The Exponential Family of Distributions and Logistic Regression

Guy Lebanon

The exponential family of distribution is the most important class of distributions in statistics. It is parameterized by a vector $\theta \in \mathbb{R}^d$ and assuming $X \in \mathcal{X}$ is defined using feature functions $f_i : \mathcal{X} \to \mathbb{R}$, $i = 1, \ldots, d$ as

$$p_\theta(X) = \frac{1}{Z(\theta)} h(X) \exp\left(\theta^\top f(X)\right) = h(X) \exp\left(\theta^\top f(X) - \log Z(\theta)\right) \tag{1}$$

where $f(X) = (f_1(X), \ldots, f_d(X))^\top$ is a column vector of feature function values and $Z(\theta)$ ensures normalization

$$Z(\theta) = \sum_{X \in \mathcal{X}} h(X) \exp(\theta^\top f(X)) \tag{2}$$

(the sum above is over all possible values of X i.e., over $\mathcal{X}$, and should be replaced with an integral if $\mathcal{X}$ is continuous rather than discrete). The distribution is thus determined via θ (the parameter vector), f (set of d feature functions), and h (carrier density). The term Z is determined automatically from θ, h, f in order to ensure normalization (note that it does not depend on X but it does depend on θ).

In practice, one determines the carrier density h as a reasonable guess of p (h should be high for highly probable X and low otherwise), from example using domain knowledge or some other estimation procedure that was done beforehand. The features f are chosen to be patterns or measurements that may or may not be useful in order to distinguish between highly probable and less probable values of X. After defining h, f the parameter θ is determined using a maximum likelihood estimator.

The representation above (1) is sometimes written in a more general form where θ_i is replaced by $g_i(\theta)$ i.e., the features are multiplied by features of the parameter vector rather than the parameter vector itself. But as this amounts to re-parameterization we can always define $\theta'_i = g_i(\theta)$ and proceed using (1) without loss of generality.

Note that the feature functions $f_1, \ldots, f_d$ may be as simple as $f_i(X) = X_i$ but in general may be nonlinear in the dimensions for example, $f_1(X) = X_1, f_2(X) = X_2, f_3(X) = X_1 X_2, f_4(X) = X_1^2, f_5(X) = X_2^2$, $f_6(X) = \log(X_1)$, etc. This way the number of features and the number of parameters d may be very high even if X is of low dimensionality. The resulting model is very powerful at modelling arbitrary distributions - assuming that enough (and the right) features were defined. It is typically better to select too many features than too little as insignificant features will be assigned a parameter 0 using the maximum likelihood estimation (assuming we have enough training data).

The exponential family includes as special cases many well known distributions such as the multivariate normal, the Poisson, Dirichlet, multinomial (which includes binomial and Bernoulli), beta, and gamma distributions. For example the univariate normal may be written as

$$p_{\mu,\sigma^2}(X) = \exp\left(-X^2/2\sigma^2 + 2X\mu/\sigma^2 - \mu^2 2/\sigma^2 - \log\sqrt{2\pi\sigma^2}\right) = \exp\left(\theta_1 f_1(X) + \theta_2 f_2(X) - \log Z(\theta)\right)$$

where $f_1(X) = X$, $f_2(X) = X^2$, $\theta_1 = 2\mu/\sigma^2$, $\theta_2 = -1/\sigma^2$ and Z is determined by (2). Notable distributions that cannot be represented by (2) are the uniform $U[0, \theta]$ and mixtures of Gaussians.

It is straightforward to derive that

$$\frac{\partial \log Z(\theta)}{\partial \theta_i} = \mathsf{E}_{p_\theta} f_i(X), \qquad \frac{\partial^2 \log Z(\theta)}{\partial \theta_i \partial \theta_j} = -\mathsf{Cov}_{p_\theta}(f_i(X), f_j(X)). \tag{3}$$

As a result the loglikelihood function

$$\ell(\theta) = -n \log Z(\theta) + \sum_{i=1}^{n} \theta^\top f(X^{(i)}) + \log h(X^{(i)})$$

has partial derivatives

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = \sum_{i=1}^{n} f_j(X^{(i)}) - n \mathsf{E}_{p_\theta}(f_j(X)), \qquad \frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j} = -n \mathsf{Cov}\left(f_i(X), f_j(X)\right).$$

which may be used with gradient descent for obtaining the maximum likelihood estimator $\hat{\theta}$. We can also observe from the equation above that the MLE $\hat{\theta}$ satisfies the following equalities between expectations and averages over the training set

$$\frac{1}{n} \sum_{i=1}^{n} f_j(X^{(i)}) = \mathsf{E}_{p_\theta}(f_j(X)) \qquad j = 1, \ldots, d. \tag{4}$$

