Tutorial on Laplacian Mean Map for Learning with Label Proportions

Giorgio Patrini
Australian National University and NICTA
Sydney, NSW, Australia
giorgio.patrini@anu.edu.au

September 23, 2015

Abstract

This tutorial complements Patrini et al. (2014) as a practical guide. We left aside most of the mathematic formalism and exemplify the case of learning by minimizing the logistic loss –equivalent to fitting a conditional exponential family via maximum likelihood estimation.

1 Framework

Hereafter, boldfaces like $\boldsymbol{u}$ denote vectors, whose coordinates are denoted u_l. Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$ be the feature and label spaces. Examples are couples (observation, label) $\in \mathcal{X} \times \mathcal{Y}$, sampled i.i.d. according to some unknown but fixed distribution $\mathcal{D}$. Let $\mathcal{S} \doteq \{(\boldsymbol{x}_i, y_i), i \in [m]\} \sim \mathcal{D}_m$ denote sample of size m. In Learning with Label Proportions (LLP), we do not observe directly $\mathcal{S}$ but $\mathcal{S}_{|y}$, which denotes $\mathcal{S}$ with labels removed; we are given its partition in $n > 0$ bags, $\mathcal{S}_{|y} = \cup_j \mathcal{S}_j, j \in [n]$, along with their respective label proportions $\hat{\pi}_j \doteq 1/m_j \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{S}_j} 1\{y_i > 0\}$ and bag proportions $\hat{p}_j \doteq m_j/m$, with $m_j = |\mathcal{S}_j|$. We do not assume to know the process that assigned examples to bags.

As if we knew the label of each example, we learn linear classifiers θ fitting the conditional exponential family of the form

$$\begin{aligned} p(y|\boldsymbol{x}) &= \exp\left(y\boldsymbol{\theta}^\top \boldsymbol{x} - g(\boldsymbol{\theta}|\boldsymbol{x})\right), \\ g(\theta|\boldsymbol{x}) &= \log \sum_{\sigma \in \{-1,1\}} \exp\left(\sigma \boldsymbol{\theta}^\top \boldsymbol{x}\right) \ , \end{aligned}$$

where g is known as the log-partition function. Maximum log-likelihood estimation gives us

$$\operatorname*{argmax}_{\boldsymbol{\theta}} \log \prod_{i=1}^{m} p(y_i|\boldsymbol{x}_i) = \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{m} g(\theta|\boldsymbol{x}_i) - \boldsymbol{\theta}^\top \sum_{i=1}^{m} y_i \boldsymbol{x}_i \ . \tag{1}$$

The key insight for learning with label proportions is here. The log-partition function does not depend on y, therefore it can be computed without knowing any label. The remaining term is the inner product between the model and a statistic called (empirical) *mean operator*, defined as

$$\boldsymbol{\mu}_{\mathcal{S}} \doteq \frac{1}{m} \sum_{i=1}^{m} y_i \boldsymbol{x}_i = \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}y] \ .$$

More precisely, the second term is *proportional* to $\boldsymbol{\theta}^\top \boldsymbol{\mu}_{\mathcal{S}}$ by $1/m$, that does not play any role in the minimization. We cannot compute the mean operator directly, as we do not know the labels. But if we can

estimate this statistic from the label proportions, then we do not need to know the labels at all. Indeed, the mean operator is a sufficient statistic for the label variable for a certain set of losses, as stated by Lemma 1 in Patrini et al. (2014); see also the Appendix. Notice that, if we knew the true mean operator, there would not be any information loss due to the knowledge of the label proportions only; in practice, the more accurate the mean operator estimate, the closer the loss is to the one computed knowing all the labels. Therefore, the strategy is to estimate the mean operator at first, and then plug it into the optimization problem (1). This is an application of the concept of *reduction* between learning problem: instead of reinventing the wheel for solving LLP, we compute the only missing quantity that prevents us to run ordinary logistic regression. We now discuss how to estimate this statistic.

