SMLDM HT 2014 - Part C Problem Sheet 5

1. An exponential family is a family of distributions parameterized by a d-dimensional vector θ, 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.

   (b) Show that
$$\nabla_\theta A(\theta) = \mathbb{E}[S(X)] \qquad\qquad \nabla^2_\theta A(\theta) = \text{Cov}[S(X), S(X)]$$
where X is a random variable with distribution given by the exponential family distribution with parameter θ.

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

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 $j = 1, \dots, d$, 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.

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

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 ith row given $x_i^\top$, and $\mathbf{Y}$ is a $n \times 1$ matrix with ith entry y_i.

   (b) Consider regularizing our empirical risk by incorporating a L_2 regularizer. That is, find β 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

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

   (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 ith 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 κ 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.