CONDITIONAL RANDOM FIELDS

JOHN THICKSTUN

LOGISTIC REGRESSION

CRFs can be seen as a generalization of logistic regression. So we will begin by reviewing logistic regression. This is the simplest example of a "log-linear model" where the log-odds of the probability of a binary label $y \in \{0,1\}$ are a linear function of the data $x \in \mathbb{R}^d$:

$$\operatorname{logit}(P(y=1|x)) = \log \frac{P(y=1|x)}{P(y=0|x)} = \log \frac{P(y=1|x)}{1-P(y=1|x)} = \beta^T x = \beta_0 + \sum_{k=1}^{d} \beta_k x_k.$$

Rearranging terms, we find that the probability that $y=1$ is given by the sigmoid function

$$P(y=1|x) = \frac{1}{1+\exp(-\beta^T x)} = S(\beta^T x).$$

The logistic regression game is to find the MLE of the weights. The likelihood for n independent observations $X = (X_1, \ldots, X_n)^T$ is

$$L_X(\beta) = \prod_{i=1}^{n} P(Y_i = 1|X_i)^{Y_i} P(Y_i = 0|X_i)^{1-Y_i}.$$

And the log-likelihood is

$$\mathcal{L}_X(\beta) = \log L_X(\beta) = \sum_{i=1}^{n} \left(Y_i \log P(Y_i = 1|X_i) + (1-Y_i) \log P(Y_i = 0|X_i)\right).$$

Proposition.

$$\nabla \mathcal{L}_X(\beta) = X^T(Y - S(X\beta)).$$

Proof. Note that $P(Y_i = 0|X_i) = 1 - P(Y_i = 1|X_i)$ so the log-likelihood can be rewritten:

$$\mathcal{L}_X(\beta) = \sum_{i=1}^{n} \big(Y_i \log S(\beta^T X_i) + (1-Y_i)\log(1 - S(\beta^T X_i))\big)$$

Calculus verifies that $\frac{d}{dz} S(z) = S(z)(1-S(z))$ and

$$\nabla \mathcal{L}_X(\beta) = \sum_{i=1}^{n} \left(Y_i \frac{\nabla S(\beta^T X_i)}{S(\beta^T X_i)} - (1-Y_i) \frac{\nabla S(\beta^T X_i)}{1 - S(\beta^T X_i)} \right)$$

$$= \sum_{i=1}^{n} \left(Y_i \frac{S(\beta^T X_i)(1-S(\beta^T X_i))}{S(\beta^T X_i)} X_i - (1-Y_i) \frac{S(\beta^T X_i)(1-S(\beta^T X_i))}{1-S(\beta^T X_i)} X_i \right)$$

$$= \sum_{i=1}^{n} \left(Y_i(1 - S(\beta^T X_i)) - (1 - Y_i)S(\beta^T X_i) \right) X_i = \sum_{i=1}^{n} \left(Y_i - S(\beta^T X_i) \right) X_i. \qquad \square$$

We can proceed from here with first-order gradient updates:

$$\beta' = \beta + t\nabla\mathcal{L}_X(\beta) = \beta + tX^T(Y - S(X\beta)).$$

As an aside, if we observe that $\mathbb{E}Y_i = P(Y_i = 1|X_i) = S(\beta^T X_i)$ then the first order optimality condition can be written

$$X^T Y = \mathbb{E}_\beta X^T Y.$$

I.e. logistic regression is the distribution in the class of log-linear models satisfying the constraint that the empirical expectation of each feature matches the model distribution.

Alternatively, we can do something second-order.

Proposition. *Let* $W = \operatorname{diag}(S(\beta^T X_i)(1 - S(\beta^T X_i)))$. *Then*

$$\nabla^2\mathcal{L}_X(\beta) = X^T W X.$$

Proof.

$$\nabla^2\mathcal{L}_X(\beta) = -\sum_{i=1}^{n} X_i^T \nabla S(\beta^T X_i) = -\sum_{i=1}^{n} X_i^T S(\beta^T X_i)(1 - S(\beta^T X_i))X_i$$

$$= X^T \operatorname{diag}(S(\beta^T X_i)(1 - S(\beta^T X_i)))X = X^T W X. \qquad \square$$

Note that $W \succ 0$ so the logistic loss is convex. This guarantees the convergence of both the first and second order methods to the MLE. In particular, the Newton updates are

$$\beta' = \beta - (X^T W X)^{-1}\nabla\mathcal{L}_X(\beta) = \beta - (X^T W X)^{-1} X^T(Y - S(X\beta)).$$

Compare this with least squares regression. In that case we want to optimize $y = X\beta$ and the closed form solution is

$$\beta = (X^T X)^{-1} X^T Y.$$

For this reason, logistic regression is sometimes considered as an iteratively re-updated weighted least squares regression.

