Machine Learning Course - CS-433

Exp Families and GL Models

Oct 24, 2017

changes by Rüdiger Urbanke 2016

minor changes by Rüdiger Urbanke 2017

Last updated: October 24, 2017

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Motivation

The logistic function (probability distribution) makes it possible to apply linear regression to binary outputs. Can we apply a similar trick to other cases (e.g., $y \in \mathbb{N}$?
And can we generalize the maximum-likelihood procedure to a more general class of distributions $p(y|\mathbf{x}^\top \mathbf{w})$?
The answer is yes. We proceed as follows. We first introduce the "right" class of distributions. It is called the *exponential family*. We will spend some time to discuss and prove some basic properties of this family. We will then see how to construct a *generalized model* based on an element of the exponential distribution. Exponential families are important in their own right and their use in ML and other areas is far greater than the simple application to generalized models we will give. We will therefore spend quite some time on exploring some of their basic properties.

Logistic regression revisited

In logistic regression we used the distribution

$$p(y|\eta) = \frac{e^{\eta y}}{1 + e^{\eta}} = \exp\left[\eta y - \log(1 + e^{\eta})\right],$$

where we assumed that y takes on values in $\{0, 1\}$ and where we wrote η as a shorthand for $\mathbf{x}^\top \mathbf{w}$. As you can see, we rewrote this distribution in a specific form. Our next step will be to generalize this form.

Exponential family

Let y be a scalar and $\boldsymbol{\eta}$ be a vector. We will say that a distribution belongs to the *exponential family* if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\psi}(y) - A(\boldsymbol{\eta})\right]. \tag{1}$$

Note that $A(\boldsymbol{\eta})$ is the logarithm of a *normalization* term. It ensures that the expression forms a proper distribution. It is sometimes called the *cumulant*. We will see shortly that despite the fact that $A(\boldsymbol{\eta})$ is *only* there for normalization purposes it plays a crucial role and contains valuable information. In some other areas (statistical physics) this term also plays a central role and it is called the *log partition* function.

Note that the expression in (1) is non-negative if $h(y) \geq 0$. So we only need to ensure that it can be properly normalized, i.e., that

$$A(\boldsymbol{\eta}) = \ln[\int_y h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\psi}(y)\right] dy] < \infty. \tag{2}$$

We will only consider parameters $\boldsymbol{\eta}$ so that $A(\boldsymbol{\eta})$ is finite.

The representation in (1) is the so-called *canonical* form. There are even more general definitions but we will not need them.

The quantity $\boldsymbol{\psi}(y)$ is called a *sufficient statistics*.[1] Note that $\boldsymbol{\psi}(y)$ can be a vector in general.

[1] What this means is that if we want to estimate the parameter $\boldsymbol{\eta}$ given iid samples from this distribution then all the information regarding the true parameter $\boldsymbol{\eta}$ is contained in the vector of samples $\boldsymbol{\psi}(\mathbf{y})$.

If you look at the definition of the exponential family, you will see that we have several "degrees of freedom" to define an element of the family. We can choose the factor $h(y)$, we can choose the vector $\boldsymbol{\psi}(y)$, and we can choose the parameter $\boldsymbol{\eta}$. For every choice we will get an element of the exponential family. The term $A(\boldsymbol{\eta})$ is then determined for each such choice and ensures that the expression is properly normalized as dicussed.

Examples

Let us look at a few examples which are probably familiar to you but you might not have seen them written in this form. *Example:* We claim that the Bernoulli distribution is a member of the exponential family. We write

$$\begin{aligned} p(y|\mu) &= \mu^y (1-\mu)^{1-y}, \text{ where } \mu \in (0,1) \\ &= \exp\left[(\ln \frac{\mu}{1-\mu}) y + \ln(1-\mu) \right]. \end{aligned}$$

Mapping this to (1) we see that

$$\begin{aligned} \psi(y) &= y, \\ \eta &= \ln \frac{\mu}{1-\mu}, \\ A(\eta) &= -\ln(1-\mu) = \ln(1+e^{\eta}), \\ h(y) &= 1. \end{aligned}$$

