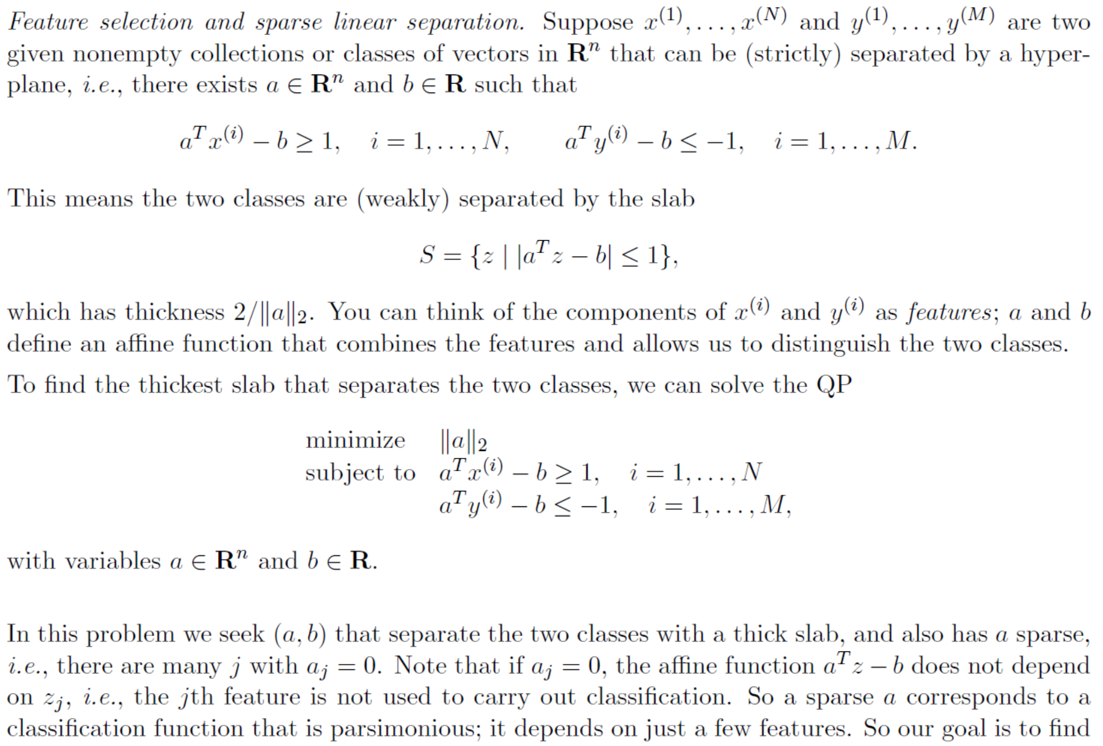
Q2



Feature selection and sparse linear separation. Suppose $x^{(1)}, \dots, x^{(N)}$ and $y^{(1)}, \dots, y^{(M)}$ are two given nonempty collections or classes of vectors in $\mathbf{R}^n$ that can be (strictly) separated by a hyperplane, *i.e.*, there exists $a \in \mathbf{R}^n$ and $b \in \mathbf{R}$ such that

$$a^T x^{(i)} - b \geq 1, \quad i = 1, \dots, N, \qquad a^T y^{(i)} - b \leq -1, \quad i = 1, \dots, M.$$

