SMLDM HT 2014 - MSc Problem Sheet 3

1. Consider using logistic regression to model the conditional distribution of binary labels $Y \in \{+1, -1\}$ given data vectors X. Suppose that the data is linearly separable, i.e. there is a hyperplane separating the two classes. Show that the maximum likelihood estimator is ill-defined.

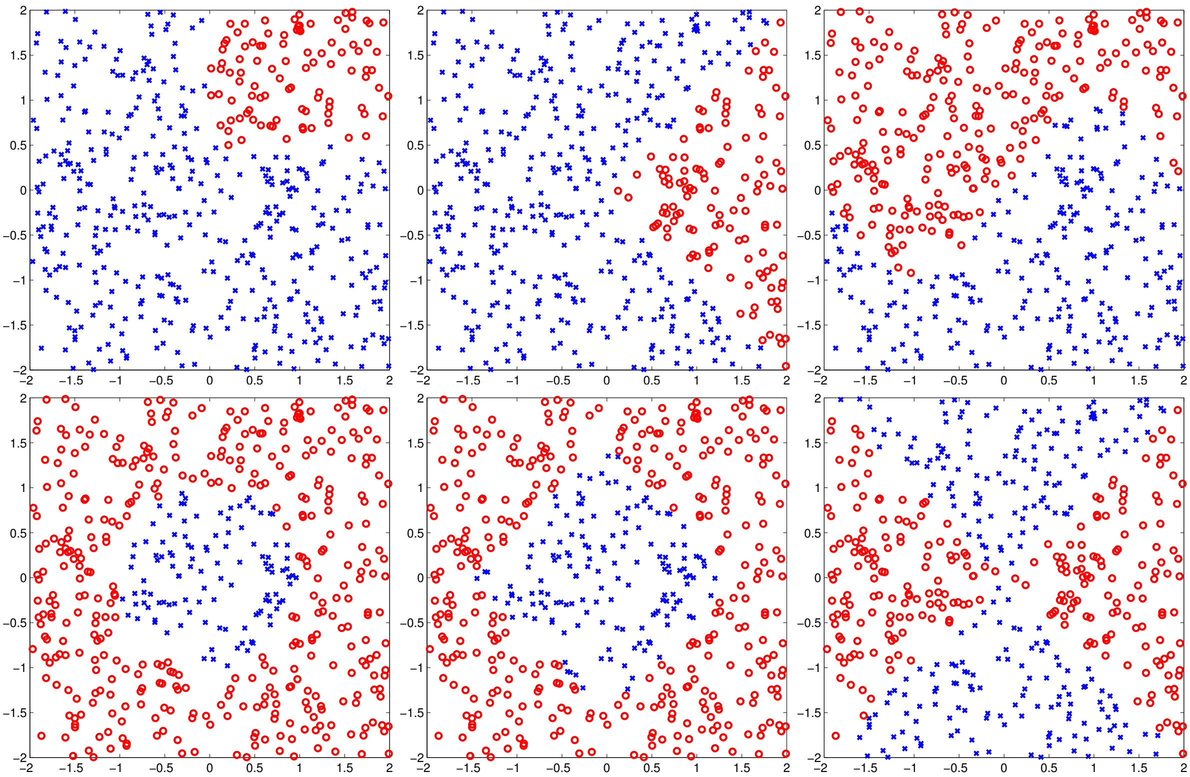
2. The receiver operating characteristic (ROC) curve plots the sensitivity against the specificity of a binary classifier as a threshold for discrimination is varied. The larger the area under the ROC curve (AUC), the better the classifier is.

   Suppose the data space is $\mathbb{R}$, the class-conditional densities are $f_0(x)$ and $f_1(x)$ for $x \in \mathbb{R}$ and for the two classes 0 and 1, and that the optimal Bayes classifier is to classify $+1$ when $x > c$ for some threshold c, which varies over $\mathbb{R}$.

   (a) Give expressions for the specificity and sensitivity of the classifier at threshold c.

   (b) Show that the AUC corresponds to the probability that $X_1 > X_0$, if data items X_1 and X_0 are independent and comes from class 1 and 0 respectively.

3. For each of the datasets below, find a non-linear function $\phi(x)$ which makes the data linearly separable, and the discriminant function (linear in $\phi(x)$) which will classify perfectly. Briefly explain your answer. You may assume, if a boundary looks like a straight line, or a function you are familiar with, that it is.



4. 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 $\mathbf{X}$.

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

(b) Show that

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

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?

5. 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.

(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.