working set contained only the correct sequences, the model wouldn't learn to discriminate between correct and incorrect outputs. Adding the most violated incorrect sequences forces the model to adjust its weights such that the correct sequences are more strongly preferred, thereby improving generalization.

Concrete Example Continued

- Suppose, after several iterations, the model has updated the weight vector $\mathbf{w}$ and the working set $\mathcal{W}$ includes a variety of sequences.
- For the sentence "The cat sits", $\mathcal{W}$ might now contain sequences like (DET, NOUN, VERB), (NOUN, VERB, DET), (DET, VERB, NOUN), etc.
- The model must ensure that the correct sequence (DET, NOUN, VERB) scores higher than any of the incorrect sequences in the working set, leading to a weight vector that properly captures the dependencies in POS tagging.

Final Convergence

The algorithm stops when no significant violations are found, indicating that the model's weight vector $\mathbf{w}$ is sufficiently trained to distinguish between correct and incorrect sequences for the given task.

2. refer to Ex 6 solutions.pdf
3. refer to Ex 6 solutions.pdf

3. Dual perceptron: a toy example

The notation used in this exercise is inspired by the notation in [Altun et al., 2003].

In this Task, we will look at the relation of primal and dual parameters of the dual perceptron for a toy example, where hidden variables are sequences of decisions of eating or not eating at the Mensa. Consider a single input sequence $\mathbf{x} = (x^1, x^2, x^3)$ comprising the set of ingredients available at Mensa on each day during three consecutive days, such that: $x^1 = \{\texttt{Rice}, \texttt{Pork}\}$, $x^2 = \{\texttt{Potato}, \texttt{Carrot}\}$, $x^3 = \{\texttt{Potato}, \texttt{Beef}, \texttt{Carrot}\}$. For this input sequence, the correct output sequence is $\mathbf{y} = (y^1, y^2, y^3) = (\texttt{Y}, \texttt{N}, \texttt{Y})$, where $y^t \in \Sigma = \{\texttt{Y}, \texttt{N}\}$, $y^t = \texttt{Y}$ denotes eating at the Mensa on the t-th day, and $\texttt{N}$ denotes not eating at the Mensa on the t-th day.

Observation Features

We define only the following five observation features to compose $\Psi(x^t) = (\psi_1(x^t), \ldots, \psi_5(x^t))$ for the t-th element x^t in $\mathbf{x}$:

- $\psi_1(x^t) = [[\texttt{Pork} \in x^t]]$
- $\psi_2(x^t) = [[\texttt{Rice} \in x^t]]$
- $\psi_3(x^t) = [[\texttt{Potato} \in x^t]]$
- $\psi_4(x^t) = [[\texttt{Carrot} \in x^t]]$

- $\psi_5(x^t) = [[\texttt{Beef} \in x^t]]$

where $[[\cdot]]$ denotes an Iverson bracket.

Following Altun et al., given an input-output pair $(\mathbf{x}, \mathbf{y})$, we define a set of combined label-observation features for each time-step $t = 1, \ldots, T$, label $\sigma \in \Sigma$ and observation feature $\psi_r(x^t)$:

$$\phi^t_{r\sigma}(\mathbf{x}, \mathbf{y}) = [[y^t = \sigma]] \cdot \psi_r(x^t).^*$$

We can also sum the observation features over time-steps $t = 1, \ldots, T$ to compute the global observation features:

$$\phi_{r\sigma}(\mathbf{x}, \mathbf{y}) = \sum_{t=1}^{T} \phi^t_{r\sigma}(\mathbf{x}, \mathbf{y}),$$

for $\sigma \in \Sigma$ and $r = 1, \ldots, d$, where d is the number of observation features.

Transition Features

In the same way as suggested in Altun et al., we here use transition features of the form:

$$\bar{\phi}^t_{\sigma\tau}(\mathbf{x}, \mathbf{y}) = [[y^t = \sigma \wedge y^{t+1} = \tau]].^\dagger$$

Again, we can sum the transition features over time-steps $t = 1, \ldots, T-1$ to compute the global transition features for $\sigma, \tau \in \Sigma$:

$$\bar{\phi}_{\sigma\tau}(\mathbf{x}, \mathbf{y}) = \sum_{t=1}^{T-1} \bar{\phi}^t_{\sigma\tau}(\mathbf{x}, \mathbf{y}).$$

Finally, if we concatenate all these observation and transition features, we obtain the global feature vector $\Phi(\mathbf{x}, \mathbf{y}) = ((\phi_{r\sigma}), (\bar{\phi}_{\sigma\tau}))$ for $r = 1, \ldots, d$ and $\sigma, \tau \in \Sigma$.[‡]

