

Mathematics behind log-linear models

Thomas Lavergne

1) Starting from the ground

We will first study most of the theory about structured log-linear model from the most simple case: models without any structure. These models are commonly named logistic regression or maxent and try to predict a single output from some observations.

1.1. Our problem

The problem is formalized as follow, we have an input space \mathcal{X} and a discrete output space \mathcal{Y} and we want, for each element x of the input space, its probability to be associated to each element of the output space: $p(y|x)$. As we cannot get this from nowhere we are given a set of N sample of item from the input space with there true label from the output space noted: $(x^i, y^i) \in \mathcal{X} \times \mathcal{Y}$ for $1 \leq i \leq N$; which should be representative of the reality. Our true problem is to modelize $p(y|x)$ as it can be observed on these data.

1.2. Representing knowledge

1.3. Logistic regression framework

So, here we come to the logistic regression framework which will be the basis of all our study. For each feature we associate a parameter θ_k – who can be seen as the k -th component of a vector θ of size K – and use a model with an exponential distribution to estimate the probabilities:

$$p_{\theta}(y|x) = \frac{\exp(\theta^{\top} \mathbf{F}(x, y))}{Z_{\theta}(x)}.$$

And, for the moment, we will train the θ parameters by maximizing the conditional log-likelihood of the model over the training data:

$$\begin{aligned} \theta^* &= \operatorname{argmax}_{\theta} \mathcal{L}_{\tilde{p}}(p_{\theta}) \\ &= \operatorname{argmax}_{\theta} - \sum_{i=1}^N \log p_{\theta}(y^i|x^i) \end{aligned}$$

As we will prove in the next section, this model shape and optimization criterion can be justified by searching for the model with maximum entropy between all the models where the features expectation match the feature expectation over the training data.

2) Primal optimization problem

We start with our primal optimization problem which is quite simple. Find the distribution with maximum entropy in the space of good distributions. This lead to a maximization problem with one kind of constrain about being a good model and two kind of constrains for ensuring the model is a probability distribution.

$$\begin{array}{ll} \text{Find} & p_* = \operatorname{argmax}_p \mathbb{H}(p) \\ \text{Subject to} & \left\{ \begin{array}{l} \forall k, \quad \sum_{x,y} \tilde{p}(x)p(y|x)f_k(x, y) = \sum_{x,y} \tilde{p}(x, y)f_k(x, y) \\ \forall x, \quad \sum_y p(y|x) = 1 \\ \forall x, y, \quad p(y|x) \geq 0 \end{array} \right. \end{array}$$

Here we maximize the entropy of the model in order to find the most uniform distribution between all those who represent well our data. Intuitively, this is the ‘‘Occam’s razor’’ principle where we search the most simple solution to our problem. The entropy is defined by:

$$H(p) = - \sum_{x,y} \tilde{p}(x)p(y|x) \log p(y|x).$$

Unfortunately for us, this problem is untractable except for very trivial cases.

3) Magic with Lagrangian multiplier

3.1. The dual problem

In order to solve this problem, we use the theory of Lagrange multipliers. These allow us to build an unconstrained optimization problem equivalent to our first problem. The first step in this process is to associate a Lagrange multiplier to each constrains called θ_k for each constrains of the first type and η_x for the second kind. As we will see later, there is no need to include the last constraint about positivity in this problem. In fact, we will see that even the η_x multipliers will not be free parameters of the model. These multipliers allow us to build the following Lagrangian:

$$\begin{aligned} \Lambda(p, \theta) = & H(p) \\ & + \sum_k \theta_k \left(\sum_{x,y} \tilde{p}(x)p(y|x)f_k(x, y) - \sum_{x,y} \tilde{p}(x, y)f_k(x, y) \right) \\ & + \sum_x \eta_x \left(\sum_y p(y|x) - 1 \right) \end{aligned}$$

Holding the multipliers fixed, we can solve the Lagrangian for p . We denote this maximum by p_θ and the value of the Lagrangian at this maximum by $\Psi(\theta)$.

$$p_\theta = \operatorname{argmax}_p \Lambda(p, \theta)$$

$$\Psi(\theta) = \Lambda(p_\theta, \theta)$$

With this we can define our dual optimization problem which is, by the Kuhn-Tucker theorem, equivalent to the primal one defined in the previous section:

$$\text{Find } \theta^* = \operatorname{argmin}_\theta \Psi(\theta).$$

This is an unconstrained optimization problem a lot more tractable than the primal one. But, before we look more closely at it, we have to solve the Lagrangian to find p_θ and to substitute back this value in the Lagrangian in order to determine the dual function $\Psi(\theta)$ we have to optimize.

