Lecture #8: Thursday, 27 January 2005
Lecturer: Prof. Charles Elkan
Scribe: Banu Dost

1 Exponential Family Definition

Let the sample space X be $\mathbb{R}^n$ for some n, i.e. a Euclidean space. The continuous distribution family P_θ on X is a member of the exponential family if and only if its density function has the following form:

$$P_\theta(x) = C(\theta) \cdot exp[Q_1(\theta) \cdot t_1(x) + ... + Q_k(\theta) \cdot t_k(x)] \cdot h(x)$$

where θ is any collection of parameters (not necessarily real valued) and the Qs and ts are real-valued functions. Note that by the factorization theorem, the vector $(t_1(x), \cdots, t_k(x))$ is sufficient.

The exponential family includes Gaussian, Poisson, and many other continous distribution families. The definition can also be extended to discrete distributions, to include the binomial, geometric and other discrete families. However, it does not include uniform distributions.

We usually have a major simplification when the parameter space is $\mathbb{R}^k$ and $Q_k(\theta) = \theta_k$.

$$P_\theta(x) = C(\theta) \cdot exp[\theta_1 \cdot t_1(x) + ... + \theta_k \cdot t_k(x)] \cdot h(x)$$

In this case, the parameters $\theta_1, \cdots, \theta_k$ are called natural parameters.

Example 1.1. *Gaussian exponential family*
Suppose $(x_1, ..., x_n)$ is an iid sample from a univariate Gaussian. The joint probability distribution of $x_1, ..., x_n$ is

$$P_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}^n} \cdot exp\left(-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right)$$

Here it looks like the parameter μ is involved with each separate x_i. However, we can rewrite the equation above as,

$$\begin{aligned} P_\theta(x) &= \frac{1}{\sqrt{2\pi\sigma^2}^n} \cdot exp\left(-\frac{\sum_{i=1}^n (x_i^2 - 2x_i\mu + \mu^2)}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}^n} \cdot exp\left(-\frac{n\mu^2}{2\sigma^2}\right) \cdot exp\left(-\frac{\sum(x_i^2 - 2x_i\mu)}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}^n} \cdot exp\left(-\frac{n\mu^2}{2\sigma^2}\right) \cdot exp\left(-\frac{\sum x_i^2}{2\sigma^2} + \frac{\mu \sum x_i}{\sigma^2}\right) \end{aligned}$$

Let $C(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}^n} \cdot exp\left(\frac{n\mu^2}{2\sigma^2}\right)$, $h(x) = 1$, $t_1(x) = \sum x_i^2$, and $t_2(x) = \sum x_i$. Thus, by definition Gaussian distribution is an exponential family.

Observe that we can describe the Gaussian family of distributions using a different definition of the parameters. Let $\phi = (-1/(2\sigma^2), \mu/\sigma^2)$. In this case, $P_\theta(x) = C'(\phi) \cdot exp\left(\phi_1 \cdot t_1(x) + \phi_2 \cdot t_2(x)\right)$. These alternative parameters ϕ are the "natural" parameters for Gaussian distribution.

2 Completeness theorem for an exponential family

Consider the exponential family of distributions

$$P_\theta(x) = C(\theta) \cdot exp\left[\theta_1 \cdot t_1(x) + \cdots + \theta_k \cdot t_k(x)\right] \cdot h(x)$$

with sufficient statistic $t(x) = (t_1(x), ..., t_k(x))$. Suppose the parameter space Θ contains a k-dimensional rectangle. Then the family P_θ^t of distributions of t is complete.

Proof: Omitted.

Note that when you define a family of distributions, you have to say not only what the parameters are (e.g. μ and σ^2) but also what the allowable ranges for these parameters are (e.g. $\mu > 0$, $\sigma^2 > \mu$).

In order to prove that the family of distributions of a statistic is complete, using this theorem, you first have to rewrite your exponential family of distribution of the data in the form of $P_\theta(x) = C(\phi) \cdot exp\left[\phi_1 \cdot t_1(x) + \cdots + \phi_k \cdot t_k(x)\right]$ where $\phi = (\phi_1, \cdots, \phi_k)$ are the natural numbers. Then, you have to find a rectangle of dimension k in the range of the natural parameters.

