Final Exam: "Discrete Inference and Learning"

22 décembre 2017

durée : 3h00

Documents autorisés: transparents / notes / slides du cours.
Authorized documents: Notes / slides from the course.

Ordinateur **non-connecté** *autorisé pour accéder les transparents du cours déjà téléchargés.*
Laptops / Computers not connected to the network are allowed to access the course slides already downloaded.

Calculatrice et téléphone portable non autorisés.
Calculators and mobile phones are not allowed.

Conseil: Veuillez lire toutes les **5** *questions avant de répondre.*
Tip: Read all the **5** questions before you begin to answer.

Part 1

[Question 1.] (6 points) Inference algorithms
Consider an inference problem, where the goal is to minimize an energy function of the following form. It is defined on N discrete variables X_i taking corresponding labels x_i, and with pairwise neighborhoods denoted by the set of edges $\mathcal{E}$.

$$E(\mathbf{x}) = \sum_{i=1}^{N} E_i(x_i) + \sum_{(i,j)\in\mathcal{E}} E_{ij}(x_i, x_j).$$

(1 points) List the inference algorithms applicable for solving this generic energy function.

Let us now restrict this energy function to further study relevant inference algorithms.

(1.5 points) Consider a binary labelling problem, where x_i can take one of two labels, for example, $x_i \in \{0, 1\}$. In this scenario, briefly describe the graph cut inference algorithm. Are there any conditions on the energy function for using graph cut? Explain if your answer is yes or no.

(1.5 points) Now, let x_i take one of multiple labels, i.e., $x_i \in \{0, 1, \dots, n-1\}$. Can the two-label graph cut algorithm be adapted to this multi-label case? Explain if your answer is yes or no.

(1 points) For the multi-label case, describe two message-passing inference algorithms briefly.

(1 point) Compare the two message-passing algorithms and explain which one is better.

[Question 2.] (5 points) Parameter learning
Let us assume that an energy function $E(\mathbf{x})$ is parameterized with $\mathbf{w}$. We can learn these parameters by minimizing a regularized risk function given by:

$$\min_{\mathbf{w}} R(\mathbf{w}) + \sum_{k=1}^{K} L_G(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}),$$

where R is a regularization function, such as the ℓ_2 norm of $\mathbf{w}$, and L_G is a loss function minimizing the "difference" between the estimated labels $\mathbf{x}^k$ and the ground truth labels $\mathbf{z}^k$ for every training sample k.

(2 points) Changing the loss function determines the type of learning. Define the loss function L_G for max-margin and maximum-likelihood learning methods.

(2 points) Discuss the advantages and disadvantages of max-margin and maximum-likelihood learning methods by comparing them.

(1 point) Solving the parameter learning problem requires efficient inference. Discuss if this statement is true or false.

Part 2

[Question 3.] (6 points) Dual decomposition

This question will help you to understand dual decomposition so that you can implement it by yourself.

Recall that the MRF energy over a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is given by

$$E(\mathbf{x}, \boldsymbol{\theta}) = \sum_{i \in \mathcal{V}} \sum_{l \in \mathcal{L}} \theta_i(l) x_i(l) + \sum_{ij \in \mathcal{E}} \sum_{l \in \mathcal{L}} \sum_{l' \in \mathcal{L}} \theta_{ij}(l, l') x_{ij}(l, l'), \tag{1}$$

where $\mathcal{L}$ is the set of labels, $\{\theta_i(\cdot)\}_{i \in \mathcal{V}}$ and $\{\theta_{ij}(\cdot, \cdot)\}_{ij \in \mathcal{E}}$ are respectively the unary and pairwise potential functions, $x_i(l)$ and $x_{ij}(l, l')$ are binary indicator functions defined by:

$$x_i(l) = 1 \Leftrightarrow \text{ label } l \text{ is assigned to node } i, \tag{2}$$

$$x_{ij}(l, l') = 1 \Leftrightarrow \text{ labels } (l, l') \text{ are assigned to edge } ij. \tag{3}$$

The vectors $\mathbf{x}$ and $\boldsymbol{\theta}$ are the concatenation of all individual values of indicator functions and potential functions, respectively.

