# SMLDM HT 2014 - Part C Problem Sheet 5

1. An exponential family is a family of distributions parameterized by a $d$-dimensional vector $\theta$, and has density of the form:

$$p(x;\theta) = h(x)\exp\left(\theta^\top S(x) - A(\theta)\right)$$

where $h(x)$ is a function that depends only on $x$, $S : \mathbb{R}^p \to \mathbb{R}^d$ is the *sufficient statistics* function, and

$$A(\theta) = \log \int_{\mathbb{R}^p} h(x)\exp\left(\theta^\top S(x)\right) dx$$

is a normalization constant. Exponential families can be defined over other spaces as well, in which case $\mathbb{R}^p$ above is replaced by some other space $\mathbb{X}$.

(a) Write the Bernoulli, normal and Poisson distributions in exponential family form, identifying the functions $h$, $S$ and $A$.

**Answer:** Bernoulli:

$$p(x;\phi) = \phi^x(1-\phi)^{1-x} = \exp(x\log\phi + (1-x)\log(1-\phi)) = \exp(x\log\frac{\phi}{1-\phi} + \log(1-\phi))$$

So $S(x) = x$, $\theta = \log\frac{\phi}{1-\phi}$, $h(x) = 1$ and

$$A(\theta) = -\log(1 - s(\theta)) = -\log(s(-\theta)) = \log(1 + \exp(\theta))$$

Normal:

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2} = e^{-\frac{1}{2\sigma^2}x^2 + \frac{1}{\sigma^2}x\mu - \frac{1}{2\sigma^2}\mu^2 - \frac{1}{2}\log(2\pi\sigma^2)}$$

So $S(x) = [x, x^2]^\top$, $\theta = [\mu/\sigma^2, -1/2\sigma^2]^\top$, $h(x) = 1$ and $A(\theta) = \frac{1}{2\sigma^2}\mu^2 + \frac{1}{2}\log(2\pi\sigma^2)$, which we'll need to express as function of $\theta$.

Poisson:

$$p(x;\lambda) = \frac{e^{-\lambda}}{x!}\lambda^x = e^{-\lambda - \log x! + x\log\lambda}$$

so $S(x) = x$, $h(x) = 1/x!$, $\theta = \log\lambda$ and $A(\theta) = \lambda = e^\theta$.

(b) Show that

$$\nabla_\theta A(\theta) = \mathbb{E}[S(X)] \qquad \nabla^2_\theta A(\theta) = \text{Cov}[S(X), S(X)]$$

where $X$ is a random variable with distribution given by the exponential family distribution with parameter $\theta$.

**Answer:** The first derivative is:

$$\nabla_\theta A(\theta) = \frac{\int h(x)\exp(\theta^\top S(x))S(x)dx}{\int h(x)\exp(\theta^\top S(x))dx} = \mathbb{E}[S(X)]$$

The second derivative is:

$$
\begin{aligned}
\nabla^2_\theta A(\theta) =& \frac{\int h(x)\exp(\theta^\top S(x))S(x)S(x)^\top dx}{\int h(x)\exp(\theta^\top S(x))dx} \\
&- \frac{\int h(x)\exp(\theta^\top S(x))S(x)dx}{\int h(x)\exp(\theta^\top S(x))dx} \frac{\int h(x)\exp(\theta^\top S(x))S(x)^\top dx}{\int h(x)\exp(\theta^\top S(x))dx} \\
=& \mathbb{E}[S(X)S(X)^\top] - \mathbb{E}[S(X)]\mathbb{E}[S(X)]^\top = \mathrm{Cov}[S(X), S(X)]
\end{aligned}
$$

(c) Suppose given a dataset $(x_i)_{i=1}^n$ we wish to perform maximum likelihood estimation of $\theta$. Explain why this is a convex optimization problem. Under what conditions is the ML estimator uniquely defined?

