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.

   Answer: Since the data is linearly separable, there is a scalar α and vector β such that $\alpha+\beta^\top X < 0$ whenever $Y = -1$ and $\alpha + \beta^\top X > 0$ whenever $Y = +1$. Let $c > 0$. the log likelihood at $a = c\alpha$, $b = c\beta$ is

   $$\sum_{i=1}^{n} -\log(1 + \exp(-y_i(c\alpha + c\beta^\top x_i)))$$

   Differentiating with respect to c,

   $$\sum_{i=1}^{n} s(cy_i(\alpha + \beta^\top x_i))y_i(\alpha + \beta^\top x_i)$$

   Noting that this is always positive, the log likelihood would be maximized only when $c \to \infty$.

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.

   Answer: At a threshold c, the sensitivity is the true positive rate, which is:

   $$\int_c^\infty f_1(x)dx$$

   while the specificity is the true negative rate:

   $$\int_{-\infty}^c f_0(x)dx$$

   (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.

   Answer: Define the function

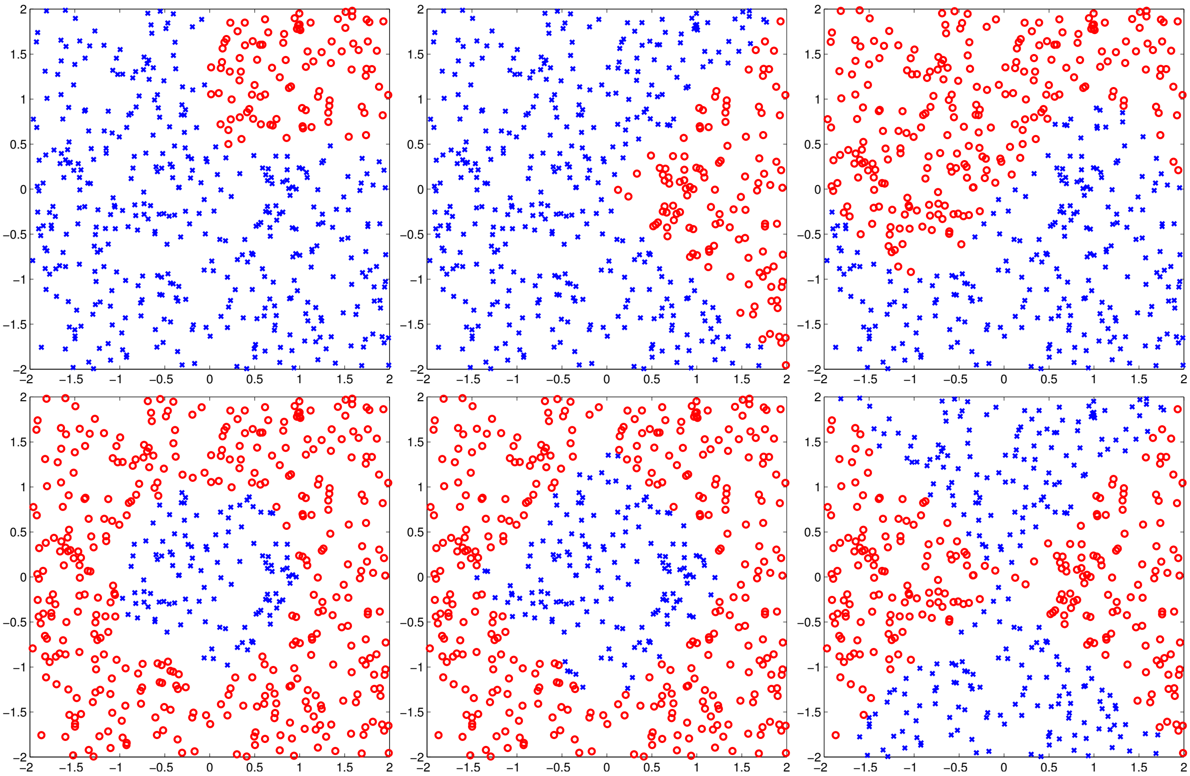
   $$F_0(c) = \int_{-\infty}^c f_0(x)dx$$

   which is the CDF of the 0 class so is invertible. At a specificity level s, the corresponding threshold is $F_0^{-1}(s)$ and so the AUC is

   $$\begin{aligned}&\int_0^1 \int_{F_0^{-1}(s)}^\infty f_1(x)dxds \\ =&\int_{-\infty}^\infty \int_z^\infty f_1(x)dxf_0(z)dz \qquad \text{by change of variable } s \mapsto F_0^{-1}(s) = z \\ =&\mathbb{P}(X_1 > X_0)\end{aligned}$$

   which is the probability of $X_1 > X_0$.

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.



Answer: From left to right and top to bottom:

(a) Looks like we want $x_1 > 0$ and $x_2 > .5$. So use $\phi_1(x) = (\text{sign}(x_1), \text{sign}(x_2 - .5))^\top$. Then perfect classification can be obtained by $\text{sign}(x_1) + \text{sign}(x_2 - .5) \geq 2$.

(b) Looks like we want $x_1 < x_2$ and $x_1 > -x_2$. Use $\phi_2(x) = (\text{sign}(x_1 - x_2), \text{sign}(x_1 + x_2))^\top$ and classify by $-\text{sign}(x_1 - x_2) + \text{sign}(x_1 + x_2) \geq 2$.

(c) Looks like $x_2 < \sin(x_1)$, so $\phi_3(x) = (x_2, \sin(x_1))^\top$ and discriminate via $\sin(x_1) - x_2 > 0$.

(d) Looks like a circle, so we want $\sqrt{x_1^2 + x_2^2} > 1$. Use $\phi_4(x) = \sqrt{x_1^2 + x_2^2} > 1$.

(e) Looks like a diamond, so we want $|x_1| + |x_2| \leq 1$. Use $\phi_5(x) = |x_1| + |x_2|$.

(f) The two lines are $x_1 - x_2 = 0$ and $x_2 + x_1 = 0$. The red region are when $(x_1 - x_2)$ and $(x_2 + x_1)$ have different signs. So $\phi_6(x) = \text{sign}((x_1 - x_2)(x_2 + x_1))$.

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 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 θ.

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 θ.

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:

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

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

Answer: The log likelihood is

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

So first term doesn't depend on θ, second is linear in θ, and third is concave in θ, 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 λ has $\lambda^\top S(x) = 0$ for all x if and only if $\lambda = 0$.

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.

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 λ and γ. 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 λ, 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 (**??**) back into the Lagrangian,

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

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,

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

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 $\lambda_j^+ + \lambda_j^- = |\lambda_j|$.