Stat 241A Statistical Learning Theory **Fall 2004**

Lecture 11: (Exponential Family (October 5, 2004)

Lecturer: Prof Jordan *Scribe: Sivakumar Rathinam*

11.1 Exponential family representations

A general representation of a exponential family is given by the following probability density function:

$$p(x|\eta) = h(x)\exp\{\eta^T T(x) - A(\eta)\} \tag{11.1}$$

where $h(x)$ is called the *base density* which is always ≥ 0, η is the *natural parameter*, $T(x)$ is the *sufficient statistic vector* and $A(\eta)$ is the *cumulant generating function* or the *log normalizer*. The choice of $T(x), h(x)$ determines the member of the exponential family. Also we know that since this is a density function,

$$1 = \int h(x)\exp\{\eta^T T(x) - A(\eta)\}dx \tag{11.2}$$

or,

$$A(\eta) = \log \int (h(x)\exp\{\eta^T T(x)\}dx) \tag{11.3}$$

For example, take a Bernoulli distribution. We have $p(x|\pi) = \pi^x(1-\pi)^{1-x}$. By some simple adjustments to the density function (apply $\exp\log p(x|\pi)$), we can show that $h(x) = 1, \eta = \log(\frac{\pi}{1-\pi}), T(x) = x$ and $A(\eta) = \log(1+\exp(-\eta))$ in this case.

For a Gaussian distribution, $p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}}\exp(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 + \frac{\mu^2}{2\sigma^2} - \log\sigma)$. In this case, $h(x) = \frac{1}{\sqrt{2\pi}}$, $\eta = [\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}]$, $T(x) = [x, x^2]$ and $A(\eta)$ is an exercise left to the reader.

11.1.1 Properties of Exponential Family

Fact 1:

$$\frac{\partial}{\partial\eta}A(\eta) = E_\eta T(x) \tag{11.4}$$

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

Fact 2:

$$\frac{\partial^2}{\partial\eta\partial\eta^T}A(\eta) = Var(T(x)) \tag{11.5}$$

Lets look at an example. If x is a Bernoulli distribution with parameter π, then $\frac{\partial}{\partial\eta}A(\eta) = \frac{1}{1+\exp{-\eta}} = \pi(\eta) = \pi = E(x) = E(T(x))$. Also, $\frac{\partial^2}{\partial\eta\partial\eta^T}A(\eta) = \pi(\eta)(1-\pi(\eta)) = \pi(1-\pi) = var(T(x))$.

In general, we can actually show that the m^{th} derivative of the cumulant generating function $A(\eta)$ is the m^{th} cumulant around the mean. This is a very useful result because we have converted the problem of trying to estimate the moments which involves integrating to a problem of differentiating a function. Differentiating is easier and hence it is worthwhile for us to study the properties of this cumulant generating function.

11.1.2 Properties of $A(\eta)$

Property 1: Domain of $A = \{\eta | A(\eta) < \inf\}$ is a convex set.

Property 2: $A(\eta)$ is a convex function of η. Proof: Note that $\frac{\partial^2}{\partial\eta\partial\eta^T}A(\eta) = Var(T(x))$ which is always positive semi-definite. Q.E.D.

In particular, say $Var(T(x))$ is positive definite, then the relationship $\mu = E(T(x)) = \frac{\partial}{\partial\eta}A(\eta)$ is invertible. That is, $\eta = [\frac{\partial}{\partial\eta}A(\eta)]^{-1}(\mu)$. This is due to the fact that the function $\frac{\partial}{\partial\eta}A(\eta)$ is one-to-one under strict convexity.

11.1.3 Sufficiency

$T(x)$ is a statistic function of data that does not involve θ, the parameter of the distribution that generated x. $T(x)$ is said to be *sufficient* for θ if *all info about θ contained in x is also contained in* $T(x)$.

For example, say x_n are i.i.d with normal distribution (μ,1). Then $T(x_1, ...x_n) = \frac{\sum_i x_i}{n}$ is sufficient for μ. Of course, Bayesians and frequentists have a different way of thinking about this sufficient statistic. For a Bayesian, all $\theta, x, T(x)$ are random variables. So they define $T(x)$ as *sufficient* if $\theta \perp\!\!\!\perp x | T(x)$. For a frequentist, θ is fixed and he defines $T(x)$ to be sufficient if the conditional distribution of x given $T(x)$ does not involve θ.

11.1.4 Neyman Factorization Theorem

$$T(x) \text{ is sufficient iff } p(x|\theta) = g(T(x), \theta)h(x, T(x)) \tag{11.6}$$

Note: This is automatically true for distributions in the exponential family as $h(x) = h(x, T(x))$ and $g(T(x), \theta) = \exp\{\theta^T T(x) - A(\theta)\}$.

11.1.5 Maximum likelihood estimation in the Exponential Family

Fact: Exponential families are closed under sampling.

Consider i.i.d samples $x_1, x_2...x_n$ which belong to a exponential family $p(x|\eta)$. Now,

$$p(x_1, x_2...x_n|\eta) = \prod_i p(x_i|\eta)$$

$$= (\prod_i h(x_i)) \exp\{\eta^T \sum_i T(x_i) - nA(\eta)\}$$

So basically, we can make the following observations: The sufficiency vector doesn't grow as the number of samples; The density function remains in the exponential family.

11.1.6 Maximum Likelihood Estimation

$$\begin{aligned} \text{(Likelihood) } l(\eta; x_1...x_n) &= \log(p(x_1...x_n|\eta)) && (11.7) \\ &= \log h(x_1..x_n) + \eta^T \sum_i T(x_i) - nA(\eta) && (11.8) \end{aligned}$$

We can easily infer that this is a concave function and also the domain is convex. Essentially what we are trying to estimate is η. If we differentiate with respect to η to find the maximum likelihood, we get:

$$\frac{\partial}{\partial \eta} l(\eta; x_1..x_n) = \sum_i T(x_i) - n \frac{\partial}{\partial \eta} A(\eta) \tag{11.9}$$

To solve for η_{ml}, we need to solve,

$$\frac{\partial}{\partial \eta} A(\eta) = \frac{\sum_i T(x_i)}{n} \tag{11.10}$$

That is we get $E_{\eta_{ml}}(T(x)) = \frac{\sum_i T(x_i)}{n}$. Recall that $\mu = \mu(\eta) = \frac{\partial}{\partial \eta} A(\eta)$. Now we have a general question: $\mu(\eta_{ml}) = \frac{\partial}{\partial \eta} A(\eta_{ml})$?. It turns out that this is true. This is a general solution to the maximum likelihood parameter estimation problem across all members of the exponential family.