3.2. The model shape

Now we have reformulated our problem, the first things we have to do is to determine the model shape by maximizing the Lagrangian for $p(y|x)$ while holding the multipliers fixed. This will give us a model in terms of the Lagrangian multipliers.

$$\begin{aligned} p_\theta = & \operatorname{argmax}_p \Lambda(p, \theta) \\ = & \operatorname{argmax}_p - \sum_{x,y} \tilde{p}(x)p(y|x) \log p(y|x) \\ & + \sum_k \theta_k \left(\sum_{x,y} \tilde{p}(x)p(y|x)f_k(x, y) - \sum_{x,y} \tilde{p}(x, y)f_k(x, y) \right) \\ & + \sum_x \eta_x \left(\sum_y p(y|x) - 1 \right) \end{aligned}$$

To solve this problem, we take the derivative of the Lagrangian relative to each assignment to x and y :

$$\begin{aligned}\frac{\partial \Lambda(p, \theta)}{\partial p(y|x)} &= -\tilde{p}(x) (\log p(y|x) + 1) + \sum_k \theta_k \tilde{p}(x) f_k(x, y) + \eta_x \\ &= -\tilde{p}(x) (\log p(y|x) + 1) + \tilde{p}(x) \theta^\top \mathbf{F}(x, y) + \eta_x \\ &= -\tilde{p}(x) (\log p(y|x) + 1 - \theta^\top \mathbf{F}(x, y)) + \eta_x\end{aligned}$$

And it just remain for us to solve this in zero:

$$\begin{aligned}-\tilde{p}(x) (\log p(y|x) + 1 - \theta^\top \mathbf{F}(x, y)) + \eta_x &= 0 \\ -\tilde{p}(x) (\log p(y|x) + 1 - \theta^\top \mathbf{F}(x, y)) &= -\eta_x \\ \log p(y|x) + 1 - \theta^\top \mathbf{F}(x, y) &= \frac{\eta_x}{\tilde{p}(x)} \\ \log p(y|x) &= \frac{\eta_x}{\tilde{p}(x)} - 1 + \theta^\top \mathbf{F}(x, y) \\ p(y|x) &= \exp\left(\frac{\eta_x}{\tilde{p}(x)} - 1 + \theta^\top \mathbf{F}(x, y)\right) \\ p(y|x) &= \frac{\exp(\theta^\top \mathbf{F}(x, y))}{\exp\left(1 - \frac{\eta_x}{\tilde{p}(x)}\right)}\end{aligned}$$

The second constrain, transposed by third term in the lagrangian, tell us that all $p(y|x)$ have to sum to one. This allow us to solve the η_x multipliers:

$$Z_\theta(x) = \exp\left(1 - \frac{\eta_x}{\tilde{p}(x)}\right) = \sum_y \exp(\theta^\top \mathbf{F}(x, y))$$

And so, the final form of our model is, as expected, a log-linear model and this ensure that the positivity constrains we have first ignored will be satisfied implicitly by the model:

$$p(y|x) = \frac{\exp(\theta^\top \mathbf{F}(x, y))}{Z_\theta(x)}$$

3.3. Back to the Lagrangian

We can now go back to the lagrangian, remove the η_x constrains as they are not needed anymore because the model shape ensure they are satisfied, and put this expression of $p(y|x)$ in it. We only do the substitution at one place as the other will simplify as we will see:

$$\begin{aligned}\mathbb{H}(p) + \sum_k \theta_k \left(\sum_{x,y} \tilde{p}(x) p(y|x) f_k(x, y) - \sum_{x,y} \tilde{p}(x, y) f_k(x, y) \right) \\ = - \sum_{x,y} \tilde{p}(x) p(y|x) \log p(y|x) + \sum_k \theta_k \left(\sum_{x,y} \tilde{p}(x) p(y|x) f_k(x, y) - \sum_{x,y} \tilde{p}(x, y) f_k(x, y) \right) \\ = - \sum_{x,y} \tilde{p}(x) p(y|x) (\theta^\top \mathbf{F}(x, y) - \log Z_\theta(x)) + \sum_k \theta_k \left(\sum_{x,y} \tilde{p}(x) p(y|x) f_k(x, y) - \sum_{x,y} \tilde{p}(x, y) f_k(x, y) \right)\end{aligned}$$

