Problem 0.

Consider a two-class classification problem. The training data is $\mathcal{L}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$, where each $y_i \in \{-1, 1\}$ and $x_i \in \mathbb{R}^d$. We consider classification made using linear models of the form $f(x) = \beta_0 + \beta^t x$: classify x as $+1$ if $f(x)$ is positive, and as -1 otherwise. We assume for now that the data is linearly separable. The SVM maximum margin classifier choses the decision boundary for which the margin is maximized: among all separating hyperplanes, it returns the one that makes the biggest gap (or margin M) between the two classes

$$\begin{aligned} \text{maximize}_{\beta_0,\beta} \quad & M \\ \text{subject to} \quad & \sum_{i=1}^{p} \beta_i^2 = 1 \\ & y_i(\beta_0 + \beta^t x_i) \geq M\,, \quad i = 1, \dots, n \end{aligned} \tag{1}$$

(a) Show that the optimisation problem (1) can be reexpressed as

$$\begin{aligned} \text{minimize}_{\beta_0,\beta} \quad & \frac{1}{2}||\beta||^2 \\ \text{subject to} \quad & y_i(\beta_0 + \beta^t x_i) \geq 1\,, \quad i = 1, \dots, n\,. \end{aligned} \tag{2}$$

(b) Write down the Lagrangian of problem (2).

(c) State the KKT conditions as they apply to this problem.

(d) Derive the dual problem of (2).

(e) Does strong duality hold? Explain why/why not.

(f) Using complementary slackness, find which points in the training data contribute to the optimal solution (that is, find the *support vectors*). Express the optimal β in terms of the training data points and the optimal Lagrange multipliers.

(g) Suggest an expression for the optimal intercept β_0.

(h) How would the expression of the dual problem derived in *(d)* change if you decided to use a kernel SVM approach?

(i) Derive an expression of the margin in terms of the optimal values of the Lagrange multipliers.

(j) Suppose that the data is linearly non-separable. Introducing *slack variables*, suggest a modification to the optimization problem (2) that allow some training points to be misclassified.

Problem 1.

(i) Derive the Lagrange dual problem of a linear program in the inequality form

$$\begin{array}{ll} \text{minimize} & c^t x \\ \text{subject to} & Ax \preceq b\,. \end{array}$$

(ii) Verify that the Lagrange dual of the dual is equivalent to the primal problem.

(iii) When does strong duality hold?

Problem 2.
Consider the following optimization problem in $\mathbb{R}^2$

$$\begin{array}{ll} \text{minimize} & J(x,y) = x + y \\ \text{subject to} & g_1(x,y) = (x-1)^2 + y^2 - 1 \leq 0 \\ & g_2(x,y) = (x+4)^2 + (y+3)^2 - 25 \leq 0 \end{array}$$

(i) Show that the feasible set is convex and sketch it.

(ii) Derive the KKT conditions for this problem, and deduce the solution(s) to the problem.

Problem 3.
Let X be a discrete random variable such that $\mathbf{P}(X = j) = p_j$, and let $A = (A_{ij})$, where $A_{ij} = f_i(j)$, for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. We want to find the distribution $p = (p_i)$ with maximum entropy (the closest to the uniform distribution), under the constraint $\mathbf{E}(f_i(X)) \leq b_i$, for $i = 1, \ldots, m$,

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{i=1}^{n} p_i \log(p_i) \\ \text{subject to} & Ap \preceq b \\ & 1^t p = 1 \end{array}$$

(i) Show that the dual problem simplifies to

$$\begin{array}{ll} \text{maximize} & -\,b^t\lambda - \log\left(\displaystyle\sum_{i=1}^{n} e^{-a_i^t\lambda}\right) \\ \text{subject to} & \lambda \succeq 0\,, \end{array}$$

where a_i is the i-th column of A.

(ii) Under which condition(s) is the optimal gap zero?

Problem 4.
In a Boolean linear program, the variable x is constrained to have components equal to 0 or 1,

$$\begin{aligned} \text{minimize} \quad & c^t x \\ \text{subject to} \quad & Ax \preceq b \\ & x_i \in \{0,1\}\,, \quad i = 1,\ldots,n\,. \end{aligned}$$

Although the feasible set is finite, this optimization problem is in general difficult to solve. We refer this problem to as the *Boolean LP*. We investigate two methods to obtain a lower bound on the optimal solution.