In this case $\psi(y)$ is a scalar, reflecting the fact that this family only depends on a single parameter. In fact, we have a 1-1 relationship between η and μ,

$$\eta = g(\mu) = \ln\frac{\mu}{1-\mu} \iff \mu = g^{-1}(\eta) = \frac{e^{\eta}}{1+e^{\eta}}.$$

This function g is known as the *link* function (it links the mean of $\boldsymbol{\psi}(y)$ to the parameter η.)
Note that this is *exactly* the same distribution that we encountered when we discussed *logistic regression.*
Example: The Gaussian distribution with mean μ and variance σ^2 as parameters is also a member of the exponential family. We write

$$\begin{aligned} p(y|\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+ \\ &= \exp\left[(\mu/\sigma^2, -1/(2\sigma^2))(y, y^2)^\top - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right]. \end{aligned}$$

Mapping this again to (1) we see that

$$\begin{aligned} \boldsymbol{\psi}(y) &= (y, y^2) \\ \boldsymbol{\eta} &= (\eta_1 = \mu/\sigma^2, \eta_2 = -1/(2\sigma^2))^\top, \\ A(\boldsymbol{\eta}) &= \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\ln(2\pi\sigma^2), \\ &= -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-\eta_2/\pi), \\ h(y) &= 1. \end{aligned}$$

Note that this time $\boldsymbol{\psi}(y)$ is a vector of length two, reflecting the fact that the distribution depends on two parameters. In fact, we have the 1-1 relationship between $\boldsymbol{\eta} = (\eta_1, \eta_2)$

and (μ, σ^2).

$$\eta_1 = \frac{\mu}{\sigma^2}; \eta_2 = -\frac{1}{2\sigma^2} \iff \mu = -\frac{\eta_1}{2\eta_2}; \sigma^2 = -\frac{1}{2\eta_2}.$$

Example: Consider the Poisson distribution with mean μ. We have, for $y \in \mathbb{N}$,

$$\begin{aligned} p(y|\mu) &= \frac{\mu^y e^{-\mu}}{y!} \\ &= \frac{1}{y!} e^{y \ln(\mu) - \mu} \\ &= h(y) e^{y\theta - e^\theta}, \end{aligned}$$

where $h(y) = 1/y!$, $\phi(y) = y$, $\theta = g(\mu) = \ln(\mu)$, and $\mu = g^{-1}(\theta) = e^\theta$.

Some useful properties of the exponential family

Convexity of $A(\boldsymbol{\eta})$

Lemma. *The cumulant $A(\boldsymbol{\eta})$ is convex as a function of $\boldsymbol{\eta}$.*

Proof. Let $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ be two parameters. Define $\boldsymbol{\eta} = \lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2$. We start with (2) and apply Hoelder's inequality. Recall that Hoelder's inequality reads $\|fg\|_1 \leq \|f\|_p \|g\|_q$, where $p, q \in [1, \infty$ and $1/p + 1/q = 1$.

We get

$$
\begin{aligned}
&e^{A(\boldsymbol{\eta})} \\
=&\int_y h(y)\exp\left[\boldsymbol{\eta}^\top \boldsymbol{\psi}(y)\right]dy \\
=&\int_y [h(y)^\lambda \exp\left[\lambda \boldsymbol{\eta}_1^\top \boldsymbol{\psi}(y)\right]][h(y)^{1-\lambda}\exp\left[(1-\lambda)\boldsymbol{\eta}_2^\top \boldsymbol{\psi}(y)\right]]dy \\
\leq&(\int_y h(y)\exp\left[\boldsymbol{\eta}_1^\top \boldsymbol{\psi}(y)\right]dy)^\lambda(\int_y h(y)\exp\left[\boldsymbol{\eta}_2^\top \boldsymbol{\psi}(y)\right]dy)^{1-\lambda} \\
=&e^{\lambda A(\boldsymbol{\eta}_1)}e^{(1-\lambda)A(\boldsymbol{\eta}_2)}.
\end{aligned}
$$

Taking the log of this chain proves the claim,

$$A(\boldsymbol{\eta}) \leq \lambda A(\boldsymbol{\eta}_1) + (1-\lambda)A(\boldsymbol{\eta}_2).$$

□

Derivatives of $A(\boldsymbol{\eta})$ and moments

Another useful property is that the gradient and Hessian (first and second derivatives) of $A(\boldsymbol{\eta})$ are related to the mean and the variance of $\boldsymbol{\psi}(y)$.