MULTINOMIAL LOGISTIC REGRESSION

We will now generalize the model from the previous section to a multiclass setting. For k classes, we can generalize the log-odds characterization of logistic regression by expressing the log-odds of each class $j > 1$ versus class 1 as a log-linear function of the data:

$$\operatorname{logit}(P(y = j)|x) = \log \frac{P(y = j|x)}{P(y = 1|x)} = \beta_j^T x.$$

Because probabilities sum to one, we must have

$$P(y = 1|x) = \frac{1}{1 + \sum_{j=2}^{k} e^{\beta_j^T x}}.$$

And a little algebra shows that for $j > 1$,

$$P(y = j|x) = \frac{e^{\beta_j^T x}}{1 + \sum_{j=2}^k e^{\beta_j^T x}}.$$

We can symmetrize the model by reparameterizing; observe that

$$P(y = 1|x) = \frac{e^{\beta_1^T x}}{e^{\beta_1^T x} + \sum_{j=2}^k e^{(\beta_1+\beta_j)^T x}} = \frac{e^{\beta_1^T x}}{\sum_{j=1}^k e^{\beta_j'^T x}}.$$

Therefore replacing β' with β, for all j

$$P(y = j|x) = \frac{e^{\beta_j^T x}}{\sum_{j'=1}^k e^{\beta_{j'}^T x}} = \frac{1}{Z_\beta(x)} e^{\beta_j^T x}.$$

This is the softmax representation of multinomial logistic regression. The log-likelihood for n independent observations is $\mathcal{L}_X : \mathbb{R}^{k \times d} \to \mathbb{R}$ defined by

$$\mathcal{L}_X(\beta) = \log L_X(\beta) = \sum_{i=1}^n \sum_{j=1}^k \log P(Y_i = j|X_i) \mathbb{1}_{Y_i=j}$$

$$= \sum_{i=1}^n \sum_{j=1}^k \left(\beta_j^T X_i - \log Z_\beta(X_i)\right) \mathbb{1}_{Y_i=j} = \sum_{i=1}^n \left(\beta_{Y_i}^T X_i - \log Z_\beta(X_i)\right).$$

Proposition.

$$\nabla_j \mathcal{L}_X(\beta) = X^T(\mathbb{1}_{Y=j} - P(Y = j|X))$$

Proof. The derivative of the log-partition function is

$$\frac{\partial \log Z_\beta(x)}{\partial \beta_{jl}} = \frac{1}{Z_\beta(x)} \sum_{j'=1}^k x_l e^{\beta_{j'}^T x} \mathbb{1}_{j'=j} = \frac{1}{Z_\beta(x)} x_l e^{\beta_j^T x} = x_l P(y = j|x).$$

And therefore the derivative of the log-likelihood is

$$\frac{\partial \mathcal{L}_X(\beta)}{\partial \beta_{jl}} = \sum_{i=1}^n \left(X_{il} \mathbb{1}_{Y_i=j} - X_{il} P(Y_i = j|X_i)\right) = \sum_{i=1}^n \left(\mathbb{1}_{Y_i=j} - P(Y_i = j|X_i)\right) X_{il}.$$

We can rewrite this as

$$\nabla_j \mathcal{L}_X(\beta) = \sum_{i=1}^n (\mathbb{1}_{Y_i=j} - P(Y_i = j|X_i)) X_i. \qquad \square$$

This gives us the the first order optimality conditions

$$X^T \mathbb{1}_{Y=j} = X^T P(Y = j|X) = X^T \mathbb{E}_\beta \mathbb{1}_{Y=j}.$$

I.e. the empirical marginal probability of each feature matches the distribution probability under the distribution parameterized by optimal β.

