STA561: Probabilistic machine learning

Introduction: exponential family, conjugacy, and sufficiency (9/2/13)

Lecturer: Barbara Engelhardt *Scribes: Melissa Dalis, Abhinandan Nath, Abhishek Dubey, Xin Zhou*

1 Review

In the previous class, we discussed the maximum likelihood estimates (MLE) and maximum a posteriori (MAP) estimates. The general ways to calculate them are as follows

$$\begin{aligned} \hat{h}_{MAP} &= \arg\max_{h \in \mathcal{H}} P(h|\mathcal{D}) \\ \hat{h}_{MLE} &= \arg\max_{h \in \mathcal{H}} P(\mathcal{D}|h) \end{aligned}$$

1.1 How to calculate the MLE?

Assume we have the observed data sequence $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$, with X_i independent and identically distributed (IID) and $X_i \in \{0, 1\}$. We also model the X_is as Bernoulli, $X_i \sim \text{Ber}(\pi)$, with parameter $\pi \in [0, 1]$ (where π represents the probability that X_i is a 1). Thus the log likelihood for $\mathcal{D}$ is

$$\begin{aligned} \ell(h; \mathcal{D}) &= \log p(\mathcal{D}|h) \\ &= \log \prod_{i=1}^{n} \pi^{x_i} (1-\pi)^{1-x_i} \\ &= \sum_{i=1}^{n} [x_i \log \pi + (1 - x_i) \log(1 - \pi)] \end{aligned}$$

The reason we use the logarithm of the likelihood is to facilitate the calculation of the first derivative of the likelihood. The log likelihood is a concave function (see Figure 1). It will first increase as π increases and reach a maximum value and then reduce as π increases. The (global) maximum value indicates that the corresponding π maximizes the log likelihood of the data; this maximum value will occur at the location where the first derivative of the log likelihood with respect to π is equal to 0 (i.e., zero slope of the log likelihood function with respect to π). The logarithm function is monotonically increasing, so maximizing $\log(f(x))$ also maximizes $f(x)$.

1.2 Examples: Calculating the $\hat{h}_{MLE}$ for the Bernoulli Distribution

In order to get the maximum of the above concave function, we take the first derivative of $\ell(h; \mathcal{D})$ with respect to π, as:

$$\frac{\partial \ell(h; \mathcal{D})}{\partial \pi} = \frac{\sum_{i=1}^{n} x_i}{\pi} - \frac{\sum_{i=1}^{n} (1 - x_i)}{1 - \pi}$$

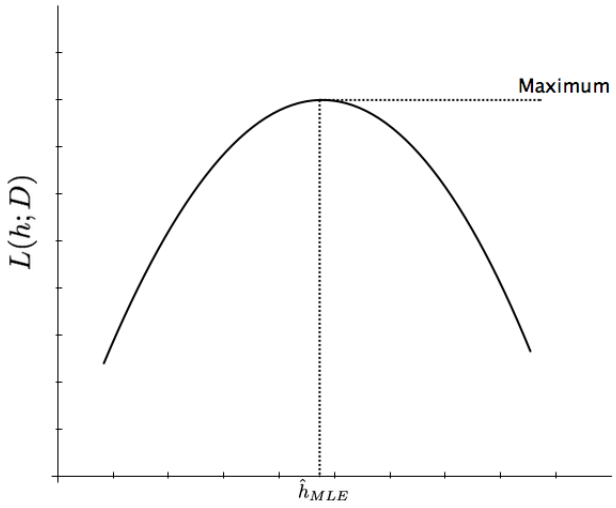

Figure 1: Example of a concave function, depicting the global maximum at the position where the slope is zero.

Then, set this to 0, we get

$$\begin{aligned}
\frac{1}{\pi}N_1 - \frac{1}{1-\pi}N_0 &= 0 \\
\frac{1-\pi}{\pi}N_1 &= N_0 \\
\frac{1}{\pi}N_1 - N_1 &= N_0 \\
N_1 + N_0 &= \frac{N_1}{\pi} \\
\hat{\pi}_{MLE} &= \frac{N_1}{N_1 + N_0}
\end{aligned}$$

where we use N_1 and N_0 as shorthand for the number of heads and the number of tails in the data set:

$$\begin{aligned}
N_1 &= \sum_{i=1}^{n} x_i \\
N_0 &= \sum_{i=1}^{n} (1 - x_i)
\end{aligned}$$

2 The Exponential Family

2.1 Why exponential family?

- The exponential family is the only family of distributions with finite-sized sufficient statistics, meaning that we can compress potentially very large amounts of data into a constant-sized summary without loss of information. This is particularly useful for online learning, where the observed data may become huge (e.g., your email inbox, where each email is a sample, and emails arrive in an ongoing way).