Lemma.

$$
\begin{aligned}
\nabla A(\boldsymbol{\eta}) &= \mathbb{E}[\boldsymbol{\psi}(y)], \\
\nabla^2 A(\boldsymbol{\eta}) &= \mathbb{E}[\boldsymbol{\psi}(y)\boldsymbol{\psi}(y)^\top] - \mathbb{E}[\boldsymbol{\psi}(y)]\mathbb{E}[\boldsymbol{\psi}(y)^\top].
\end{aligned}
$$

Before we prove this, let us check this for our two running examples. Recall that for the Bernoulli distribution $\boldsymbol{\psi}(y)$ is a scalar, namely y. So in this case the first derivative should be the mean of the Bernoulli distribution and the second derivative the variance. Let us verify this. We get

$$\begin{aligned}\frac{dA(\eta)}{d\eta} &= \frac{d\ln(1+e^{\eta})}{d\eta} = \frac{e^{\eta}}{1+e^{\eta}} = \sigma(\eta) = \mu, \\ \frac{d^2A(\eta)}{d\eta^2} &= \frac{d\sigma(\eta)}{d\eta} = \sigma(\eta)(1-\sigma(\eta)) = \mu(1-\mu),\end{aligned}$$

which confirms the claim.
For the Gaussian distribution our vector $\boldsymbol{\psi}(y)$ is of the form (y, y^2). So the first derivative (gradient) should give us the mean and the scond moment of the Gaussian. The second derivative should give us the variance of various moments of y. We get

$$\begin{aligned}\frac{\partial A(\boldsymbol{\eta})}{d\eta_1} &= \frac{\partial(-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-\eta_2/\pi))}{\partial\eta_1} = -\frac{\eta_1}{2\eta_2} = \mu, \\ \frac{\partial A(\boldsymbol{\eta})}{d\eta_2} &= \frac{\partial(-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-\eta_2/\pi))}{\partial\eta_2} = (\frac{\eta_1^2 - 2\eta_2}{4\eta_2^2}) = \mu^2 + \sigma^2,\end{aligned}$$

which are exactly the expected value and the second moment of y, as claimed. To do one more computation, let us compute

$$\frac{\partial^2 A(\boldsymbol{\eta})}{d\eta_1^2} = \frac{\partial(-\frac{\eta_1}{2\eta_2})}{\partial\eta_1} = -\frac{1}{2\eta_2} = \sigma^2,$$

which is the variance of y, again as expected.

Proof. Let us just write down the proof regarding the first derivative. The proof for the second derivative proceeds in a similar fashion. We have

$$
\begin{aligned}
\nabla A(\boldsymbol{\eta}) &= \nabla \ln[\int_y h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\psi}(y)\right] dy] \\
&= \frac{\int_y \nabla h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\psi}(y)\right] dy}{\int_y h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\psi}(y)\right] dy} \\
&= \frac{\int_y h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\psi}(y)\right] \boldsymbol{\psi}(y) dy}{\exp(A(\boldsymbol{\eta}))} \\
&= \int_y h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\psi}(y) - A(\boldsymbol{\eta})\right] \boldsymbol{\psi}(y) dy \\
&= \mathbb{E}[\boldsymbol{\psi}(y)].
\end{aligned}
$$

In the second step we have exchange the derivative with the integral. Note that the exchange of differentiation and integration is permitted if the resulting integral is finite (which it is in our case).

□

Link function

As we have seen already in specific cases, there is a relationship between the "mean" $\boldsymbol{\mu} := \mathbb{E}[\boldsymbol{\psi}(y)]$ and $\boldsymbol{\eta}$ defined using a so-called *link function* $\mathbf{g}$.

$$
\boldsymbol{\eta} = \mathbf{g}(\boldsymbol{\mu}) \iff \boldsymbol{\mu} = \mathbf{g}^{-1}(\boldsymbol{\eta}).
$$

See the table of link functions and many other examples of exponential family in the KPM book chapter on "Generalized Linear Model".