**Answer:** The log likelihood is

$$
\begin{aligned}
&\sum_{i=1}^n \log h(x_i) + \theta^\top S(x_i) - A(\theta) \\
=& \left(\sum_{i=1}^n \log h(x_i)\right) + \theta^\top \left(\sum_{i=1}^n S(x_i)\right) - nA(\theta)
\end{aligned}
$$

So first term doesn't depend on $\theta$, second is linear in $\theta$, and third is concave in $\theta$, since second derivative of $A$ is positive semidefinite. Thus the objective is concave. The ML estimator is uniquely defined if the second derivative is positive definite. This happens if the entries of $S(x)$ are linearly independent, that is, a vector $\lambda$ has $\lambda^\top S(x) = 0$ for all $x$ if and only if $\lambda = 0$.

2. Consider the following *maximum-entropy* problem. Suppose we have a dataset $(x_i)_{i=1}^n$, from which we can calculate a number of statistics, say

$$T_j = \frac{1}{n}\sum_{i=1}^n S_j(x_i)$$

for $j = 1, \ldots, d$, and functions $S_j : \mathbb{R}^p \to \mathbb{R}$. For example, when $p = 1$, we can take $S_1(x) = x$, $S_2(x) = x^2$. We wish to find the density $f(x)$ which maximizes the differential entropy

$$\mathcal{H}[f] = -\int_{\mathbb{R}^p} f(x)\log f(x) dx$$

subject to the constraints:

$$\int_{\mathbb{R}^p} f(x) S_j(x) dx = T_j$$

(a) Formulate the maximum entropy problem as a convex optimization problem, and show that the maximum entropy problem is equivalent to the problem of maximum likelihood estimation in an exponential family.

**Answer:** This is a convex optimization problem because the entropy is concave, which we want to maximize. Negating, the negative entropy is to be minimized and it is convex. The constraints are linear in $f(x)$.

The Lagrangian is

$$\mathcal{L}(f,\lambda,\gamma) = \int_{\mathbb{R}^p} f(x)\log f(x)dx + \sum_{j=1}^d \lambda_j \left(T_j - \int_{\mathbb{R}^p} f(x)S_j(x)dx\right) + \gamma\left(1 - \int_{\mathbb{R}^p} f(x)dx\right)$$

with Lagrange multipliers $\lambda$ and $\gamma$. Solving for $f$, the derivative wrt $f(x)$ is

$$0 = \log f(x) + 1 - \sum_{j=1}^d \lambda_j S_j(x) - \gamma \tag{1}$$

$$f(x) = e^{\gamma-1}\exp\left(\sum_{j=1}^d \lambda_j S_j(x)\right)$$

So $f(x)$ is an exponential family distribution with sufficient statistics $S(x) = [S_1(x), \ldots, S_d(x)]^\top$ and parameters $\lambda$, and $e^{\gamma-1}$ is the normalization constant, i.e.

$$e^{1-\gamma} = \int_{\mathbb{R}^p} \exp\left(\sum_{j=1}^d \lambda_j S_j(x)\right) dx \tag{2}$$

The dual objective is obtained by substituting (1) back into the Lagrangian,

$$\begin{aligned}
&- \int_{\mathbb{R}^p} f(x)dx + \sum_{j=1}^d \lambda_j T_j + \gamma \\
=& \sum_{j=1}^d \lambda_j T_j + \gamma - 1 \\
=& \sum_{j=1}^d \lambda_j T_j - \log \int_{\mathbb{R}^p} \exp\left(\sum_{j=1}^d \lambda_j S_j(x)\right) dx \qquad \text{by (2)}
\end{aligned}$$

We wish to maximize this dual objective. If we multiply by $n$, the dataset size, and take $T_j$ to be the empirical mean of $S_j(x)$ under the dataset, this is the objective function we would get under ML estimation.

(b) Suppose that we are not certain about the statistics collected, and wish to introduce a degree of uncertainty into our method. Say we relax our equality constraints by interval constraints,

$$T_j - C \leq \int_{\mathbb{R}^p} f(x)S_j(x)dx \leq T_j + C$$

for a positive number $C > 0$. Show that this problem is equivalent to a regularized maximum likelihood estimation problem in an exponential family, with an $L_1$ regularization.