And it remain to try to simplify this and check what it can be. We first expand and simplify the first term:

$$\begin{aligned}
&= - \sum_{x,y} \tilde{p}(x)p(y|x)\theta^\top \mathbf{F}(x,y) + \sum_{x,y} \tilde{p}(x)p(y|x) \log Z_\theta(x) \\
&\quad + \sum_k \theta_k \left(\sum_{x,y} \tilde{p}(x)p(y|x)f_k(x,y) - \sum_{x,y} \tilde{p}(x,y)f_k(x,y) \right) \\
&= - \sum_{x,y} \tilde{p}(x)p(y|x)\theta^\top \mathbf{F}(x,y) + \sum_x \tilde{p}(x) \log Z_\theta(x) \\
&\quad + \sum_k \theta_k \left(\sum_{x,y} \tilde{p}(x)p(y|x)f_k(x,y) - \sum_{x,y} \tilde{p}(x,y)f_k(x,y) \right)
\end{aligned}$$

Next, we develop and simplify the second part:

$$\begin{aligned}
&= - \sum_{x,y} \tilde{p}(x)p(y|x)\theta^\top \mathbf{F}(x,y) + \sum_x \tilde{p}(x) \log Z_\theta(x) \\
&\quad + \sum_k \theta_k \sum_{x,y} \tilde{p}(x)p(y|x)f_k(x,y) - \sum_k \theta_k \sum_{x,y} \tilde{p}(x,y)f_k(x,y) \\
&= - \sum_{x,y} \tilde{p}(x)p(y|x)\theta^\top \mathbf{F}(x,y) + \sum_x \tilde{p}(x) \log Z_\theta(x) \\
&\quad + \sum_{x,y} \tilde{p}(x)p(y|x)\theta^\top \mathbf{F}(x,y) - \sum_{x,y} \tilde{p}(x,y)\theta^\top \mathbf{F}(x,y)
\end{aligned}$$

The first and third terms cancel and it remains:

$$\begin{aligned}
\Psi(\theta) &= \sum_x \tilde{p}(x) \log Z_\theta(x) - \sum_{x,y} \tilde{p}(x,y)\theta^\top \mathbf{F}(x,y) \\
&= - \sum_{x,y} \tilde{p}(x,y) \log \frac{\exp(\theta^\top \mathbf{F}(x,y))}{Z_\theta(x)} \\
&= - \sum_{x,y} \tilde{p}(x,y) \log p_\theta(y|x) \\
&= \mathbf{H}(\tilde{p}, p_\theta)
\end{aligned}$$

So, our dual optimization problem is minimize the cross entropy between the empirical distribution and the model or:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathbf{H}(\tilde{p}, p_\theta)$$

3.4. Relation with the likelihood

An interesting thing to note is that the cross entropy is equal to the conditional log-likelihood times a constant:

$$\begin{aligned}
\mathbf{H}(\tilde{p}, p_\theta) &= - \sum_{i=1}^N \tilde{p}(x^i, y^i) \log p_\theta(y^i|x^i) \\
&= -N \sum_{i=1}^N \log p_\theta(y^i|x^i) \\
&= -N \cdot \mathcal{L}_{\tilde{p}}(p_\theta)
\end{aligned}$$

This imply that, instead of minimizing the cross entropy, we can maximize the conditional log-likelihood in order to train our model:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \mathcal{L}_{\tilde{p}}(p_\theta)$$

4) Solving the dual

Historically, the most classical way to solve this optimization problem was to use iterative scaling but the formulation in terms of an optimization problem shown before, suggest that classical optimization technics can be used.

