SC4/SM8 Advanced Topics in Statistical Machine Learning
Problem Sheet 1

1. For a given loss function L, the risk R of real-valued $f : \mathcal{X} \to \mathbb{R}$ is given by the expected loss

$$R(f) = \mathbb{E}\left[L(Y, f(X))\right].$$

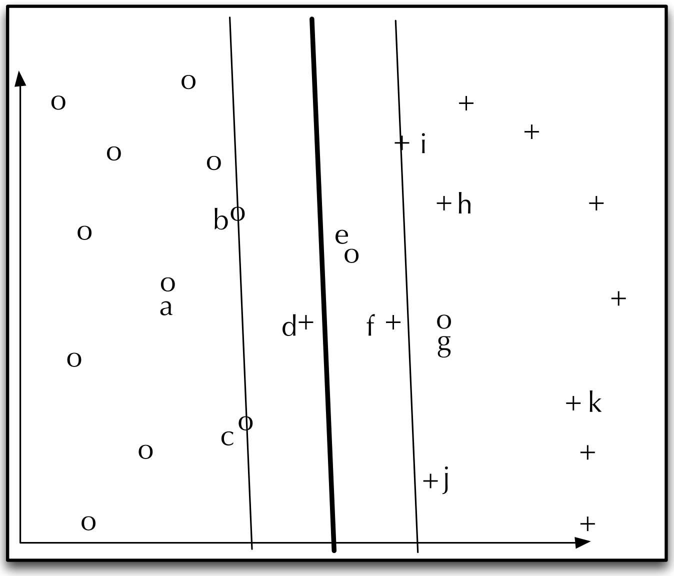
Derive the optimal regression functions (which minimize the true risk) for the following losses:

   (a) The squared error loss
$$L(Y, f(X)) = (Y - f(X))^2$$

   (b) The τ-pinball loss, for general $\tau \in (0, 1)$, given by
$$L(Y, f(X)) = 2 \max\left\{\tau(Y - f(X)), (\tau - 1)(Y - f(X))\right\}.$$

   What happens in the case $\tau = 1/2$?

2. The figure below shows a binary classification dataset and the optimal the decision boundary and margins of a soft-margin C-SVM for some value C.



   (a) Which of the points a,…,k are support vectors? Which ones are margin support vectors?

   (b) For points a, b and d what are the range of possible values for the corresponding dual variables?

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. Consider the regression problem to the real-valued output $y \in \mathbb{R}$. Let $\epsilon > 0$ and define the ϵ-insensitive loss function L_ϵ as

$$
L_\epsilon(y, f(x)) = \begin{cases} 0 & \text{if } |y - f(x)| < \epsilon, \\ |y - f(x)| - \epsilon & \text{otherwise,} \end{cases}
$$

and the regularized empirical risk objective defined as

$$
J(w, b) = C \sum_{i=1}^{n} L_\epsilon(y_i, f(x_i)) + \frac{1}{2} \|w\|_2^2,
$$

where we used a linear model $f(x) = w^\top x + b$ for regression functions.

   (a) Introduce the slack variables $\xi_i^+ = \max\{y_i - f(x_i) - \epsilon, 0\}$ and $\xi_i^- = \max\{f(x_i) - y_i - \epsilon, 0\}$. Verify that $L_\epsilon(y_i, f(x_i)) = \xi_i^+ + \xi_i^-$.

   (b) Re-express the regularized empirical risk objective $J(w, b)$ as a constrained optimization problem over w, b, ξ^+ and ξ^-. Write down Lagrangian and show that the dual problem can be written as

$$
\max_{\alpha^+, \alpha^-} \left\{ -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-) x_i^\top x_j + \sum_{i=1}^{n} (\alpha_i^+ - \alpha_i^-) y_i - \epsilon \sum_{i=1}^{n} (\alpha_i^+ + \alpha_i^-) \right\},
$$

   subject to

$$
\sum_{i=1}^{n} (\alpha_i^+ - \alpha_i^-) = 0, \quad \alpha_i^+ \in [0, C], \quad \alpha_i^- \in [0, C], \quad i = 1, \ldots, n.
$$

   (c) Considering derivatives of the Lagrangian and complementary slackness, express the weight vector w using dual coefficients α_i^+ and α_i^-. Show that those examples (x_i, y_i) which lie outside of the ϵ-insensitive tube around f, must have corresponding $\alpha_i^+ = C$ or $\alpha_i^- = C$ and that those examples (x_i, y_i) for which $|f(x_i) - y_i)| < \epsilon$ (they lie strictly inside the ϵ-tube), must have $\alpha_i^+ = \alpha_i^- = 0$. How can you compute b using the dual solution?

5. **(Kernel Ridge Regression)** Let $(x_i, y_i)_{i=1}^n$ be our dataset, with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Classical linear regression can be formulated as empirical risk minimization, where the model is to predict y using a class of functions $f(x) = w^\top x$, parametrized by vector $w \in \mathbb{R}^p$ using the squared loss, i.e. we minimize

$$
\hat{R}(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - w^\top x_i)^2.
$$

(a) Show that the optimal parameter vector is

$$\hat{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

where $\mathbf{X}$ is a $n \times p$ matrix with ith row given by $x_i^\top$, and $\mathbf{y}$ is a $n \times 1$ column vector with i-th entry y_i.

(b) Consider regularizing our empirical risk by incorporating an L_2 regularizer. That is, find w minimizing

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

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

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

(c) Suppose that we now wish to introduce nonlinearities into the model, by transforming $x \mapsto \varphi(x)$. Let $\mathbf{\Phi}$ be a matrix with ith row given by $\varphi(x_i)^\top$. The optimal parameters $\hat{w}$ would then be given by (previous part):

$$\hat{w} = (\mathbf{\Phi}^\top \mathbf{\Phi} + \lambda I)^{-1} \mathbf{\Phi}^\top \mathbf{y}.$$

Can we make predictions without computing $\hat{w}$?

First, express the predicted y values on the training set, $\mathbf{\Phi}\hat{w}$, 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. Then, compute an expression for the value of $y_\star$ predicted by the model at an unseen test vector $x_\star$.

[*Hint: 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.]

6. Denote $\sigma(t) = 1/(1 + e^{-t})$. Verify that the ERM corresponding to the logistic loss over the functions of the form $f(x) = w^\top \varphi(x)$ can be written as

$$\min_w \sum_{i=1}^{n} -\log \sigma(y_i w^\top \varphi(x_i)) + \lambda \|w\|_2^2 \tag{1}$$

and is a convex optimisation problem in w. Assume that you can write $w = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$. Show that the criterion in (1) is also convex in the so called dual coefficients $\alpha \in \mathbb{R}^n$. [*Hint:* $\sigma'(t) = \sigma(t)\sigma(-t)$]