Kaggle Competition:Planet: Understanding the Amazon from Space

June 7, 2017

1 Problem Statement

Given a label satellite image, the task is to map the image to a particular atmospheric condition and various classes of land cover/land use. The labels for this task can broadly be broken into three groups: atmospheric conditions, common land cover/land use phenomena, and rare land cover/land use phenomena.

Let us denote our input image by the vector $\boldsymbol{x}$ and the corresponding target labels by the binary vector $\boldsymbol{y}$. The vector $\boldsymbol{y}$ is a sequence of binary random variables i.e $\boldsymbol{y} = (y_1, y_2 \cdots y_K)$ where each variable y_j denotes the presence of absence of a particular label. In this representation, the first 4 random variables $(y_1..y_4)$ correspond to atmospheric labels while the rest of the sequence corresponds to variables belonging to the remaining two groups.

In the task description, it is stated that each chip will have one atmospheric label and zero or more common and rare labels. This means the variables in the sub-sequence $(y_1..y_4)$ are mutually exclusive. Furthermore, it is also stated that on days when its is cloudy, nothing else can be observed. This means that $y_1 = 1 \implies y_j = 0 \; \forall j \neq 1$. To enforce this constraint, let us partition the vectors $\boldsymbol{y}$ into two clusters : { cloudy, not-cloudy} where the class cloudy only contains the binary variable $\boldsymbol{y} = (1, 0, 0, 0 \cdots 0)$

2 Modelling the conditional distribution

We wish to train our Deep Convolutional Network to learn to distribution $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ from seen examples. Before we proceed any further, for notational clarity, let us represent the parameters of the model by $\boldsymbol{\theta} = \{\boldsymbol{\theta}_c, \boldsymbol{\theta}_A, \boldsymbol{\theta}_5..\boldsymbol{\theta}_K, \hat{\boldsymbol{\theta}}\}$ where

- $\hat{\boldsymbol{\theta}}$ corresponds to the parameters of the part of the network that generates a transformed input $\phi(\boldsymbol{x})$
- $\{\boldsymbol{\theta}_5..\boldsymbol{\theta}_K\}$ correspond to individual parameters associated with logistic classifiers for each label that doesnt belong to the atmospheric group
- $\boldsymbol{\theta}_c$ corresponds to the parameters of the logistic classifier that predicts the probability of the cluster *cloudy*

- $\boldsymbol{\theta}_A$ are the parameters of the soft-max classifier that predicts the other atmospheric conditions given that the weather is not *cloudy*.

Under the framework described in section 1, we can factorise the distribution learned by our model as follows:

$$\begin{aligned} P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) =& \sum_i P(\boldsymbol{y}, C=i|\boldsymbol{x},\boldsymbol{\theta}) \\ =& \sum_i P(\boldsymbol{y}|C=i,\boldsymbol{x},\boldsymbol{\theta}) P(C=i|\boldsymbol{x},\boldsymbol{\theta}) \\ =& \text{ Since only cluster C is compatible with any value of } \boldsymbol{y} \\ =& P(\boldsymbol{y}|C=c(\boldsymbol{y}),\boldsymbol{x},\boldsymbol{\theta}) P(C=c(\boldsymbol{y})|\boldsymbol{x},\boldsymbol{\theta}) \end{aligned} \tag{1}$$

Let us first consider the probability $P(C=c(\boldsymbol{y})|\boldsymbol{x},\boldsymbol{\theta})$. Since we have only two clusters, we can represent this probability as a Bernoulli distribution by taking into account the i.id distribution of $\boldsymbol{x}$:

$$P(C=c(\boldsymbol{y})|\boldsymbol{x},\boldsymbol{\theta}) = Ber(C=c(\boldsymbol{y})|\zeta_c(\phi(\boldsymbol{x})) \tag{2}$$

where

$$\begin{aligned} \zeta_c(\phi(\boldsymbol{x})) &= E(c(\boldsymbol{y})=1|\phi(\boldsymbol{x})) \\ &= P(c(\boldsymbol{y})=1|\phi(\boldsymbol{x})) \\ &= sig(\boldsymbol{\theta}_c^T \phi(\boldsymbol{x})) \\ &= \sigma_c \end{aligned}$$

