SDMML HT 2016 - Part C Problem Sheet 4

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

2. 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.

3. 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).

4. 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.

5. (*Representer Theorem*) Consider general regularized empirical risk minimisation problem

$$\min_{w,b} R^{\text{emp}}(f_{w,b}) + \Omega(\|w\|_2^2),$$

where $R^{\text{emp}}(f_{w,b}) = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f_{w,b}(x_i)\right)$ is the empirical risk, Ω is a non-decreasing function, and $f_{w,b}(x) = \text{sign}(w^\top \varphi(x) + b)$, for some feature map $x \mapsto \varphi(x)$.

(a) Show that the optimal w lies in the subspace spanned by the training feature vectors, i.e., it takes form

$$w = \sum_{i=1}^{n} a_i \varphi(x_i).$$

(b) Denote $k(x, x') = \varphi(x)^\top \varphi(x')$. Express the decision function $f_{w,b}(x)$ in terms of k and the regularization term $\Omega(\|w\|_2^2)$ in terms of the kernel matrix.

Optional

6. (*One-Class SVM*) A Gaussian RBF kernel on $\mathcal{X} = \mathbb{R}^p$ is given by

$$k(x, x') = \exp\left(-\gamma \left\|x - x'\right\|^2\right). \tag{1}$$

(i) What is $k(x, x)$ for this kernel? What can you conclude about the norm of the features $\varphi(x)$ of x? What values can the angles between $\varphi(x)$ and $\varphi(x')$ take? Sketch the set $\{\varphi(x) : \ x \in \mathcal{X}\}$ as if the features lived in a 2D space.

(ii) Let $\{x_i\}_{i=1}^n$ be a set of points in $\mathcal{X} = \mathbb{R}^p$ (no labels are given). The one-class Support Vector Machine (SVM) is a method for outlier detection which in its primal form is defined as

$$\min_{w,\xi,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n} \xi_i - \rho, \quad \text{subject to } \langle w, \varphi(x_i)\rangle \geq \rho - \xi_i, \ \xi_i \geq 0,$$

where ν is a given SVM parameter, features $\varphi(x)$ correspond to the RBF kernel in (1), and ξ_i's are the non-negative slack variables. The fitted hyperplane $\langle w, \varphi(x)\rangle - \rho$ in the feature space separates the majority of points from the origin (while pushing away from the origin as much as possible) and is used to determine "atypical" x-instances.

Using the 2D intuition from (i), sketch the corresponding hyperplane in the feature space and annotate with ρ, w and a non-zero slack ξ_j for an "outlier" x_j. Would it make sense to use the one-class SVM with a linear kernel?

(iii) Write the dual form of the one-class SVM.