Primal Model Parameters

For each label $\sigma \in \Sigma$ and each observation feature $\psi_r(\cdot)$, we have an associated observation parameter $w_{r\sigma}$. And, for each pair of labels $\sigma, \tau \in \Sigma$, we have a transition parameter $\bar{w}_{\sigma\tau}$. We then define the complete set of *primal* parameters just as the concatenation of all these parameters, that is $\mathbf{w} = ((w_{r\sigma}), (\bar{w}_{\sigma\tau}))$, for $r = 1, \ldots, d$ and $\sigma, \tau \in \Sigma$.[§] Then we can define the linear discriminant function as:

$$F(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{y}) \rangle.$$

[*] Observe that Altun et al. define a more general feature $\phi^{st}_{r\sigma}(\mathbf{x}, \mathbf{y})$ but, shortly after that, they mention that, in fact, they will be restricted to $s = t$, that is $\phi^{tt}_{r\sigma}(\mathbf{x}, \mathbf{y})$. Here, $\phi^t_{r\sigma}(\mathbf{x}, \mathbf{y})$ is equivalent to $\phi^{tt}_{r\sigma}(\mathbf{x}, \mathbf{y})$ in the paper.

[†] Similarly to the observation features, we have that $\bar{\phi}^t_{\sigma\tau}(\mathbf{x}, \mathbf{y})$ is equivalent to $\phi^{t(t+1)}_{\sigma\tau}(\mathbf{x}, \mathbf{y})$ in the paper.

[‡] i.e., with the set of labels $\Sigma = \{\texttt{N}, \texttt{Y}\}$, features $\Phi(\mathbf{x}, \mathbf{y})$ correspond to the vector $[\Phi_{\text{obs}}^\top, \Phi_{\text{trans}}^\top]^\top$ with $\Phi_{\text{obs}} = [\phi_{1\texttt{N}}, ..., \phi_{d\texttt{N}}, \phi_{1\texttt{Y}}, ..., \phi_{d\texttt{Y}}]^\top$ and $\bar{\Phi}_{\text{trans}} = [\bar{\phi}_{\texttt{NN}}, \bar{\phi}_{\texttt{NY}}, \bar{\phi}_{\texttt{YN}}, \bar{\phi}_{\texttt{YY}}]^\top$.

[§] i.e., with the set of labels $\Sigma = \{\texttt{N}, \texttt{Y}\}$ the weights $\mathbf{w}$ correspond to the vector $[\mathbf{w}_{\text{obs}}^\top, \mathbf{w}_{\text{trans}}^\top]^\top$ with $\mathbf{w}_{\text{obs}} = [w_{1\texttt{N}}, ..., w_{d\texttt{N}}, w_{1\texttt{Y}}, ..., w_{d\texttt{Y}}]^\top$ and $\mathbf{w}_{\text{trans}} = [\bar{w}_{\texttt{NN}}, \bar{w}_{\texttt{NY}}, \bar{w}_{\texttt{YN}}, \bar{w}_{\texttt{YY}}]^\top$.

Dual Model Parameters

The dual model representation consists of defining, for each training input-output pair $(\mathbf{x}_i, \mathbf{y}_i)$, a set of dual variables $\alpha_i(\bar{\mathbf{y}}) \in \mathbb{R}$, where $\bar{\mathbf{y}} \in \mathcal{Y}(\mathbf{x}_i)$ and $\mathcal{Y}(\mathbf{x}_i)$ is the set of all possible output sequences for input $\mathbf{x}_i$. Then we can define the dual discriminant function as:

$$F(\mathbf{x}, \mathbf{y}) = \sum_i \sum_{\bar{\mathbf{y}}} \alpha_i(\bar{\mathbf{y}}) \langle \Phi(\mathbf{x}_i, \bar{\mathbf{y}}), \Phi(\mathbf{x}, \mathbf{y}) \rangle.$$

Obtaining Primal Parameters from Dual Parameters

Let us use the input-output pair given in the beginning of this exercise as a training instance $(\mathbf{x}_1, \mathbf{y}_1)$, such that $\mathbf{x}_1 = (x_1^1, x_1^2, x_1^3)$ and:

$$\begin{aligned} x_1^1 &= \{\texttt{Rice}, \texttt{Pork}\} \\ x_1^2 &= \{\texttt{Potato}, \texttt{Carrot}\} \\ x_1^3 &= \{\texttt{Potato}, \texttt{Beef}, \texttt{Carrot}\}. \end{aligned}$$

Additionally, let the following three output sequences below, associated with $\mathbf{x}_1$, be the only ones with $\alpha_i(\cdot) \neq 0$.

$$\begin{aligned} \mathbf{y}_1 &= (\texttt{Y}, \texttt{N}, \texttt{Y}) & \alpha_1(\mathbf{y}_1) &= +2 \\ \mathbf{y}_{1.1} &= (\texttt{N}, \texttt{N}, \texttt{N}) & \alpha_1(\mathbf{y}_{1.1}) &= -1 \\ \mathbf{y}_{1.2} &= (\texttt{Y}, \texttt{Y}, \texttt{Y}) & \alpha_1(\mathbf{y}_{1.2}) &= -1. \end{aligned}$$