Most of the classical optimization technics and eventually the ones best suited for training CRF require at least computing the gradient of the function to be optimized. We have seen that optimizing the cross entropy is equivalent to optimizing the conditional log-likelihood and the later is simpler so we choose it:

$$\begin{aligned}\mathcal{L}_{\bar{p}}(p_{\theta}) &= \sum_{i=1}^N \log p_{\theta}(y^i | x^i) \\ &= \sum_{i=1}^N \log \frac{\exp(\theta^{\top} \mathbf{F}(x^i, y^i))}{Z_{\theta}(x^i)} \\ &= \sum_{i=1}^N \left(\theta^{\top} \mathbf{F}(x^i, y^i) - \log Z_{\theta}(x^i) \right)\end{aligned}$$

We get its gradient by taking its derivative according to each coordinate of the Lagragian multipliers:

$$\begin{aligned}\frac{\partial \mathcal{L}_{\bar{p}}(p_{\theta})}{\partial \theta_k} &= \sum_{i=1}^N \left(f_k(x^i, y^i) - \sum_{y'} f_k(x^i, y') \frac{\exp(\theta^{\top} \mathbf{F}(x^i, y'))}{Z_{\theta}(x^i)} \right) \\ &= \sum_{i=1}^N \left(f_k(x^i, y^i) - \sum_{y'} f_k(x^i, y') p_{\theta}(y' | x^i) \right) \\ &= \sum_{i=1}^N \left(\mathbb{E}_{\bar{p}(y|x)} [f_k(x^i, y^i)] - \mathbb{E}_{p_{\theta}(y|x)} [f_k(x^i, y)] \right)\end{aligned}$$

An interesting thing to note is that this gradient will be all zero when the constrains of our first model are met, which is exactly what we want. The complexity of computing this gradient is $\mathcal{O}(NK|\mathcal{Y}|)$ which mean that for complex label set, whose cardinality can be exponential, all cases will not be tractable. For some very usefull cases like chains or tree, some dynamic programming can be used to make them efficient. For more general case, it is possible to fallback to some heuristics algorithms like belief propagation.

In order to use second order optimization technics, we also need to compute the hessian of the conditional

log-likelihood. Starting from the gradient we just computed, we derive it a second time:

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\bar{p}}(p_{\theta})}{\partial \theta_k} &= \sum_{i=1}^N \left(f_k(x^i, y^i) - \frac{1}{Z_{\theta}(x^i)} \sum_{y'} f_k(x^i, y') \exp(\theta^{\top} \mathbf{F}(x^i, y')) \right) \\
\frac{\partial^2 \mathcal{L}_{\bar{p}}(p_{\theta})}{\partial \theta_k \partial \theta_l} &= - \sum_{i=1}^N \frac{1}{Z_{\theta}(x^i)^2} \left[Z_{\theta}(x^i) \sum_{y'} f_k(x^i, y') f_l(x^i, y') \exp(\theta^{\top} \mathbf{F}(x^i, y')) \right. \\
&\quad \left. - \sum_{y'} f_k(x^i, y') \exp(\theta^{\top} \mathbf{F}(x^i, y')) \sum_{y'} f_l(x^i, y') \exp(\theta^{\top} \mathbf{F}(x^i, y')) \right] \\
&= - \sum_{i=1}^N \left[\sum_{y'} \frac{f_k(x^i, y') f_l(x^i, y') \exp(\theta^{\top} \mathbf{F}(x^i, y'))}{Z_{\theta}(x^i)} \right. \\
&\quad \left. - \sum_{y'} \frac{f_l(x^i, y') \exp(\theta^{\top} \mathbf{F}(x^i, y'))}{Z_{\theta}(x^i)} \sum_{y'} \frac{f_k(x^i, y') \exp(\theta^{\top} \mathbf{F}(x^i, y'))}{Z_{\theta}(x^i)} \right] \\
&= - \sum_{i=1}^N \mathbf{E}_{p_{\theta}(y|x)} [f_k(x^i, y) f_l(x^i, y)] - \mathbf{E}_{p_{\theta}(y|x)} [f_k(x^i, y)] \mathbf{E}_{p_{\theta}(y|x)} [f_l(x^i, y)] \\
&= - \sum_{i=1}^N \text{cov}_{p_{\theta}(y|x)} [f_k(x^i, y), f_l(x^i, y)]
\end{aligned}$$

As shown, the hessian is the negated covariance matrix. Covariance matrix are always at least positive semi-definite, so is our hessian. True second-order method are impractical in the general case as the hessian is too big to be computed but quasi-Newton methods, like L-BFGS, can be used instead.