Thus,

$$P(C=c(\boldsymbol{y})|\boldsymbol{x},\boldsymbol{\theta}) = \sigma_c^{y_1}(1-\sigma_c)^{1-y_1}$$

On the other hand, due to the way we have partitioned our classes:

$$P(\boldsymbol{y}|C=c(\boldsymbol{y}),\boldsymbol{x},\boldsymbol{\theta}) = \begin{cases} 1 & \text{if } y_1 = 1 \\ P(\boldsymbol{y}|C=0,\boldsymbol{x},\boldsymbol{\theta}) & \text{if } y_1 = 0 \end{cases}$$

The random variables $(y_2, .y_4)$ in $\boldsymbol{y}$ correspond to atmospheric conditions and are mutually exclusive. In addition, these variables are mutually independent from the other labels. Hence, we can represent the distribution of $(y_2, .y_4)$ by a categorical distribution assuming that weather is not cloudy. The labels corresponding to various land covers and uses are mutually independent to each other. Therefore, the distribution of the individual variables can be expressed Bernoulli distribution assuming the fact that $y_1 = 0$.
Formally,

$$\begin{aligned} P(\boldsymbol{y}|C=0,\boldsymbol{x},\boldsymbol{\theta}) &= Cat(\boldsymbol{y}_{2:4}|\boldsymbol{\theta}_A,\phi(\boldsymbol{x})) \prod_{j=5}^{K} Ber(y_j|\zeta_j(\phi(\boldsymbol{x})) \\ &= \prod_{k=2}^{4} \mu_k^{y_k} \prod_{j=5} \sigma_j^{y_j}(1-\sigma_j)^{1-y_j} \end{aligned}$$

Thus,

$$P(\boldsymbol{y}|C = c(\boldsymbol{y}), \boldsymbol{x}, \boldsymbol{\theta}) = 1^{y_1} \left[\prod_{k=2}^{4} \mu_k^{y_k} \prod_{j=5} \sigma_j^{y_j} (1-\sigma_j)^{1-y_j} \right]^{1-y_1}$$

We now have representations for the probabilities $P(\boldsymbol{y}|C = c(\boldsymbol{y}), \boldsymbol{x}, \boldsymbol{\theta})$ and $P(C = c(\boldsymbol{y})|\boldsymbol{x}, \boldsymbol{\theta})$. Therefore,

$$\begin{aligned} P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) &= P(\boldsymbol{y}|C = c(\boldsymbol{y}), \boldsymbol{x}, \boldsymbol{\theta}) P(C = c(\boldsymbol{y})|\boldsymbol{x}, \boldsymbol{\theta}) \\ &= 1^{y_1} \left[\prod_{k=2}^{4} \mu_k^{y_k} \prod_{j=5} \sigma_j^{y_j} (1-\sigma_j)^{1-y_j} \right]^{1-y_1} \sigma_c^{y_1} (1-\sigma_c)^{1-y_1} \\ &= \left[\prod_{k=2}^{4} \mu_k^{y_k} \prod_{j=5} \sigma_j^{y_j} (1-\sigma_j)^{1-y_j} \right]^{1-y_1} \sigma_c^{y_1} (1-\sigma_c)^{1-y_1} \end{aligned} \tag{3}$$

$$\begin{aligned} \log P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) &= log \left(\left[\prod_{k=2}^{4} \mu_k^{y_k} \prod_{j=5} \sigma_j^{y_j} (1-\sigma_j)^{1-y_j} \right]^{1-y_1} \sigma_c^{y_1} (1-\sigma_c)^{1-y_1} \right) \\ &= (1-y_1) \sum_{k=2}^{4} y_k \log \mu_k + (1-y_1) \left[\sum_{j=5}^{K} y_j \log \sigma_j + (1-y_j) \log (1-\sigma_j) \right] + y_1 \log \sigma_c + (1-y_1) \log (1-\sigma_c) \end{aligned} \tag{4}$$