We start by expanding the mean operator in its bag-wise label-wise components:

$$\begin{aligned}\boldsymbol{\mu}_{\mathcal{S}} &= \sum_{j=1}^{n} \hat{p}_j \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}y|j] \\ &= \sum_{j=1}^{n} \hat{p}_j \sum_{\sigma\in\{-1,+1\}} \sigma\hat{\pi}_j \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|\sigma,j] \\ &= \sum_{j=1}^{n} \hat{p}_j \left(\hat{\pi}_j \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|+1,j] - (1-\hat{\pi}_j)\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|-1,j]\right) \ . \end{aligned} \tag{2}$$

The quantities $\hat{p}_j$ and $\hat{\pi}_j$ defined above are given. So our problem is turned into the estimation of $2n$ vectors of unknowns $\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|\sigma,j]$ of dimension d, as the feature space. Those vectors are the solution of a linear system expressed by mean of law of total probability:

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j] &= \sum_{\sigma\in\{-1,+1\}} \mathbb{E}_{\mathcal{S}}[\boldsymbol{x},\sigma|j] \\ &= \sum_{\sigma\in\{-1,+1\}} \hat{\pi}_j \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|\sigma,j] \ , \end{aligned}$$

and again, every $\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j]$ can be computed even without label knowledge as they are the bag-wise average feature vectors. For convenience, we rewrite the system in matrix form. Let $\boldsymbol{b}_j^{\sigma} = \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|\sigma,j]$ and $\boldsymbol{b}_j = \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j]$. The $2n$ $\boldsymbol{b}_j^{\sigma}$s are solution of

$$\mathrm{B} - \Pi^{\top}\mathrm{B}^{\pm} \quad = \quad \mathbf{0} \ , \tag{3}$$

where $\mathrm{B} \doteq [\boldsymbol{b}_1|\boldsymbol{b}_2|...|\boldsymbol{b}_n]^{\top} \in \mathbb{R}^{n\times d}$, $\Pi \doteq [\textsc{Diag}(\hat{\boldsymbol{\pi}})|\textsc{Diag}(\mathbf{1}-\hat{\boldsymbol{\pi}})]^{\top} \in \mathbb{R}^{2n\times n}$ and $\mathrm{B}^{\pm} \in \mathbb{R}^{2n\times d}$ is the matrix of unknowns:

$$\mathrm{B}^{\pm} \quad \doteq \quad \Big[\underbrace{\boldsymbol{b}_1^{+1}|\boldsymbol{b}_2^{+1}|...|\boldsymbol{b}_n^{+1}}_{(\mathrm{B}^{+})^{\top}} \Big| \underbrace{\boldsymbol{b}_1^{-1}|\boldsymbol{b}_2^{-1}|...|\boldsymbol{b}_n^{-1}}_{(\mathrm{B}^{-})^{\top}}\Big]^{\top} \ . \tag{4}$$

System (3) is underdetermined, as it is made of $n \times d$ equations on $2n \times d$ unknowns. This fact should not be a surprise as it expresses the ill-posed nature of the problem of learning a classifier with label proportions only. To solve the system we need to resort to additional assumptions. We remark that System (3) is the first estimation step, followed by Equation (2), that recovers the mean operator, and finally by the minimization problem in (1). We now focus on solutions to the linear system.

2 The Mean Map algorithm of Quadrianto et al. (2009)

The framework illustrated above was first presented by Quadrianto et al. (2009) to construct the Mean Map algorithm. The original method enforces a *homogeneity assumption*, that is a statement of conditional

independence for j:

$$\forall j,\ \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|\sigma, j] = \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|\sigma]\ ,$$

Under this hypothesis, the number of unknowns falls to $2d$. This is enough for obtaining a well-formed system, granted that $n \geq 2$. In this case, the linear system is simplified by defining $\mathrm{B}^{\pm}_{MM} \doteq [\boldsymbol{b}^{+1}, \boldsymbol{b}^{-1}]^\top \in \mathbb{R}^{n\times d}$ and $\Pi_{MM} \doteq [\hat{\boldsymbol{\pi}}|\mathbf{1} - \hat{\boldsymbol{\pi}}] \in \mathbb{R}^{n\times 2}$, and the solution is found by pseudo-inversion as $\tilde{\mathrm{B}}^{\pm}_{MM} = \Pi^{\dagger}_{MM}\mathrm{B}$. The mean operator is then computed by $\boldsymbol{\mu}^{MM}_{\mathcal{S}} = \sum_{y\in\{-1,+1\}} yp(y)\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|y]$, where $p(y)$ is the probability of the label over the whole $\mathcal{S}$ and can be derived by summing the proportions.