- The exponential family is the only family of distributions for which conjugate priors exist, which simplifies the computation of the posterior.
- They are the core of generalized linear models and variational methods, which we will learn about in this course.
- Expectations are simple to compute, as we will see today, making our life simple. This is part of the 'exponential family machinery' that we can exploit in our work to make the computation and mathematics simpler.

The exponential family includes many of the distributions we've seen already, including: normal, exponential, gamma, beta, Dirichlet, Bernoulli, Poisson, and many others. An important distribution that does not strictly belong to the exponential family is the uniform distribution.

2.2 Definition

A distribution is in the exponential family if its pdf or pmf $p(x|\eta)$, for $x = \{x_1, x_2, \ldots, x_n\} \in \mathcal{R}^n$ and $\eta \in \mathcal{H} \subseteq \mathbb{R}^d$ can be written in the following form:

$$p(x|\eta) = h(x) \exp\left\{\eta^T T(x) - A(\eta)\right\}$$

where:

$$\begin{aligned} \eta &= \text{natural parameters} \\ A(\eta) &= \text{log partition function} \\ T(x) &= \text{sufficient statistics} \\ h(x) &= \text{base measure} \\ \mu &= \text{mean parameter} \end{aligned}$$

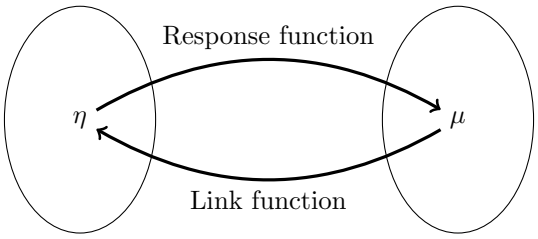

Figure 2: Link and Response functions

The relationship between μ and η is shown in Figure 2: the link function and the response function are invertible functions that allow a mapping between the mean parameters and the natural parameters. This is a one-to-one mapping (the reason it is one-to-one will become clear later). This fact is enables us to work in terms of either the the space of natural or mean parameters (whichever is mathematically most convenient), since converting between them is straightforward.

2.3 Examples

In this section, we will represent the Bernoulli and the Gaussian in the exponential family form. We will also derive the link and response functions.

2.3.1 Bernoulli distribution in the exponential family

We write the probability mass function for a Bernoulli random variable $x \sim Ber(\pi)$ in exponential form as below, where π is the mean parameter of the random variable X, e.g., the probability of a single coin flip coming up heads. The general way to derive this is to take the $\exp(\log(pdf()))$ of the pmf (or pdf) and organize the resulting variables to match the form of the exponential family distribution.

$$\begin{aligned} Ber(x|\pi) &= \pi^x(1-\pi)^{1-x} \\ &= \exp\left[\log(\pi^x(1-\pi)^{1-x})\right] \\ &= \exp\left[x\log\pi + (1-x)\log(1-\pi)\right] \\ &= \exp\left[x(\log\pi - \log(1-\pi)) + \log(1-\pi))\right] \\ &= \exp\left[x\log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right] \end{aligned}$$

On comparing the above formula with the exponential family form, we have

$$\begin{aligned} h(x) &= 1 \\ T(x) &= x \\ \eta &= \log\left(\frac{\pi}{1-\pi}\right) \\ A(\eta) &= \log\left(\frac{1}{1-\pi}\right) \\ &= \log(1+\exp(\eta)) \end{aligned}$$

Converting between η and μ: we can use the *logit function*, $\eta = \log(\frac{\pi}{1-\pi})$, to compute the natural parameter η from the mean parameter π. We can use the function $\pi = \frac{1}{1+\exp(-\eta)}$, called the *logistic function*, which is just the inverse of the previous relation to compute the mean parameter from the natural parameter.

$$\begin{aligned} A(\eta) &= \log(1+e^\eta) = -\log(1-\pi) \\ \frac{1}{1+e^\eta} &= 1-\pi \\ \frac{e^\eta}{1+e^\eta} &= \pi \\ \pi &= \frac{1}{1+e^{-\eta}}. \end{aligned}$$