CONDITIONAL RANDOM FIELDS

Define a (log-linear) potential over observations $x \in \mathcal{X}$ and labelings $y \in \mathcal{Y}$ by

$$\psi_t(y,x) = \exp\left(\beta^T \varphi_t(y,x)\right)$$

Here φ_t is a feature map. We can think of ψ_t as an un-normalized probability distribution and consider the product distribution

$$\Psi(y,x) = \prod_{t=1}^{T} \psi_t(y,x).$$

The chain-structured CRF is then given by

$$P(Y=y|x) = \frac{1}{Z(x,\beta)}\Psi(y,x) = \frac{1}{Z(x,\beta)}\prod_{t=1}^{T}\psi_t(y,x)$$

$$= \frac{1}{Z(x,\beta)}\exp\left(\beta^T\sum_{t=1}^{T}\varphi_t(y,x)\right) = \frac{1}{Z(x,\beta)}\exp\left(\beta^T\Phi(y,x)\right).$$

Above we define

$$\Phi(y,x) = \sum_{t=1}^{T}\varphi_t(y,x).$$

And the partition function is

$$Z(x,\beta) = \sum_y \Psi(y,x) = \sum_y \exp\left(\beta^T\Phi(y,x)\right),$$

TRAINING

Given supervised training sequences (x^i, y^i), the log-likelihood is

$$L(\beta) = \frac{1}{n}\sum_{i=1}^{n}\log p(y^i|x^i) = \frac{1}{n}\sum_{i=1}^{n}\left(\beta^T\Phi(y^i,x^i) - \log Z(x^i,\beta)\right).$$

And the first order optimality conditions are

$$0 = \frac{\partial L}{\partial \beta_k} = \frac{1}{n}\sum_{i=1}^{n}\left(\Phi_k(y^i,x^i) - \frac{\partial}{\partial\beta_k}\log Z(x^i,\beta)\right).$$

Where

$$\frac{\partial}{\partial\beta}\log Z(x^i,\beta) = \frac{1}{Z(x^i,\beta)}\sum_{y'}\Phi_k(y',x^i)\exp\left(\beta^T\Phi(y',x^i)\right) = \mathbb{E}_\beta\Phi_k(Y,x^i).$$

So the first order condition reduces to

$$\sum_{i=1}^{n}\Phi(y^i,x^i) = \mathbb{E}_\beta\Phi(Y,x^i).$$

This is the maximum entropy condition: find the weights β that make the distribution expectation match the empirical expectation. We can find the optimal weights with first-order gradient updates:

$$\beta' = \beta + t\nabla L(\beta) = \beta + t\left(\frac{1}{n}\sum_{i=1}^{n}\Phi(y^i, x^i) - \mathbb{E}_\beta \Phi(Y, x^i)\right).$$

How do we compute $\mathbb{E}_\beta \Phi_k(Y, x^i)$? This is intractable in general because we need to sum over all possible sequences y'. But if we impose additional structure on our feature map, the computation may simplify.

logistic regression revisited. For example, suppose there is some φ' such that

$$\varphi_t(y, x) = \varphi'(y_t, x, t).$$

That is, each φ_t depends only on y_t, rather than the whole label space. In this case the expectation reduces to

$$\mathbb{E}_\beta \Phi(Y, x) = \sum_y \sum_{t=1}^{T} \varphi'(y_t, x, t) P(Y = y|x) = \sum_{t=1}^{T}\sum_{y_t} \varphi'(y_t, x, t) \sum_{\substack{y_s \\ s\neq t}} P(Y = y|x).$$

The marginal probability is

$$P(Y_t = y_t|x) = \sum_{\substack{y_s \\ s\neq t}} p(y|x) = \frac{1}{Z(x,\beta)} \sum_{\substack{y_s \\ s\neq t}} \prod_{s=1}^{T} \psi_s(y, x) = \frac{\psi_t(y, x)}{Z(x, \beta)} \sum_{\substack{y_s \\ s\neq t}} \prod_{s\neq t} \psi_s(y, x).$$

And the product reduces to

$$\prod_{s\neq t} \psi_s(y, x) = \prod_{s\neq t} \exp\left(\beta^T \varphi'(y_s, x, s)\right) = \exp\left(\beta^T \sum_{s\neq t} \varphi'(y_s, x, s)\right).$$