3 Laplacian Mean Map

The Laplacian Mean Map algorithm aims to solve System (3) at its full extent. In order to do that, we relax the homogeneity assumption as

$$\forall j, j' \text{ if } j \approx j' \text{ then } \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|y, j] \approx \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|y, j']\ .$$

Here we are relaxing the homogeneity assumption in a precise sense: instead of having every $\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|y, j]$ equal for each bag j for a given y, we assume them close when bags are similar, while none is constrained to be equal. The similarity between bags j and j' is a domain-specific parameter of the algorithm and we will refer to it by $v_{j,j'} \geq 0$. To incorporate the assumption into the estimation, we solve System (3) by least square minimization with regularization as

$$\underset{\boldsymbol{b}^1_j, \boldsymbol{b}^{-1}_j}{\operatorname{argmin}} \sum_j (\boldsymbol{b}_j - \hat{\pi}_j \boldsymbol{b}^1_j + (1-\hat{\pi}_j)\boldsymbol{b}^{-1}_j)^2 + \gamma \sum_{j,j'} v_{j,j'}[(\boldsymbol{b}^1_j - \boldsymbol{b}^1_{j'})^2 + (\boldsymbol{b}^{-1}_j - \boldsymbol{b}^{-1}_{j'})^2]\ . \tag{5}$$

Depending on the regularization strength $\gamma \geq 0$, the more the bags are similar ($v_{j,j'}$ higher), the more their relative estimates are close each other. We can restate the problem in matrix form by the Laplacian of the symmetric matrix $V \in \mathbb{R}^{n\times n}$, which is the adjacency matrix of the graph induced by the similarity $v_{j,j'}$. The Laplacian is defined as $L_a = D - V$, and D is a diagonal matrix such that $D_j = \sum_{j'}^{N} v_{j,j'}$. For any vector $u \in \mathbb{R}^n$, it holds that

$$\begin{aligned} u^\top L_a u &= u^\top D u - u^\top V u \\ &= \sum_j D_j u_j^2 - \sum_{j,j'} v_{j,j'} u_j u_{j'} \\ &= \frac{1}{2}\Big(\sum_j D_j u_j^2 - 2\sum_{j,j'} v_{j,j'} u_j u_{j'} + \sum_{j'} D_{j'} u_{j'}^2\Big) \\ &= \sum_{j,j'} v_{j,j'}(u_j - u_{j'})^2 \end{aligned} \tag{6}$$

Expression (6) is obtained by applying the definition of D_j. This is a standard result in spectral graph theory; see for example Von Luxburg (2007). We make use of this equivalence for both $\boldsymbol{b}^1_j$ and $\boldsymbol{b}^{-1}_j$, but separately, as the two vectors are subject to distinct constraints. Thanks to the structure of matrix $\tilde{\mathrm{B}}^{\pm}$, we can define

$$\mathrm{L} \doteq \varepsilon \mathrm{I} + \left[\begin{array}{c|c} \mathrm{L}_a & 0 \\ 0 & \mathrm{L}_a \end{array}\right] \in \mathbb{R}^{2n\times 2n}\ , \tag{7}$$

(with $\epsilon > 0$ to assure non-singularity and numerical stability, see Patrini et al. (2014)), such that

$$\tilde{\mathrm{B}}^{\pm} = \underset{\mathrm{X}\in\mathbb{R}^{2n\times d}}{\operatorname{argmin}} \ \operatorname{tr}\left((\mathrm{B}^\top - \mathrm{X}^\top\Pi)\mathrm{D}_{\boldsymbol{w}}(\mathrm{B} - \Pi^\top\mathrm{X})\right) + \gamma \operatorname{tr}\left(\mathrm{X}^\top \mathrm{L}\mathrm{X}\right)\ . \tag{8}$$

$\mathrm{D}_{\boldsymbol{w}} \doteq \textsc{Diag}(\boldsymbol{w})$ is a user-fixed bias matrix with non-negative element on the diagonal $\boldsymbol{w}$. For example, we can re-weight the importance of the linear equations by the size of the respective bags m_j, by $\boldsymbol{w} = \hat{\boldsymbol{p}}$. The second term assume the form of a manifold regularizer of Belkin et al. (2006), which allows us to reinterpret our assumption from a geometrical standpoint: the bag label-wise feature averages, the $\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|y,j]$s, live on a low-dimensional manifold parametrized by the unidimensional similarity function $v_{j,j'}$. The Laplacian matrix L_a is an empirical approximation of this manifold.

Problem (8) admits global optimum in closed form:

$$\tilde{\mathrm{B}}^{\pm} = \left(\Pi \mathrm{D}_{\boldsymbol{w}} \Pi^{\top} + \gamma \mathrm{L}\right)^{-1} \Pi \mathrm{D}_{\boldsymbol{w}} \mathrm{B} \ .$$

The size of the Laplacian is $O(n^2)$, which is small compared to $O(m^2)$ if there are not many bags. This is in contrast with traditional approaches for semi-supervised learning, where the Laplacian matrix is formed on top of similarity between examples, instead of bags.