2.3.2 Gaussian

The Gaussian probability density function can be written in exponential form as follows:

$$\begin{aligned} \mathcal{N}(x|\mu,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp[-\frac{1}{2\sigma^2}(x-\mu)^2] \\ &= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp[-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2] \\ &= \frac{1}{(2\pi)^{1/2}}\exp[-\log(\sigma) - \frac{1}{2\sigma^2}\mu^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2] \end{aligned}$$

On comparing the above with the exponential family form, we have:

$$\begin{aligned} h(x) &= \frac{1}{(2\pi)^{1/2}} \\ \eta &= \begin{pmatrix} \frac{\mu}{\sigma^2} \\ \frac{-1}{2\sigma^2} \end{pmatrix} \\ T(x) &= \begin{pmatrix} x \\ x^2 \end{pmatrix} \\ A(\eta) &= \log\sigma + \frac{1}{2\sigma^2}\mu^2 \\ &= \frac{-\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2) \end{aligned}$$

2.4 Properties of the exponential family

For the exponential family, we have

$$\int_{-\infty}^{\infty} h(x)\exp(\eta^T T(x) - A(\eta))dx = 1 \text{ (as the integrand is a probability distribution)}$$

which implies

$$A(\eta) = \log \int h(x)\exp(\eta^T T(x))dx$$

We will use this property below to compute the mean and variance of a distribution in the exponential family. For this reason, the function $A(\eta)$ is often called the *log normalizer*.

2.4.1 Expected Value

In the exponential family, we can take the derivatives of the log partition function in order to obtain the cumulants of the sufficient statistics. This is why $A(\eta)$ is often called a *cumulant function*. Below we will show how to calculate the first and second cumulants of a distribution, which are the mean $E[\cdot]$ and variance $var[\cdot]$, in this case, of the sufficient statistics.

The first derivative of $A(\eta)$ is the expected value, as shown below.

$$\begin{aligned} \frac{\partial A(\eta)}{\partial \eta} &= \frac{\partial}{\partial \eta}\left(\log \int h(x)\exp\{\eta^T T(x)\}dx\right) \\ &= \frac{\int h(x)\exp\{\eta^T T(x)\}T(x)dx}{\exp(A(\eta))} \\ &= \int h(x)\exp\{\eta^T T(x) - A(\eta)\}T(x)dx \\ &= \int p(x|\eta)T(x)dx \\ &= E_\eta[T(x)] \end{aligned}$$

2.4.2 Variance

It is also simple to calculate the variance, which is equal to the second derivative of the log partition function with respect to the natural parameter, as proved below.

$$\begin{aligned}
\frac{\partial^2 A(\eta)}{\partial\eta\partial\eta^T} &= \int h(x)\exp\{\eta^T T(x) - A(\eta)\}T(x)\left(T(x) - \frac{\partial A(\eta)}{\partial\eta}\right)dx \\
&= \int p(x|\eta)T(x)(T(x) - A'(\eta))dx \\
&= \int p(x|\eta)T^2(x)dx - A'(\eta)\int p(x|\eta)T(x)dx \\
&= E[T^2(x)] - E^2[T(x)] \\
&= var[T(x)]
\end{aligned}$$

Example: the Bernoulli distribution
Let X be a Bernoulli random variable with probability π. Then $A(\eta) = \log(1 + e^\eta)$ and $T(x) = x$ as shown in a previous section. As explained above, we can then find the expectation and variance by taking the first and second derivatives of $A(\eta)$:

$$\frac{\partial A(\eta)}{\partial\eta} = \frac{1}{1+e^{-\eta}} = \pi(\eta) = E[x] = E[T(x)]$$

$$\frac{\partial^2 A(\eta)}{\partial\eta\partial\eta^T} = \frac{1}{1+e^{-\eta}} - \frac{1}{1+e^{-\eta}}\frac{1}{1+e^{-\eta}} = \pi(1-\pi) = var[T(x)]$$

The property of the log partition function can be generalize: The m_{th} derivative of $A(\eta)$ is the m_{th} cumulant around the mean, so that the problems of estimating moments (integration) are simplified to differentiation, making the computation much easier.

Also, $A(\eta)$ is a convex function of η, since its second derivative is var[T(x), which is always positive (semi)definite, and when $var[T(x)]$ is positive definite, under strict convexity, we are guaranteed that $\frac{\partial A(\eta)}{\partial\eta}$ is one-to-one, which means that $\mu = E[T(x)] = \frac{\partial A(\eta)}{\partial\eta}$ is invertible, i.e. one-to-one mapping between the mean parameter and the natural parameter.