For example, if $\varphi'(y_s, x, s) = x_s y_s$ and $y_s \in \{0, 1\}$ then we recover a simple logistic regression model:

$$x^T y = \Phi(y, x) = \mathbb{E}_\beta \Phi(Y, x) = \sum_{y'} x^T y' P(Y = y'|x) = x^T \mathbb{E}_\beta Y.$$

In particular, the partition function factors:

$$Z(x, \beta) = \prod_{t=1}^{T} \sum_{y_t} \exp\left(\beta^T x_t y_t\right) = \prod_{t=1}^{T} \left(1 + \exp(\beta^T x_t)\right).$$

And the expectation $\mathbb{E}_\beta Y_t = p(Y_t = 1|x)$ can be written as

$$\frac{\exp\left(\beta^T x_t\right)}{Z(x, \beta)} \sum_{\substack{y_s \\ s\neq t}} \prod_{s\neq t} \exp(\beta^T x_s y_s) = S(\beta^T x_t) \sum_{\substack{y_s \\ s\neq t}} \prod_{s\neq t} \frac{\exp(\beta^T x_s y_s)}{1 + \exp(\beta^T x_s)} = S(\beta^T x_t).$$

quadratic interactions. We get a much richer model when we allow interaction terms between the labels y_t but there is a delicate balance between the richness of these interactions and tractable computations. In the case of pairwise interactions, the computations remain tractable. Here we let

$$\varphi_t(y,x) = \varphi''(y_t, y_{t-1}, x, t).$$

By marginalization,

$$\begin{aligned}
\mathbb{E}_\beta \Phi(Y,x) &= \sum_{t=1}^T \sum_{y_t, y_{t-1}} \varphi(y_t, y_{t-1}, x) \sum_{\substack{y_j \\ j \notin \{t-1,t\}}} P(Y=y|x) \\
&= \sum_{t=1}^T \sum_{y_t, y_{t-1}} \varphi(y_t, y_{t-1}, x) P(Y_t = y_t, Y_{t-1} = y_{t-1}|x).
\end{aligned}$$

The probabilities $P(Y_t = y_t, Y_{t-1} = y_{t-1}|x)$ can be computed by dynamic programming, known in this context as the forward-backward algorithm. From definitions and algebra,

$$\begin{aligned}
P(Y_t = y_t, Y_{t-1} = y_{t-1}|x) = \sum_{\substack{y_j \\ j \notin \{t-1,t\}}} P(Y=y|x) = \frac{1}{Z(x,\beta)} \sum_{\substack{y_j \\ j \notin \{t-1,t\}}} \prod_{s=1}^T \psi(y_s, y_{s-1}, x) \\
= \frac{1}{Z(x,\beta)} \psi(y_t, y_{t-1}, x) \left(\sum_{\substack{y_j \\ j < t-1}} \prod_{s=1}^{t-1} \psi(y_s, y_{s-1}, x) \right) \left(\sum_{\substack{y_j \\ j > t}} \prod_{s=t+1}^{T} \psi(y_s, y_{s-1}, x) \right).
\end{aligned}$$

The latter two terms can be calculated recursively; define

$$\alpha_t(y_t) = \sum_{\substack{y_j \\ j < t}} \prod_{s=1}^t \psi(y_s, y_{s-1}, x), \qquad \gamma_t(y_t) = \sum_{\substack{y_j \\ j > t}} \prod_{s=t+1}^T \psi(y_s, y_{s-1}, x).$$

If we define $\alpha_{-1}(y_t) = 1$ and $\gamma_{T+1}(y_t) = 1$ then

$$\alpha_t(y) = \sum_{y'} \psi(y, y', x) \alpha_{t-1}(y'), \qquad \gamma_t(y) = \sum_{y'} \psi(y', y, x) \gamma_{t+1}(y').$$

To summarize,

$$P(Y_t = y_t, Y_{t-1} = y_{t-1}|x) = \frac{1}{Z(x,\beta)} \alpha_{t-1}(y_{t-1}) \psi(y_t, y_{t-1}, x) \gamma_t(y_t).$$

And because probabilities sum to one, the normalizing constant is

$$Z(x,\beta) = \sum_y \prod_{t=1}^T \psi(y_t, y_{t-1}, x) = \sum_y \alpha_T(y).$$