This means the two classes are (weakly) separated by the slab

$$S = \{z \mid |a^T z - b| \leq 1\},$$

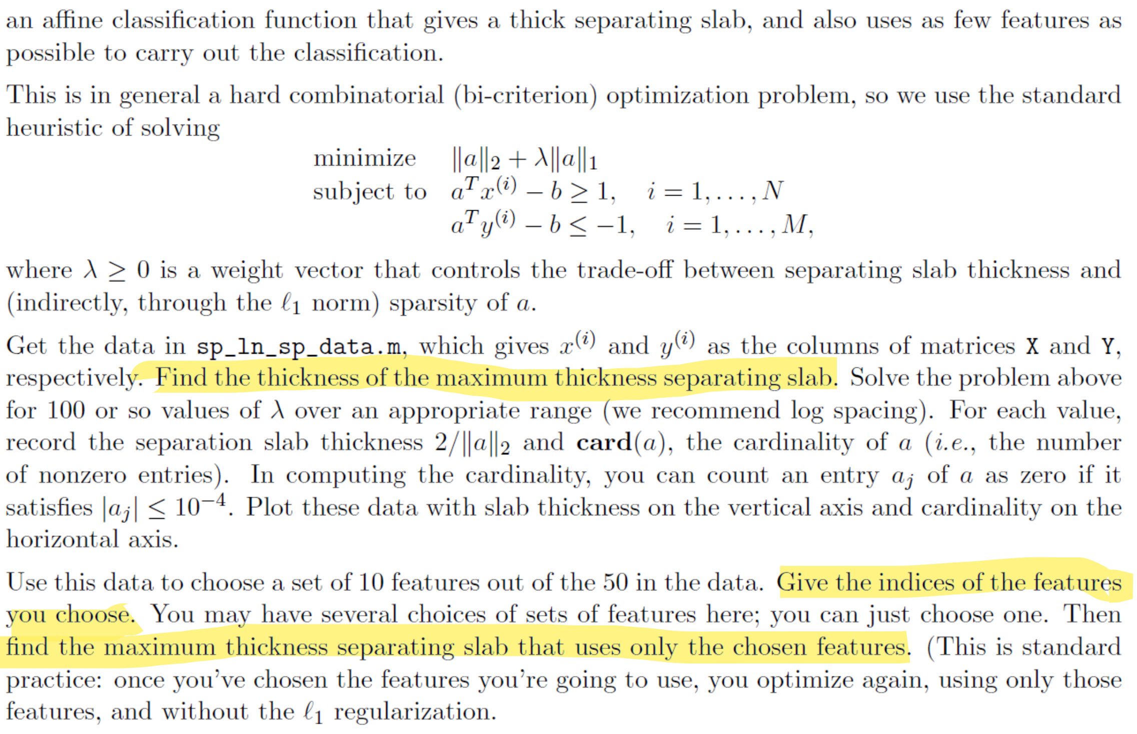
which has thickness $2/\|a\|_2$. You can think of the components of $x^{(i)}$ and $y^{(i)}$ as *features*; a and b define an affine function that combines the features and allows us to distinguish the two classes.

To find the thickest slab that separates the two classes, we can solve the QP

$$\begin{array}{ll} \text{minimize} & \|a\|_2 \\ \text{subject to} & a^T x^{(i)} - b \geq 1, \quad i = 1, \dots, N \\ & a^T y^{(i)} - b \leq -1, \quad i = 1, \dots, M, \end{array}$$

with variables $a \in \mathbf{R}^n$ and $b \in \mathbf{R}$.

In this problem we seek (a, b) that separate the two classes with a thick slab, and also has a sparse, *i.e.*, there are many j with $a_j = 0$. Note that if $a_j = 0$, the affine function $a^T z - b$ does not depend on z_j, *i.e.*, the jth feature is not used to carry out classification. So a sparse a corresponds to a classification function that is parsimonious; it depends on just a few features. So our goal is to find



an affine classification function that gives a thick separating slab, and also uses as few features as possible to carry out the classification.

This is in general a hard combinatorial (bi-criterion) optimization problem, so we use the standard heuristic of solving

$$\begin{array}{ll} \text{minimize} & \|a\|_2 + \lambda \|a\|_1 \\ \text{subject to} & a^T x^{(i)} - b \geq 1, \quad i = 1, \dots, N \\ & a^T y^{(i)} - b \leq -1, \quad i = 1, \dots, M, \end{array}$$

where $\lambda \geq 0$ is a weight vector that controls the trade-off between separating slab thickness and (indirectly, through the ℓ_1 norm) sparsity of a.

Get the data in `sp_ln_sp_data.m`, which gives $x^{(i)}$ and $y^{(i)}$ as the columns of matrices `X` and `Y`, respectively. Find the thickness of the maximum thickness separating slab. Solve the problem above for 100 or so values of λ over an appropriate range (we recommend log spacing). For each value, record the separation slab thickness $2/\|a\|_2$ and $\mathbf{card}(a)$, the cardinality of a (*i.e.*, the number of nonzero entries). In computing the cardinality, you can count an entry a_j of a as zero if it satisfies $|a_j| \leq 10^{-4}$. Plot these data with slab thickness on the vertical axis and cardinality on the horizontal axis.

Use this data to choose a set of 10 features out of the 50 in the data. Give the indices of the features you choose. You may have several choices of sets of features here; you can just choose one. Then find the maximum thickness separating slab that uses only the chosen features. (This is standard practice: once you've chosen the features you're going to use, you optimize again, using only those features, and without the ℓ_1 regularization.