**Answer:** These are inequality constraints, so we will need to introduce Lagrange multipliers

$\lambda_j^+ \geq 0, \lambda_j^- \geq 0$ for both sides of the inequalities. The Lagrangian is

$$\begin{aligned}
\mathcal{L}(f, \lambda^+, \lambda^-, \gamma) = & \int_{\mathbb{R}^p} f(x) \log f(x) dx \\
& + \sum_{j=1}^d \lambda_j^+ \left( T_j - C - \int_{\mathbb{R}^p} f(x) S_j(x) dx \right) \\
& + \sum_{j=1}^d \lambda_j^- \left( \int_{\mathbb{R}^p} f(x) S_j(x) dx - T_j - C \right) \\
& + \gamma \left( 1 - \int_{\mathbb{R}^p} f(x) dx \right)
\end{aligned}$$

Again setting the derivative wrt $f(x)$ to zero, we find that

$$f(x) = e^{\gamma - 1} \exp \left( \sum_{j=1}^d (\lambda_j^+ - \lambda_j^-) S_j(x) \right)$$

which is of exponential family form, with parameters $\lambda_j = \lambda_j^+ - \lambda_j^-$. Substituting back into the Lagrangian, we get the dual objective which is to be maximized:

$$\sum_{j=1}^d \lambda_j T_j - \log \int_{\mathbb{R}^p} \exp \left( \sum_{j=1}^d \lambda_j S_j(x) \right) dx - C \left( \sum_{j=1}^d \lambda_j^+ + \lambda_j^- \right)$$

Multiplying by $n$, the dataset size again, the first two terms are again the log likelihood. The last term is

$$-nC \left( \sum_{j=1}^d \lambda_j^+ + \lambda_j^- \right)$$

The claim is now that the sum inside is $\|\lambda\|_1$, so we get the $L_1$ regularization term. Here we can use the complementary slackness property, which gives, for each $j$,

$$\begin{aligned}
\lambda_j^+ \left( T_j - C - \int_{\mathbb{R}^p} f(x) S_j(x) dx \right) &= 0 \\
\lambda_j^- \left( \int_{\mathbb{R}^p} f(x) S_j(x) dx - T_j - C \right) &= 0
\end{aligned}$$

Now $\lambda_j^+ > 0$ implies that the integral equals $T_j - C$, so it cannot equal $T_j + C$, so that $\lambda_j^- = 0$. Likewise, $\lambda_j^- > 0$ impiles $\lambda_j^+ = 0$. Hence $\lambda_j^+ + \lambda_j^- = |\lambda_j|$.

3. Let $(x_i, y_i)_{i=1}^n$ be our dataset, with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Linear regression can be formulated as empirical risk minimization, where the model is to predict $y$ as $x^\top \beta$, and we use the squared loss:

$$R^{\text{emp}}(\beta) = \sum_{i=1}^n \frac{1}{2} (y_i - x_i^\top \beta)^2$$

(a) Show that the optimal parameter is

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

where $\mathbf{X}$ is a $n \times p$ matrix with $i$th row given $x_i^\top$, and $\mathbf{Y}$ is a $n \times 1$ matrix with $i$th entry $y_i$.