Show that the dual model corresponding to these parameters is equivalent to the primal model with the following observation parameters:

r	$w_{r\texttt{N}}$	$w_{r\texttt{Y}}$	$\psi_r(x^t)$
1	-1	+1	$[[\texttt{Pork} \in x^t]]$
2	-1	+1	$[[\texttt{Rice} \in x^t]]$
3	0	0	$[[\texttt{Potato} \in x^t]]$
4	0	0	$[[\texttt{Carrot} \in x^t]]$
5	-1	+1	$[[\texttt{Beef} \in x^t]]$

and the following transition parameters:

$\bar{w}_{\texttt{NN}} = -2$	$\bar{w}_{\texttt{NY}} = +2$
$\bar{w}_{\texttt{YN}} = +2$	$\bar{w}_{\texttt{YY}} = -2$

Remember that:

$$\mathbf{w} = \sum_i \sum_{\bar{\mathbf{y}}} \alpha_i(\bar{\mathbf{y}}) \Phi(\mathbf{x}_i, \bar{\mathbf{y}}).$$

Solution

First, we should recall that $\mathbf{w}$ and $\Phi(\mathbf{x}, \mathbf{y})$ are aligned such that they share the same dimension. And they comprise the following elements, respectively:

$$\begin{aligned} & w_{r\sigma}, \phi_{r\sigma} && \text{for } r = 1, \ldots, d \text{ and } \sigma \in \Sigma \\ & \bar{w}_{\sigma\tau}, \bar{\phi}_{\sigma\tau} && \text{for } \sigma, \tau \in \Sigma \end{aligned}$$

Then, we can write the equation for each individual element in $\mathbf{w}$ as a function of the corresponding element in $\Phi(\mathbf{x}, \mathbf{y})$. Thus, for $r = 1, \dots, d$ and $\sigma \in \Sigma$, we have:

$$\begin{aligned} w_{r\sigma} &= \sum_i \sum_{\bar{\mathbf{y}}} \alpha_i(\bar{\mathbf{y}}) \phi_{r\sigma}(\mathbf{x}_i, \bar{\mathbf{y}}) \\ &= \sum_i \sum_{\bar{\mathbf{y}}} \alpha_i(\bar{\mathbf{y}}) \sum_{t=1}^{T} \phi_{r\sigma}^t(\mathbf{x}_i, \bar{\mathbf{y}}). \end{aligned}$$

And, for $\sigma, \tau \in \Sigma$, we have that:

$$\begin{aligned} \bar{w}_{\sigma\tau} &= \sum_i \sum_{\bar{\mathbf{y}}} \alpha_i(\bar{\mathbf{y}}) \bar{\phi}_{\sigma\tau}(\mathbf{x}_i, \bar{\mathbf{y}}) \\ &= \sum_i \sum_{\bar{\mathbf{y}}} \alpha_i(\bar{\mathbf{y}}) \sum_{t=1}^{T-1} \bar{\phi}_{\sigma\tau}^t(\mathbf{x}_i, \bar{\mathbf{y}}). \end{aligned}$$

Therefore, we have that:

r	$w_{r\texttt{N}}$	$w_{r\texttt{Y}}$	$\psi_r(x^t)$
1	$+2 \cdot 0 - 1 \cdot 1 - 1 \cdot 0 = -1$	$+2 \cdot 1 - 1 \cdot 0 - 1 \cdot 1 = +1$	$[[\texttt{Pork} \in x^t]]$
2	$+2 \cdot 0 - 1 \cdot 1 - 1 \cdot 0 = -1$	$+2 \cdot 1 - 1 \cdot 0 - 1 \cdot 1 = +1$	$[[\texttt{Rice} \in x^t]]$
3	$+2 \cdot 1 - 1 \cdot 2 - 1 \cdot 0 = 0$	$+2 \cdot 1 - 1 \cdot 0 - 1 \cdot 2 = 0$	$[[\texttt{Potato} \in x^t]]$
4	$+2 \cdot 1 - 1 \cdot 2 - 1 \cdot 0 = 0$	$+2 \cdot 1 - 1 \cdot 0 - 1 \cdot 2 = 0$	$[[\texttt{Carrot} \in x^t]]$
5	$+2 \cdot 0 - 1 \cdot 1 - 1 \cdot 0 = -1$	$+2 \cdot 1 - 1 \cdot 0 - 1 \cdot 1 = +1$	$[[\texttt{Beef} \in x^t]]$

and the following transition parameters:

$\bar{w}_{\sigma\tau}$	N	Y
N	$+2 \cdot 0 - 1 \cdot 2 - 1 \cdot 0 = -2$	$+2 \cdot 1 - 1 \cdot 0 - 1 \cdot 0 = +2$
Y	$+2 \cdot 1 - 1 \cdot 0 - 1 \cdot 0 = +2$	$+2 \cdot 0 - 1 \cdot 2 - 1 \cdot 0 = -2$

References

[Altun et al., 2003] Altun, Y., Tsochantaridis, I., and Hofmann, T. (2003). Hidden markov support vector machines. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 3–10.