When a family of distributions is highly restricted, completeness may fail since you may not be able to find a rectangle of full dimension, i.e. of dimension k. Suppose, for instance, we only consider the Gaussians with restriction $\Theta = \{\theta : \mu = \sigma^2\}$. Then, the space for the natural parameters $\phi = (-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2})$ will be $\Phi = \{\phi : \phi_1 < 0, \phi_2 = 1\}$ and Φ does not contain a two dimensional rectangle. In this case, we can not use the theorem above to deduce completeness of the family of distributions of t.

3 Completeness theorem for an exponential family - v2

Let $x_1, \cdots, x_n$ be iid observations from an exponential family with probability density or mass function of the form

$$P_\theta(x) = C(\theta) \cdot exp\left[w_1(\theta_1) \cdot t_1(x) + \cdots + w_k(\theta_k) \cdot t_k(x)\right] \cdot h(x)$$

where $\theta = (\theta_1, \cdots, \theta_k)$. Then, the statistic

$$T(X) = \left(\sum_{i=1}^{n} t_1(x_1), \cdots, \sum_{i=1}^{n} t_k(x_k) \right)$$

is complete as long as the parameter space Θ contains an open set in $\mathbb{R}^k$. The main difference between the two versions of the completeness theorem is that the second one considers the pdf of a single random variable x_i while the first one considers the joint pdf of $x_1, \cdots, x_n$ as an exponential family.

4 Maximum Likelihood

The method of maximum likelihood is a technique to derive estimators. The main idea behind this method is to estimate the true parameter θ by any parameter at which the observed sample is most likely.

We call the function $P(x, \theta)$ the likelihood function. The distinction between the likelihood function and the probability density function is which variable is considered fixed and which is varying. The usual pdf considers θ as fixed and x as the variable. However, the likelihood function considers x as fixed and θ as varying over all possible parameter values. We would like to find the value of θ that maximizes the probability of the observed sample x. Thus, maximum likelihood estimator for θ is defined by the equation,

$$\theta_{ML} = argmax_\theta P(x, \theta)$$

In order to find the value of θ that maximizes the likelihood function, we need to differentiate this function. Since sometimes differentiating the logarithm of a density function is easier than differentiating the original function, we also consider the log-likelihood function $\ln P(x, \theta)$. Then, our equation for MLE becomes,

$$\theta_{ML} = argmax_\theta \ln P(x, \theta)$$

5 The Score Function

The derivative with respect to θ of the log-likelihood is called the score function.

$$s(x, \theta) = \frac{d}{d\theta} \ln P(x, \theta)$$

Using the fact that $\frac{d}{dx} ln(x) = 1/x$,

$$s(x, \theta) = \frac{1}{P(x, \theta)} \cdot \frac{d}{d\theta} P(x, \theta)$$

We are looking for the maximum likelihood estimator of θ, i.e. the point at which log likelihood function $L(x, \theta)$ has a global maxima. At this point, the first derivative of $L(x, \theta)$ is zero, in other words $s(x, \theta) = 0$. Note that $s(x, \theta) = 0$ is a necessary condition for global maxima, but not sufficient. Because this condition holds for local or global maxima, local or global minima, or inferior points.

Example 5.1. *Bernoulli Maximum Likelihood Estimator*
Let $x_1, \cdots, x_n$ be iid Bernoulli with $p = \theta$. Then, the likelihood function is

$$\begin{aligned} P(x, \theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \end{aligned}$$

Take the logarithm for log-likelihood function,

$$L(x, \theta) = \sum_{i=1}^{n} x_i \ln \theta + (n - \sum_{i=1}^{n} x_i) \ln(1-\theta)$$

Then, we get the score function by differentiating wrt θ,

$$\begin{aligned} s(x, \theta) &= \frac{d}{d\theta} L(x, \theta) \\ &= \frac{\sum x_i}{\theta} + \frac{\sum x_i - n}{\theta} \\ &= \frac{\sum x_i - n\theta}{\theta(1-\theta)} \end{aligned}$$

Set $s(x, \theta) = 0$.

$$s(x, \theta) = \frac{d}{d\theta} L(x, \theta) = 0 \quad \text{iff } \theta = \frac{\sum x_i}{n}$$

Since the second derivative of L(x, θ) is negative, and $\theta = \frac{\sum x_i}{n}$ is the only point at which $s(x, \theta) = 0$, then $\theta = \frac{\sum x_i}{n}$ is a global maxima. Thus, the MLE for θ is $\frac{\sum x_i}{n}$.