Name: ____________________

University of Illinois Spring 2020

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 $\mathcal{Y}_i$.

Solution:

$$\text{(partition function/normalization constant)} \qquad Z = \sum_{y\in\mathcal{Y}} \exp F(y)$$

Number of summands: $\prod_{i\in\{1,2\}} |\mathcal{Y}_i|$

(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)?

Solution:
Lagrangian:

$$L() = \sum_{y\in\mathcal{Y}} \hat{p}(y)F(y) - \sum_{y\in\mathcal{Y}} \epsilon\hat{p}(y)\log\hat{p}(y) + \lambda\left(1 - \sum_{y\in\mathcal{Y}} \hat{p}(y)\right)$$

Derivative:

$$\frac{\partial L}{\partial \hat{p}(y)} = F(y) - \epsilon - \epsilon\log\hat{p}(y) - \lambda$$

Optimal solution:

$$\hat{p}^*(y) = \frac{\exp F(y)/\epsilon}{\sum_{\hat{y}\in\mathcal{Y}} \exp F(\hat{y})/\epsilon}$$

Optimal cost function value:

$$\epsilon \log \sum_{\hat{y}\in\mathcal{Y}} \exp F(\hat{y})/\epsilon$$

For $\epsilon = 1$ optimal value equals $\log Z$

(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?

Solution:
$f_1 : |\mathcal{Y}_1|$
$f_2 : |\mathcal{Y}_2|$
$f_{1,2} : |\mathcal{Y}_1||\mathcal{Y}_2|$

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

$$\max_b \sum_{r,y_r} b_r(y_r) f_r(y_r) \qquad \text{s.t.} \qquad \left\{ \begin{array}{ll} b_r(y_r) \in \{0,1\} & \forall r, y_r \\ \sum_{y_r} b_r(y_r) = 1 & \forall r \\ \sum_{y_p \setminus y_r} b_p(y_p) = b_r(y_r) & \forall r, p \in P(r), y_r \end{array} \right. . \tag{2}$$

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 $b_1(y_1 = 0)$.)

Solution:

$$\max_b \quad b_1(y_1=0)f_1(y_1=0) + b_1(y_1=1)f_1(y_1=1) + b_2(y_2=0)f_2(y_2=0) + b_2(y_2=1)f_2(y_2=1) +$$
$$b_{1,2}(y_1=0,y_2=0)f_{1,2}(y_1=0,y_2=0) + b_{1,2}(y_1=1,y_2=0)f_{1,2}(y_1=1,y_2=0) +$$
$$b_{1,2}(y_1=0,y_2=1)f_{1,2}(y_1=0,y_2=1) + b_{1,2}(y_1=1,y_2=1)f_{1,2}(y_1=1,y_2=1)$$

s.t.

$$b_1(y_1=0) \in \{0,1\}, b_1(y_1=1) \in \{0,1\}, b_2(y_2=0) \in \{0,1\}, b_2(y_2=1) \in \{0,1\}$$
$$b_{1,2}(y_1=0,y_2=0) \in \{0,1\}, b_{1,2}(y_1=0,y_2=1) \in \{0,1\}, b_{1,2}(y_1=1,y_2=0) \in \{0,1\}, b_{1,2}(y_1 =$$
$$b_1(y_1=0) + b_1(y_1=1) = 1$$
$$b_2(y_2=0) + b_2(y_2=1) = 1$$
$$b_{1,2}(y_1=0,y_2=0) + b_{1,2}(y_1=1,y_2=0) + b_{1,2}(y_1=0,y_2=1) + b_{1,2}(y_1=1,y_2=1) = 1$$
$$b_{1,2}(y_1=0,y_2=0) + b_{1,2}(y_1=0,y_2=1) = b_1(y_1=0)$$
$$b_{1,2}(y_1=1,y_2=0) + b_{1,2}(y_1=1,y_2=1) = b_1(y_1=1)$$
$$b_{1,2}(y_1=0,y_2=0) + b_{1,2}(y_1=1,y_2=0) = b_2(y_2=0)$$
$$b_{1,2}(y_1=0,y_2=1) + b_{1,2}(y_1=1,y_2=1) = b_2(y_2=1)$$

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

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

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

Name:

Solution:

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ -1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & -1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 & 1 \end{bmatrix} \quad \text{and} \quad c = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

(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?

Solution:
Solver solves a minimization problem rather than a maximization.

$$b = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Cost function value: 5
This is the largest possible score for the given f