**Answer:** We can write the empirical risk as

$$\frac{1}{2}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$$

Differentiating wrt $\beta$ and setting to 0,

$$\begin{aligned}(\mathbf{Y} - \mathbf{X}\beta)^\top \mathbf{X} &= 0 \\ \mathbf{Y}^\top \mathbf{X} - \beta^\top (\mathbf{X}^\top \mathbf{X}) &= 0 \\ \hat{\beta} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}\end{aligned}$$

(b) Consider regularizing our empirical risk by incorporating a $L_2$ regularizer. That is, find $\beta$ minimizing

$$\frac{C}{2}\|\beta\|_2^2 + \sum_{i=1}^{n} \frac{1}{2}(y_i - x_i^\top \beta)^2$$

Show that the optimal parameter is given by the ridge regression estimator

$$\hat{\beta} = (CI + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

**Answer:** The objective becomes:

$$\frac{1}{2}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \frac{C}{2}\|\beta\|_2^2$$

Again differentiating and setting derivative to 0,

$$\begin{aligned}(\mathbf{Y} - \mathbf{X}\beta)^\top \mathbf{X} + C\beta^\top &= 0 \\ \mathbf{Y}^\top \mathbf{X} - \beta^\top (CI + \mathbf{X}^\top \mathbf{X}) &= 0 \\ \hat{\beta} &= (CI + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}\end{aligned}$$

(c) Suppose we wish to introduce nonlinearities into the model, by transforming $x \mapsto \phi(x)$. Show how this transformation may be achieved using the kernel trick. That is, let $\mathbf{\Phi}$ be a matrix with $i$th row given by $\phi(x_i)^\top$. The optimal parameters $\hat{\beta}$ would then be given by (previous part):

$$\hat{\beta} = (CI + \mathbf{\Phi}^\top \mathbf{\Phi})^{-1} \mathbf{\Phi}^\top \mathbf{Y}$$

Express the predicted $y$ values on the training set, $\mathbf{\Phi}\hat{\beta}$, only in terms of $\mathbf{Y}$ and the Gram matrix $G = \mathbf{\Phi}\mathbf{\Phi}^\top$, with $G_{ij} = \phi(x_i)^\top \phi(x_j) = \kappa(x_i, x_j)$ where $\kappa$ is some kernel function.

Compute an expression for the value of $y_0$ predicted by the model at a test vector $x_0$.

You will find the Woodbury matrix inversion formula useful:

$$(A + UBV)^{-1} = A^{-1} - A^{-1}U(B^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

where $A$ and $B$ are square invertible matrices of size $n \times n$ and $p \times p$ respectively, and $U$ and $V$ are $n \times p$ and $p \times n$ rectangular matrices.

**Answer:** Using $\mathbf{\Phi}$ instead of $\mathbf{X}$, we would get

$$\hat{\beta} = (CI + \mathbf{\Phi}^\top \mathbf{\Phi})^{-1}\mathbf{\Phi}^\top \mathbf{Y}$$

instead. Multiply by $\mathbf{\Phi}$,

$$\begin{aligned}
\mathbf{\Phi}\hat{\beta} =& \mathbf{\Phi}(CI + \mathbf{\Phi}^\top \mathbf{\Phi})^{-1}\mathbf{\Phi}^\top \mathbf{Y} \\
=& \mathbf{\Phi}(C^{-1}I - C^{-1}\mathbf{\Phi}^\top(I + \mathbf{\Phi}(C^{-1}I)\mathbf{\Phi}^\top)^{-1}\mathbf{\Phi}C^{-1})\mathbf{\Phi}^\top \mathbf{Y} \\
=& C^{-1}(\mathbf{\Phi}\mathbf{\Phi}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top(CI + \mathbf{\Phi}\mathbf{\Phi}^\top)^{-1}\mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{Y} \\
=& C^{-1}(G - G(CI + G)^{-1}G)\mathbf{Y}
\end{aligned}$$

Finally, for a test vector $x_0$, let $\phi_0 = \phi(x)$. Then the prediction is $\phi_0^\top \hat{\beta}$, which gives

$$C^{-1}(\phi_0^\top \mathbf{\Phi}^\top - \phi_0^\top \mathbf{\Phi}^\top (CI + G)^{-1}G)\mathbf{Y}$$

where we note that $\phi_0^\top \mathbf{\Phi}^\top$ is a row vector with $i$th entry $\kappa(x_0, x_i)$.

In particular, the nonlinear model can be "kernelized" and all computations can be carried out without explicit computation of $\phi(x)$.