It is straightforward to see that the energy (1) can be written as $\boldsymbol{\theta} \cdot \mathbf{x}$, the dot product between $\boldsymbol{\theta}$ and $\mathbf{x}$ (in this question we use the dot product notation instead of the transpose $\boldsymbol{\theta}^\top \mathbf{x}$ for clarity purpose). The MRF optimization problem can then be reformulated as:

$$\min \quad \boldsymbol{\theta} \cdot \mathbf{x} \quad \text{ s.t. } \quad \mathbf{x} \in \bar{\mathcal{X}}_{\mathcal{G}}, \tag{4}$$

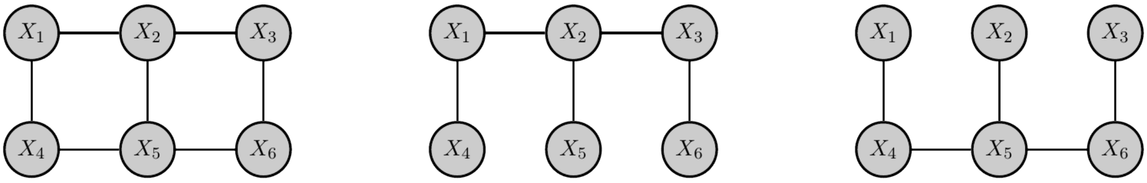
where $\bar{\mathcal{X}}_{\mathcal{G}}$ is called the *marginal polytope*, defined by

$$\bar{\mathcal{X}}_{\mathcal{G}} = \left\{ \mathbf{x} \;\middle|\; \begin{array}{ll} \sum_{l \in \mathcal{L}} x_i(l) = 1, & \forall i \in \mathcal{V}, \\ \sum_{l' \in \mathcal{L}} x_{ij}(l, l') = x_i(l), & \forall ij \in \mathcal{E}, l \in \mathcal{L}, \\ \sum_{l \in \mathcal{L}} x_{ij}(l, l') = x_j(l'), & \forall ij \in \mathcal{E}, l' \in \mathcal{L}, \\ x_i(l) \in \{0,1\}, x_{ij}(l, l') \in \{0,1\} & \forall ij \in \mathcal{E}, l \in \mathcal{L}, l' \in \mathcal{L}. \end{array} \right\}. \tag{5}$$

If we replace the constraints $x_i(l) \in \{0,1\}, x_{ij}(l, l') \in \{0,1\}$ in the above by $x_i(l) \geq 0, x_{ij}(l, l') \geq 0$, we obtain a new set $\mathcal{X}_{\mathcal{G}}$, called the *local polytope*. Minimizing the MRF energy over $\mathcal{X}_{\mathcal{G}}$ is known as the *standard LP relaxation*:

$$\min \quad \boldsymbol{\theta} \cdot \mathbf{x} \quad \text{ s.t. } \quad \mathbf{x} \in \mathcal{X}_{\mathcal{G}}. \tag{6}$$

Now let us consider the graph $\mathcal{G}$ in Figure 1a and derive a dual decomposition solution based on the the graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in Figures 1b and 1c.



(a) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (b) $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ (c) $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$

Figure 1: We consider the decomposition of $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ into $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ and $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$.

Note that by definition (5), the dimensions of the vectors in $\mathcal{X}_\mathcal{G}$ and $\mathcal{X}_{\mathcal{G}_1}$ (or $\mathcal{X}_{\mathcal{G}_2}$) are different because $\mathcal{G}$ and $\mathcal{G}_1$ (or $\mathcal{G}_2$) do not have the same number of edges. For ease of presentation, we can assume that the dimensions are the same (by adding variables to the missing edges without adding constraints on them).

(a) (0.5 points) Show that $\mathcal{X}_\mathcal{G} = \mathcal{X}_{\mathcal{G}_1} \cap \mathcal{X}_{\mathcal{G}_2}$.

(b) (1 point) Show that the corresponding MRF standard LP relaxation can be rewritten as

$$\begin{array}{ll} \min & \boldsymbol{\theta}^1 \cdot \mathbf{x} + \boldsymbol{\theta}^2 \cdot \mathbf{x} \\ \text{s.t.} & \mathbf{x} \in \mathcal{X}_{\mathcal{G}_1},\ \mathbf{x} \in \mathcal{X}_{\mathcal{G}_2}, \end{array} \tag{7}$$

with suitably chosen $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$, which can be seen as potentials corresponding to $\mathcal{G}_1$ and $\mathcal{G}_2$ (give explicit relations of the node/edge potentials between $\boldsymbol{\theta}^1$,$\boldsymbol{\theta}^2$ and $\boldsymbol{\theta}$).

(c) (1.5 points) Using two auxiliary variables $\mathbf{x}^1$ and $\mathbf{x}^2$, prove that a dual function for (7) is

$$g(\boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2) = \min_{\mathbf{x}^1 \in \mathcal{X}_{\mathcal{G}_1}} (\boldsymbol{\theta}^1 + \boldsymbol{\lambda}^1) \cdot \mathbf{x}^1 + \min_{\mathbf{x}^2 \in \mathcal{X}_{\mathcal{G}_2}} (\boldsymbol{\theta}^2 + \boldsymbol{\lambda}^2) \cdot \mathbf{x}^2 + \min_{\mathbf{x}} -(\boldsymbol{\lambda}^1 + \boldsymbol{\lambda}^2) \cdot \mathbf{x}. \tag{8}$$

Explain in a few lines how to solve each of the above three minimization subproblems.