3 MLE for the Exponential Family

A nice property of the exponential family is that exponential families are closed under sampling. The sufficient statistics $T(x)$ are finite (independent of the size of the data set), i.e., the size of $T(x)$ does not grow as $n = |D| \to \infty$. To see why, consider a sequence of observations $X = \{x_1, x_2, \ldots, x_n\}$ (all x_is are i.i.d.). We look at $T(x)$ as $n \to \infty$. We do this by writing the likelihood in the exponential family form.

$$\begin{aligned}
p(x_1, x_2, \ldots, x_n|\eta) &= \prod_{i=1}^n p(x_i|\eta) \\
&= \left(\prod_{i=1}^n h(x_i)\right)\exp\left(\eta^T\sum_{i=1}^n T(x_i) - nA(\eta)\right)
\end{aligned}$$

The sufficient statistics are thus $\{n, \sum_{i=1}^n T_j(x)\}$, where $j = \{1, \ldots, |T(x)|\}$, which has exactly $|T(x)| + 1$ components. For examples: the sufficient statistics for Bernoulli distribution are: $\{\sum_{i=1}^n \mathbb{1}(x_i = 1), n\}$, and the sufficient statistics for Gaussian distribution are: $\{\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2, n\}$

3.1 Computing MLE

We define the log likelihood with respect to the exponential family form as:

$$\begin{aligned} \ell(\eta; D) &= \log p(D|\eta) \\ &= \log\left(\prod_{i=1}^n h(x_i)\right) + \eta^T \sum_{i=1}^n T(x_i) - nA(\eta) \end{aligned}$$

This is a concave function of η and hence must have a global maximum, which can be found by equating the derivative of the log likelihood with respect to natural parameter η to 0:

$$\begin{aligned} & \frac{\partial}{\partial \eta}\ell(\eta; x_1, \ldots, x_n) = 0 \\ \Rightarrow \quad & \sum_{i=1}^n T(x_i) - n\frac{\partial A(\eta)}{\partial \eta} = 0 \\ \Rightarrow \quad & E_{\eta_{MLE}}[T(x)] = \frac{1}{n}\sum_{i=1}^n T(x_i) \\ \Rightarrow \quad & \mu(\eta_{MLE}) = \frac{1}{n}\sum_{i=1}^n T(x_i) \end{aligned}$$

We see that the theoretical expected sufficient statistics of the model are equal to the empirical average of the sufficient statistics.

4 Conjugacy

When the prior and the posterior have the same form, we say that the prior is a conjugate prior for the corresponding likelihood. For all distributions in the exponential family, we can derive a conjugate prior for the natural parameters. Let the prior be $p(\eta|\tau)$, where τ denotes the hyper-parameters. The posterior can be written as:

$$p(\eta|X) \propto p(X|\eta)p(\eta|\tau)$$

The likelihood of the exponential family is:

$$p(X|\eta) = \left(\prod_{i=1}^n h(x_i)\right) \exp\left(\eta^T \sum_{i=1}^n T(x_i) - nA(\eta)\right)$$

To make the prior conjugate to the likelihood, the prior $p(\eta|\tau)$ must be in the exponential family form with two terms, one term $\tau = \{\tau_1, ..., \tau_k\}$ multiplying the η, the other term τ_0 multiplying $A(\eta)$, as:

$$p(\eta|\tau) \propto \exp\left\{\eta^T \tau - \tau_0 A(\eta)\right\}$$

Then the posterior can be written as:

$$\begin{aligned} p(\eta|X) &\propto p(X|\eta)p(\eta|\tau) \\ &\propto \exp\left(\eta^T \sum_{i=1}^{n} T(x_i) - nA(\eta)\right) \exp\left(\eta^T \tau - \tau_0 A(\eta)\right) \\ &= \exp\left\{\eta^T \left(\sum_{i=1}^{n} T(x_i) + \tau\right) - (n+\tau_0)A(\eta)\right\} \end{aligned}$$

So we see the posterior has the same exponential family form as the prior, and the posterior hyper-parameters are adding the sum of the sufficient statistics to hyper-parameters of the conjugate prior. The exponential family is the only family of distributions for which the conjugate priors exist. This is a convenient property of the exponential family because conjugate priors simplify computation of the posterior.