We discuss the choice of V in the experimental section below [TBA]. Next, we state the Laplacian Mean Map algorithm. The last step of convex optimization for logistic regression is expressed with L_2 regularization.

Algorithm 1 Laplacian Mean Map (LMM)

Input $\mathcal{S}_j, \hat{\pi}_j, \hat{p}_j, j \in [n]$; $\gamma > 0$; $\boldsymbol{w}$; V; $\lambda > 0$;
Step 1 : let $\tilde{\mathrm{B}}^{\pm} \leftarrow \left(\Pi \mathrm{D}_{\boldsymbol{w}} \Pi^{\top} + \gamma \mathrm{L}\right)^{-1} \Pi \mathrm{D}_{\boldsymbol{w}} \mathrm{B}$;
Step 2 : let $\tilde{\boldsymbol{\mu}}_{\mathcal{S}} \leftarrow \frac{1}{m} \sum_j \hat{p}_j (\hat{\pi}_j \tilde{\boldsymbol{b}}_j^1 - (1 - \hat{\pi}_j) \tilde{\boldsymbol{b}}_j^{-1})$;
Step 3 : let $\tilde{\boldsymbol{\theta}}_* \leftarrow \mathrm{argmin}_{\boldsymbol{\theta}} \frac{1}{m} \sum_{i=1}^m g(\theta|\boldsymbol{x}_i) - \boldsymbol{\theta}^{\top} \tilde{\boldsymbol{\mu}}_{\mathcal{S}} + \lambda \|\boldsymbol{\theta}\|_2^2$;
Return $\tilde{\boldsymbol{\theta}}^*$

A Link to Proper Losses

Patrini et al. (2014) proved that the "mean operator trick", *i.e.* the decomposition of the loss in label-dependent and label-independent terms, is a general property for a set of losses called Symmetric Proper Losses; logistic and square loss belong to the set, see Nock and Nielsen (2009). Here we have shown the trick in the case of maximum likelihood estimation for the conditional exponential family, following Quadrianto et al. (2009). Indeed, it turns out the cost function of problem of (1) is equivalent the empirical risk built on logistic loss, as we prove here.

$$\begin{aligned}
&\sum_{i=1}^m \log \sum_{\sigma \in \{-1,1\}} \exp(\sigma \boldsymbol{\theta}^{\top} \boldsymbol{x}) - \boldsymbol{\theta}^{\top} \sum_{i=1}^m y_i \boldsymbol{x}_i \\
&= \sum_{i=1}^m \log \sum_{\sigma \in \{-1,1\}} \exp\left(\sigma \boldsymbol{\theta}^{\top} \boldsymbol{x}\right) - \sum_{i=1}^m \log \exp\left(y_i \boldsymbol{\theta}^{\top} \boldsymbol{x}_i\right) \\
&= \sum_{i=1}^m \log \left(\frac{\exp(\boldsymbol{\theta}^{\top} \boldsymbol{x}) + \exp(-\boldsymbol{\theta}^{\top} \boldsymbol{x})}{\exp(y_i \boldsymbol{\theta}^{\top} \boldsymbol{x}_i)} \right) \\
&= \sum_{i=1}^m \log \left(1 + \exp(-2 y_i \boldsymbol{\theta}^{\top} \boldsymbol{x})\right) \qquad (9) \\
&\propto \frac{1}{m} \sum_{i=1}^m \log \left(1 + \exp(-y_i \boldsymbol{\theta'}^{\top} \boldsymbol{x})\right) \qquad (10)
\end{aligned}$$

We obtain Step (9) by observing that the label always takes value in $\{-1, +1\}$. In the last step (10), we operate a change of variable by expressing the new model $\boldsymbol{\theta}' \leftarrow 2\boldsymbol{\theta}$, and rescale the cost by $1/m$ to obtain the empirical risk for the logistic loss in Table (1) from Patrini et al. (2014).

References

Belkin, M., Niyogi, P., and Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7:2399–2434.

Nock, R. and Nielsen, F. (2009). Bregman divergences and surrogates for learning. *IEEE Trans.PAMI*, 31:2048–2059.

Patrini, G., Nock, R., Rivera, P., and Caetano, T. (2014). (Almost) no label no cry. In *NIPS*27*.

Quadrianto, N., Smola, A.-J., Caetano, T.-S., and Le, Q.-V. (2009). Estimating labels from label proportions. *JMLR*, 10:2349–2374.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.