(d) (0.5 points) Show that the corresponding dual problem is equivalent to $\min_{\boldsymbol{\lambda} \in \Lambda} f(\boldsymbol{\lambda})$ where $f(\boldsymbol{\lambda}) = -g(\boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2)$ and $\Lambda = \left\{\boldsymbol{\lambda} = (\boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2) \mid \boldsymbol{\lambda}^1 + \boldsymbol{\lambda}^2 = \mathbf{0}\right\}$.

(e) (1 point) Denote $g_1(\boldsymbol{\lambda}^1) = \min_{\mathbf{x}^1 \in \mathcal{X}_{\mathcal{G}_1}} (\boldsymbol{\theta}^1 + \boldsymbol{\lambda}^1) \cdot \mathbf{x}^1$ and let $\mathbf{x}^1(\boldsymbol{\lambda}^1)$ denote the optimal solution of this minimization subproblem. Prove that $-\mathbf{x}^1(\boldsymbol{\lambda}^1)$ is a subgradient of $-g_1$ at $\boldsymbol{\lambda}^1$. Deduce a subgradient of f at $\boldsymbol{\lambda}$. Reminder: $\mathbf{s}$ is a subgradient of h at $\mathbf{u}$ if and only if $h(\mathbf{v}) \geq h(\mathbf{u}) + \mathbf{s} \cdot (\mathbf{v} - \mathbf{u}) \quad \forall \mathbf{v}$.

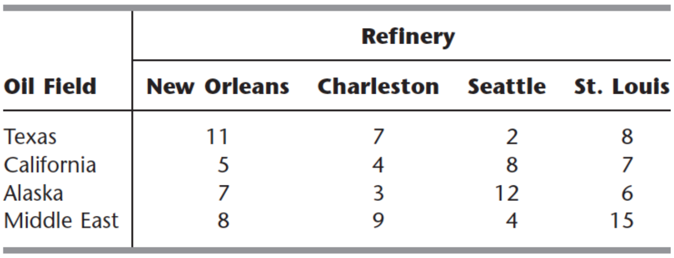
(f) (Bonus, 0.5 points) The projection of a vector $\boldsymbol{\lambda}$ onto the set Λ is the closest point of Λ to $\boldsymbol{\lambda}$, defined by $\text{Proj}_\Lambda(\boldsymbol{\lambda}) := \text{argmin}_{\boldsymbol{\lambda}' \in \Lambda} \|\boldsymbol{\lambda}' - \boldsymbol{\lambda}\|$. Prove that $\text{Proj}_\Lambda(\boldsymbol{\lambda}) = \boldsymbol{\lambda}_*$ where $\boldsymbol{\lambda}^1_* = \boldsymbol{\lambda}^1 - \frac{\boldsymbol{\lambda}^1 + \boldsymbol{\lambda}^2}{2}$ and $\boldsymbol{\lambda}^2_* = \boldsymbol{\lambda}^2 - \frac{\boldsymbol{\lambda}^1 + \boldsymbol{\lambda}^2}{2}$.

(g) (1.5 points) For minimizing the non-differentiable convex function $f(\boldsymbol{\lambda})$ over the closed convex set Λ, the projected subgradient method uses the iteration $\boldsymbol{\lambda}^{(k+1)} = \text{Proj}_\Lambda\left(\boldsymbol{\lambda}^{(k)} - \alpha^{(k)}\mathbf{s}^{(k)}\right)$, where $\mathbf{s}^{(k)}$ is a subgradient of f at $\boldsymbol{\lambda}^{(k)}$, and $\alpha^{(k)} > 0$ is the k^{th} step-size (that must follow some rules to guarantee the convergence). Based on the above elements, describe the dual decomposition algorithm for solving the standard LP relaxation of the MRF corresponding to the graph G in Figure 1.

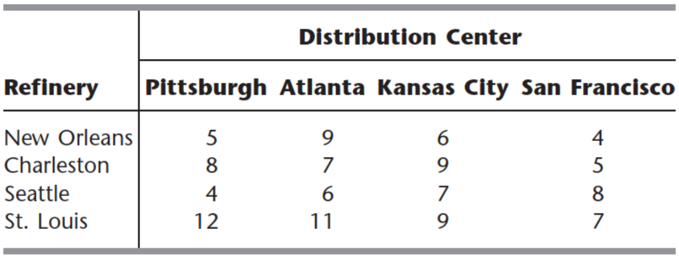
Part 3

[Question 4.] (4 points) Network Optimization

The Texago Corporation has four oil fields, four refineries, and four distribution centers. A major strike involving the transportation industries now has sharply curtailed Texago's capacity to ship oil from the oil fields to the refineries and to ship petroleum products from the refineries to the distribution centers. Using units of thousands of barrels of crude oil (and its equivalent in refined products), the following tables show the maximum number of units that can be shipped per day from each oil field to each refinery, and from each refinery to each distribution center.