In the first method, called *relaxation*, the constraint that x_i is 0 or 1 is replaced with the linear inequalities $0 \leq x_i \leq 1$,

$$\begin{aligned} \text{minimize} \quad & c^t x \\ \text{subject to} \quad & Ax \preceq b \\ & 0 \leq x_i \leq 1\,, \quad i = 1,\ldots,n\,. \end{aligned}$$

We refer to this problem as the *LP relaxation*. This problem is by far easier to solve than the original problem.

(i) Show that the optimal value of the LP relaxation is a lower bound on the optimal value of the Boolean LP.

(ii) What can you say about the Boolean LP if the LP relaxation is infeasible?

The Boolean LP can be reformulated as

$$\begin{aligned} \text{minimize} \quad & c^t x \\ \text{subject to} \quad & Ax \preceq b \\ & x_i(1-x_i) = 0\,, \quad i = 1,\ldots,n\,, \end{aligned} \tag{3}$$

which has quadratic equality constraints.

(iii) Find the Lagrange dual function of problem (3), and show that the dual problem can be written

$$\begin{aligned} \text{maximize} \quad & -b^t\lambda + \sum_{i=1}^{n} \min\{0, c_i + a_i^t\lambda\} \\ \text{subject to} \quad & \lambda \succeq 0\,, \end{aligned}$$

where a_i represents the i-th column of A.

(iv) The optimal value of the dual of problem (3) provides a lower bound on the optimal value of the Boolean LP. This method of finding a lower bound is called *Lagrangian relaxation*. Show that the lower bound obtained using Lagrangian relaxation is the same as the lower bound obtained using LP relaxation.

Problem 5.

We extend SVM to regression problems. In regularised linear regression, the error function is given by

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda||\beta||^2\,, \quad \text{where} \quad f(x_i) = \beta_0 + x_i^t\,\beta\,.$$

The quadratic error function is replaced with an error function which gives zero error if $|y_i - f(x_i)|$ is less than some $\epsilon > 0$, and a linear penalty otherwise,

$$E_\epsilon(y_i - f(x_i)) = \begin{cases} 0 & \text{if } |y_i - f(x_i)| < \epsilon \\ |y_i - f(x_i)| - \epsilon & \text{otherwise}\ . \end{cases}$$

The problem is now to minimise the following regularised error function

$$C\sum_{i=1}^{n} E_\epsilon(y_i - f(x_i)) + \frac{1}{2}||\beta||^2\,, \quad \text{where} \quad f(x_i) = \beta_0 + x_i^t\,\beta\,,$$

with $C > 0$ some regularisation parameter. For each observation x_i, we introduce two slack variables $\xi_i \geq 0$ and $\hat{\xi}_i \geq 0$, where $\xi_i > 0$ corresponds to a point for which $y_i > f(x_i) + \epsilon$, and $\hat{\xi}_i > 0$ to a point for which $y_i < f(x_i) - \epsilon$

(i) Show that the error function to minimise for support vector regression can be reexpressed as

$$C\sum_{i=1}^{n}(\xi_i + \hat{\xi}_i) + \frac{1}{2}||\beta||^2\,,$$

subject to $\xi_i \geq 0$, $\hat{\xi}_i \geq 0$, $y_i \leq f(x_i) + \epsilon + \xi_i$ and $y_i \geq f(x_i) - \epsilon - \hat{\xi}_i$.

(ii) Write down the expression of the Lagrangian function, and the associated KKT conditions.

(iii) Write down the dual optimisation problem.

(iv) Express the optimal solution $f^*(x_i)$ for the optimal vector of coefficients β^*. You do not need to return an expression for the optimal intercept β_0^* at this stage.

(v) Provide a detailed analysis of the solution: which points have Lagrange multipliers strictly positive? equal to zero? equal to C? Which points are support vectors?

(vi) Suggest an expression for the optimal intercept β_0^*.