SMLDM HT 2014 - Part C Problem Sheet 4

1. Complete Question 6 of Problem Sheet 3 on implementing the EM algorithm to model a collection of handwritten digits using a mixture of products of Bernoullis.

2. Assume you have used some data $D = \{x_i, y_i\}_{i=1}^n$ where $x_i \in \mathbb{R}^p$ and $y_i \in \{1, 2, ..., K\}$ to train a classifier. We are interested in classifying a new input vector x_t. However we have only been able to collect $p-1$ features, say $(x_{t2}, ..., x_{tp})$ and x_{t1} is missing. Explain whether or not it is possible to use your classifier to classify this incomplete input vector in the following scenarios. If it is possible, how do you classify the incomplete test vector?

   You may find the Gaussian identities cheat sheet useful if you are not very familiar with properties of the multivariate normals. You do not need to calculate integrals in this question.

   (a) A naïve Bayes model, with
   $$f(x|\phi_k) = \prod_{j=1}^{p} f(x_j|\phi_{kj}),$$
   i.e. conditioned upon $Y = k$, you assume that the features are independent and feature x_j follows a distribution with density/probability mass function $f(x_j|\phi_{kj})$.

   (b) A LDA model, i.e.
   $$f(x|\phi_k) = \mathcal{N}(x; \mu_k, \Sigma)$$

   (c) Generally, which conditions on $f(x|\phi_k)$ are necessary to allow us to implement easily, that is without using numerical integration, a probabilistic classifier like LDA and naïve Bayes in the presence of missing features?

   (d) A logistic regression model, i.e.
   $$p(Y = y|X = x) = s(y(a + b^\top x))$$
   where $y \in \{+1, -1\}$.

3. Consider using logistic regression to model the conditional distribution of binary labels $Y \in \{+1, -1\}$ given data vectors X.

   (a) 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.

   (b) Suppose the first entry in X is binary, i.e. it takes on only values 0 or 1. Suppose that in the dataset, whenever $y_i = -1$, the corresponding entry $x_{i1} = 0$, but there are data cases with $y_i = +1$, and x_{i1} taking on both values. Show that the maximum likelihood estimator of b_1 is ∞, but that the dataset need not be linearly separable.

4. 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.

5. 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.