We are typically interested in this relationship when $\psi(y)$ is a scalar, i.e., when the family has a single degree of freedom. In this case we can combine the link function with our previous observation regaring the derivative of $A(\eta)$ to get

$$\frac{dA(\eta)}{d\eta} = \mathbb{E}[\psi(y)] = g^{-1}(\eta).$$

Generalized linear models

It remains to discuss one important reason why we are considering this family of distributions. Given an element from this family, we can construct from this a data model by assuming that a sample $(\mathbf{x}, y)$ follows the distribution

$$p(y \mid \mathbf{x}, \mathbf{w}) = h(y) e^{\mathbf{x}^\top \mathbf{w} \psi(y) - A(\mathbf{x}^\top \mathbf{w})},$$

where $\psi(y)$ is assumed to be a *scalar*. We call such a model a *generalized linear model*. As we will now discuss, for such a model the maximum likelihood problem is particularly easy to solve.

Maximum likelihood for generalized linear models

Assume that we have given a training set S_t consisting of N iid samples $(\mathbf{x}_n, y_n)$. Assume further that we fit a generalized linear model to this data. This means that we assume that samples obey a distribution of the form

$$p(y_n \mid \mathbf{x}_n, \mathbf{w}) = h(y_n) e^{\eta_n \psi(y_n) - A(\eta_n)}$$

with $\eta_n = \mathbf{x}_n^\top \mathbf{w}$. Given S_t, we then write down the likelihood and look for that weight vector $\mathbf{w}$ that maximizes this likelihood.

In more detail, we consider the cost function

$$\begin{aligned}\mathcal{L}(\mathbf{w}) &= -\sum_{n=1}^{N} \ln p(y_n|\mathbf{x}_n^\top \mathbf{w}) \\ &= -\sum_{n=1}^{N} \ln(h(y_n)) + \mathbf{x}_n^\top \mathbf{w} \psi(y_n) - A(\mathbf{x}_n^\top \mathbf{w}).\end{aligned}$$

We want to minimize this cost function (we added a minus sign). Therefore, let us take the gradient of this expression,

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = -\sum_{n=1}^{N} \mathbf{x}_n \psi(y_n) - \mathbf{x}_n g^{-1}(\mathbf{x}_n^\top \mathbf{w}),$$

where in the last step we have use of the fact that $\frac{dA(\eta)}{d\eta} = g^{-1}(\eta)$.
If we set this equation to zero we get the condition of optimality. In particular, if we rewrite this sum by using our matrix notation we get

$$\nabla \mathcal{L}(\mathbf{w}) = \mathbf{X}^\top \left[g^{-1}(\mathbf{X}\mathbf{w}) - \psi(\mathbf{y}) \right] = 0,$$

where, as before, the scalar functions (g^{-1} and ψ) are applied to each vector component-wise.
To compare, for the case of the logistic regression we got the equation

$$\nabla \mathcal{L}(\mathbf{w}) = \mathbf{X}^\top \left[\sigma(\mathbf{X}\mathbf{w}) - \mathbf{y} \right] = 0.$$

As we have discussed, for the logistic case (Bernoulli distribution) we have the relationship $g^{-1} = \sigma$, which confirms that our previous derivation was just a special case.

Note also that we have already shown that $A(\mathbf{x}^\top \mathbf{w})$ is a convex function (A is convex and $A(\mathbf{x}^\top \mathbf{w})$ is the composition of a linear function with a convex function). Therefore $\boldsymbol{\mathcal{L}}(\mathbf{w})$ is convex (the other terms are constant or linear), just as we have seen this for the logistic regression. As a consequence, greedy iterative algorithms (like gradient descent) to find the optimum weight vector $\mathbf{w}$ are expected to work well in this context.

ToDo

1. Read the following sections in the KPM book: Section 9.2.1 to 9.2.4, Section 9.3.1 to 9.3.2.

2. Derive exponential family form for the multinomial distributions.

3. Derive the generalized linear model for regression with Poisson distribution for count data.