CS 446/ECE 449 Machine Learning Homework 6: Structured Prediction

Due on Thursday March 12 2020, noon Central Time

1. [28 points] Structured Prediction

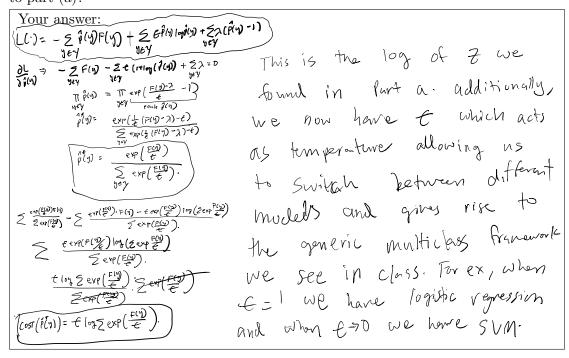
We are interested in jointly predicting/modeling two discrete random variables $y = (y_1, y_2) \in \mathcal{Y}$ with $y_i \in \mathcal{Y}_i = \{0, 1\}$ for $i \in \{1, 2\}$ and $\mathcal{Y} = \prod_{i \in \{1, 2\}} \mathcal{Y}_i$. We define the joint probability distribution to be $p(y) = p(y_1, y_2) = \frac{1}{Z} \exp F(y)$.

(a) (3 points) What is the value of Z (in terms of F(y)) and what is Z called? How many configurations do we need to sum over? Provide the expression using \mathcal{Y}_i .

(b) (6 points) Next we want to solve (for any hyperparameter ϵ)

$$\max_{\hat{p} \in \Delta_{\mathcal{Y}}} \sum_{y \in \mathcal{Y}} \hat{p}(y) F(y) - \sum_{y \in \mathcal{Y}} \epsilon \hat{p}(y) \log \hat{p}(y), \tag{1}$$

where $\Delta_{\mathcal{Y}}$ denotes the probability simplex, *i.e.*, \hat{p} is a valid probability distribution over its domain \mathcal{Y} . Using general notation, write down the Lagrangian and compute its derivative w.r.t. $\hat{p}(y) \ \forall y \in \mathcal{Y}$. Subsequently, find the optimal \hat{p}^* . What is the resulting optimal cost function value for the program given in Eq. (1)? How does this result relate to part (a)?



(c) (3 points) For the program in Eq. (1) assume now $\epsilon = 0$, *i.e.*, we are searching for that configuration $y^* = \arg\max_{\hat{y} \in \mathcal{Y}} F(\hat{y})$ which maximizes F(y). Assume $F(y) = f_1(y_1) + f_2(y_2) + f_{1,2}(y_1, y_2)$. How many different values can the functions f_1 , f_2 and $f_{1,2}$ result in?

Your answer:
$$f_{1} = \left| \begin{cases} 0 & 1 \end{cases} \right| = \lambda$$

$$f_{2} = \left| \begin{cases} 0 & 1 \end{cases} \right| = \lambda$$

$$f_{12} = \left| \begin{cases} (0p)_{1} (01)_{1} (1,0)_{1} (1,1) \right| = 1 \end{cases}$$
Values.

(d) (9 points) As discussed in class, finding the global maximizer can be equivalently written as the following integer linear program:

$$\max_{b} \sum_{r, y_r} b_r(y_r) f_r(y_r) \qquad \text{s.t.} \qquad \begin{cases} b_r(y_r) \in \{0, 1\} & \forall r, y_r \\ \sum_{y_r} b_r(y_r) = 1 & \forall r \\ \sum_{y_p \setminus y_r} b_p(y_p) = b_r(y_r) & \forall r, p \in P(r), y_r \end{cases}$$
(2)

Using the decomposition $F(y) = f_1(y_1) + f_2(y_2) + f_{1,2}(y_1, y_2)$, i.e., for $r \in \{\{1\}, \{2\}, \{1, 2\}\}\}$, explicitly state the integer linear program and all its constraints for the special case that $\mathcal{Y}_i = \{0, 1\}$ for $i \in \{1, 2\}$. (**Hint:** The parent sets are as follows: $P(\{1\}) = \{1, 2\}$ and $P(\{2\}) = \{1, 2\}$. Use notation such as $f_1(y_1 = 0)$ and $f_2(y_1 = 0)$.)

Your answer:

$$\begin{cases}
b_{1}(y_{1}=0) \\
b_{1}(y_{2}=0) \\
b_{2}(y_{2}=0) \\
b_{12}(y_{1}=0,y_{2}=0)
\end{cases}$$

$$b_{12}(y_{1}=0,y_{2}=0)$$

$$b_{12}(y_{2}=0,y_{2}=0)$$

$$b_{12}(y_{2}=0$$

(e) (3 points) Let b be the vector

$$b = [b_1(y_1 = 0), b_1(y_1 = 1), b_2(y_2 = 0), b_2(y_2 = 1), b_{1,2}(y_1 = 0, y_2 = 0), b_{1,2}(y_1 = 1, y_2 = 0), b_{1,2}(y_1 = 0, y_2 = 1), b_{1,2}(y_1 = 1, y_2 = 1)]^{\top}.$$

Specify all but the integrality constraints of part (d) using matrix vector notation, *i.e.*, provide A and c for Ab = c.

(f) (4 points) Complete A6_Structure.py where we approximately solve the integer linear program using the linear programming relaxation. Implement the constraints. Why do we provide -f as input to the solver? What is the obtained result b for the relaxation of the program given in Eq. (2) and its cost function value? Is this the configuration y^* which has the largest score?