This equality makese sense intuitively as the averages (left hand side) converge to their expectations (right hand side) as $n \to \infty$ due to the law of large numbers. A very useful property of the exponential family is that the loglikelihood is concave which ensures that there is at most a single maximum likelihood (no multiple local maxima). The concavity of $\ell(\theta)$ can be proved by showing that the matrix of second order derivatives of $\ell(\theta)$ (the Hessian) is negative definite[1] (see (3))

$$\begin{aligned} v^\top H(\theta) v &= -n v^\top \mathsf{Var}\left(f_1(X), \ldots, f_d(X)\right) v = -n v^\top \mathsf{E}\left((f(X) - \mathsf{E}\, f(X))(f(X) - \mathsf{E}\, f(X))^\top\right) v \\ &= -n \mathsf{E}\left(v^\top (f(X) - \mathsf{E}\, f(X))(f(X) - \mathsf{E}\, f(X))^\top v\right) = -n \mathsf{E}\left((f(X) - \mathsf{E}\, f(X))^\top v\right)^\top (f(X) - \mathsf{E}\, f(X))^\top v \\ &= -n \mathsf{E}\, W^\top W \leq 0. \end{aligned}$$

Logistic Regression

The discriminative modeling task of predicting Y given X is often done by assuming an exponential family for the random vector $Z = (X, Y)$ $p_\theta(X, Y) = h(X, Y) \exp(\theta^\top f(X, Y)^\top - \log Z(\theta))$ and obtaining an estimate of $p_\theta(Y|X) = p_\theta(X, Y) / \sum_Y p_\theta(X, Y) = p_\theta(X, Y)/Z(\theta, X)$ by maximizing the conditional loglikelihood $\hat{\theta} = \arg\max \sum_{i=1}^{n} p_\theta(Y^{(i)}|X^{(i)})$. The resulting $\hat{\theta}$ is plugged in to obtain $p_{\hat{\theta}}(Y|X)$ which is useful for deriving the optimal Bayes rule prediction (given X it predicts X that minimizes the Bayes risk).

The features $f_1, \ldots, f_d$ are in this case functions of both X and Y. In the case of logistic regression ($Y \in \{+1, -1\}$ we have $f_i(X, Y) = Y \tilde{f}_i(X)/2$ (for some features $\tilde{f}$ of X, that are often taken to be $\tilde{f}_i(X) = X_i$) to obtain

$$p_\theta(Y|X) = \frac{1}{Z(\theta, X)} \exp\left(Y \theta^\top \tilde{f}(X)/2\right) = \frac{e^{Y \theta^\top \tilde{f}(X)/2}}{e^{Y \theta^\top \tilde{f}(X)/2} + e^{-Y \theta^\top \tilde{f}(X)/2}} = \left(1 + e^{-Y \theta^\top \tilde{f}(X)}\right)^{-1}. \tag{5}$$

Logistic regression is sometimes written as $\log \frac{p_\theta(1|X)}{1 - p_\theta(1|X)} = \theta^\top \tilde{f}(X)$ which is equivalent to (5) (this may be seen by solving the equation $\log(p/(1-p)) = \theta^\top \tilde{f}(X)$ for $p = p_\theta(1|X)$). The estimate $\hat{\theta}$ obtained by maximizing the conditional loglikelihood

$$\hat{\theta} = \arg\max_\theta \sum_{i=1}^{n} \log p_\theta(Y^{(i)}|X^{(i)}) = \arg\min_\theta \sum_{i=1}^{n} \log(1 + e^{-Y^{(i)} \theta^\top \tilde{f}(X^{(i)})})$$

results in one of the best performing classifiers in practice (in many cases better than Fisher's LDA, naive Bayes, nearest neighbors, decision trees). It is linear in $\tilde{f}_1(X), \ldots, \tilde{f}_d(X)$ and the decision boundary is a linear hyperplane in that space. Its Bayes rule prediction in the case of the 0/1 loss is $\hat{Y} = \arg\max_y p_\theta(y|X)$.

[1] A matrix H is negative definite if for all vectors v, we have $v^\top H v < 0$. For symmetric matrices this is equivalent to all eigenvalues being negative.