SDMML HT 2015 - MSc Problem Sheet 3

1. Show that under a Naïve Bayes model with binary predictors from the lectures, the Bayes classifier $f_{\text{Bayes}}(x)$ minimizing the total risk for the $0-1$ loss has a linear discriminant function of the form

$$f_{\text{Bayes}}(x) = \arg\max_{k=1,2} \; a_k + b_k^\top x.$$

and find the values of a_k, b_k.

2. Data in the table gives attributes of 14 customers of an electronics shop and a class label on whether or not they purchased a computer. Using a Naïve Bayes classifier, predict whether a new customer $x = (\texttt{age=youth}, \texttt{income=medium}, \texttt{student=yes}, \texttt{credit=fair})$ will buy a computer?

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

3. Assume we trained a classifier using data $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ where $x_i \in \mathbb{R}^p$ and $y_i \in \{1, 2, ..., K\}$. We are interested in classifying a new input vector $\tilde{x}$. However, we have only been able to collect $p-1$ features, say $\left(\tilde{x}^{(2)}, ..., \tilde{x}^{(p)}\right)$ and $\tilde{x}^{(1)}$ is missing. Explain whether or not it is possible to use the trained classifier to classify this incomplete input vector in the cases listed below. If it is possible, how do you classify the incomplete test vector?

   Note: You do not need to calculate any integrals in this question.

   (a) A naïve Bayes model, with

$$g_k(x) = \prod_{j=1}^{p} p(x^{(j)}|\phi_{kj}),$$

   i.e. conditioned upon $Y = k$, you assume that the features are independent and feature $x^{(j)}$ has probability mass function/density $p(x^{(j)}|\phi_{kj})$.

   (b) An LDA model, i.e.

$$g_k(x) = \mathcal{N}(x; \mu_k, \Sigma)$$

(c) Generally, what condition on the conditional density/pmf $g_k(x)$ would allow easy classification (i.e, without numerical integration) in the presence of missing features for generative classifiers like LDA or naïve Bayes?

(d) A logistic regression model, i.e.

$$p(Y = y|X = x) = s(y(a + b^\top x))$$

where $y \in \{+1, -1\}$.

4. Consider using logistic regression model

$$p(Y = y|X = x) = s(y(a + b^\top x))$$

for 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 $X^{(1)}$ 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 $x_i^{(1)} = 0$, but there are data cases with $y_i = +1$, and $x_i^{(1)}$ taking on both values. Show that the maximum likelihood estimator of b_1 is ∞, but that the dataset need not be linearly separable.

5. Parameter C in C-SVM can sometimes be hard to interpret. An alternative parametrization is given by ν-SVM:

$$\min_{w,b,\rho,\xi} \left(\frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^{n} \xi_i \right)$$

subject to

$$\begin{aligned} \rho &\geq 0, \\ \xi_i &\geq 0, \\ y_i \left(w^\top x_i + b \right) &\geq \rho - \xi_i. \end{aligned}$$

(note that we now directly adjust the constraint threshold ρ).

Using complementary slackness, show that ν is an upper bound on the proportion of non-margin support vectors (margin errors) and a lower bound on the proportion of all support vectors with non-zero weight (both those on the margin and margin errors). You can assume that $\rho > 0$ at the optimum (non-zero margin).

6. 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

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

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

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

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

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

Express the predicted y values on the training set, $\mathbf{\Phi}\hat{\beta}$, only in terms of $\mathbf{Y}$ and the Gram matrix $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^\top$, with $\mathbf{K}_{ij} = \varphi(x_i)^\top \varphi(x_j) = k(x_i, x_j)$ where k 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.