	Refinery			
Oil Field	**New Orleans**	**Charleston**	**Seattle**	**St. Louis**
Texas	11	7	2	8
California	5	4	8	7
Alaska	7	3	12	6
Middle East	8	9	4	15



	Distribution Center			
Refinery	**Pittsburgh**	**Atlanta**	**Kansas City**	**San Francisco**
New Orleans	5	9	6	4
Charleston	8	7	9	5
Seattle	4	6	7	8
St. Louis	12	11	9	7

The Texago management now wants to determine a plan for how many units to ship from each oil field to each refinery and from each refinery to each distribution center that will maximize the total number of units reaching the distribution centers.

(2 points.) Draw the distribution network, formulate and solve maximum-flow problem to determine a plan for the Texago management.

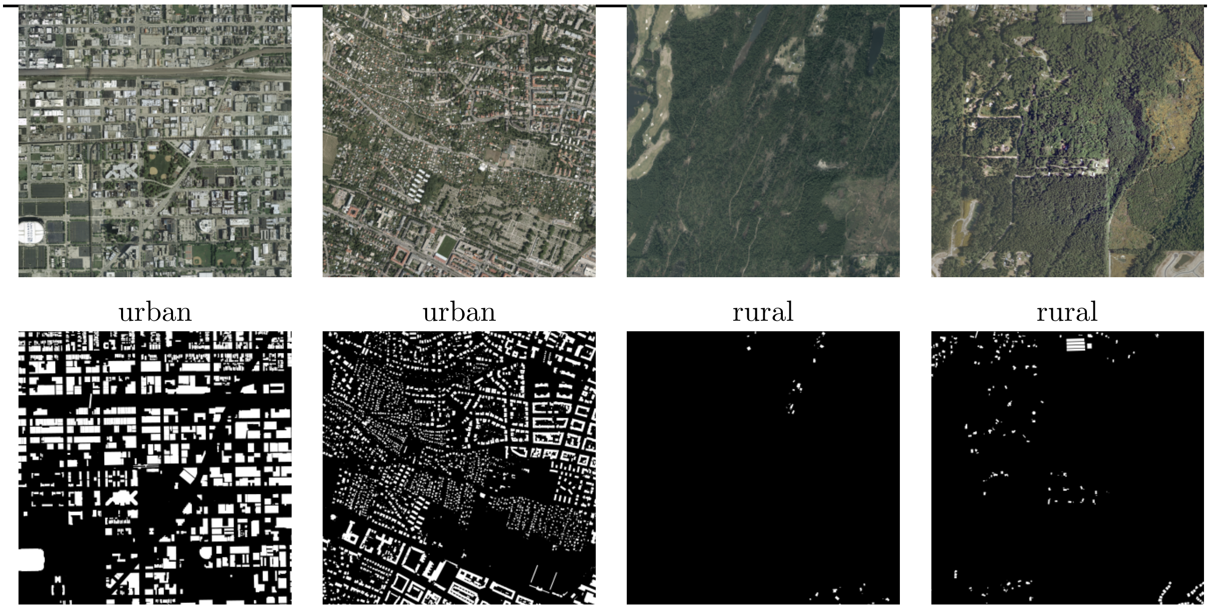
(2 points.) Is the algorithm you use guaranteed or not to terminate in a finite number of iterations when it is applied to a max-flow graph whose edge capacities are all non-negative integers? Prove your statement.

[Question 5.] (5 points) Learning

You are given a dataset of 360 aerial RGB images of spatial dimensions 5000 × 5000, acquired over cities and villages in different parts of the world. Half of these images are labeled in two ways: (1) A label is given for the whole image to indicate if it is urban or rural area. (2) A pixelwise labeling of each image assigns every pixel to one of two classes: building or non-building. Examples of images and labelings are given below.

First, we consider categorization problem, i.e. we want to learn assigning urban/rural labels for new images.

(a) (1.5 points) Which classification method you would use for this task? Propose a classification scheme. Explain what would be the discriminative features and which optimizer would be used for the proposed classification algorithm.



urban urban rural rural

Now we consider dense labeling problem, where we want to classify every pixel of new images to building/non-building class.

(b) (1 point) Explain advantages and disadvantages of using random forest or convolutional neural network-based classifier for this task.

(c) (1 point) Propose a feature selection strategy for random forest classifier.

(d) (1.5 points) Propose a CNN architecture for this dense labeling task. Explain which loss function and optimization algorithm can be well suited for this network.