15.094/SMA5223 Systems Optimization: Models and Computation
Assignment 4 (100 points)
Due April 13, 2004

1 Deriving the Dual of the Linear Classification Problem with Penalties (20 points)

The quadratic model formulation for the linear classification problem with penalties is:

$$\begin{aligned}
\text{QP3}: \quad \text{minimize}_{x,\alpha,\xi} \quad & \frac{1}{2}x^T x + C\sum_{i=1}^{k+m}\xi_i & \\
\text{s.t.} \quad & x^T a^i - \alpha \geq 1 - \xi_i & i = 1,\ldots,k \\
& \alpha - x^T b^j \geq 1 - \xi_{k+j} & j = 1,\ldots,m \\
& \xi_i \geq 0 & i = 1,\ldots,k+m.
\end{aligned}$$

Write down the Lagrangian for this problem, using multipliers $\lambda_1,\ldots,\lambda_k$ for the first set of inequality constraints and $\gamma_1,\ldots,\gamma_m$ for the second set of inequality constraints. Present a complete derivation of the dual problem. Your completed dual problem should look like the following by the time you have finished:

$$\begin{aligned}
\text{D3}: \quad \text{maximize}_{\lambda,\gamma} \quad & \sum_{i=1}^{k}\lambda_i + \sum_{j=1}^{k}\gamma_j - \frac{1}{2}\left(\sum_{i=1}^{k}\lambda_i a^i - \sum_{j=1}^{m}\gamma_j b^j\right)^T\left(\sum_{i=1}^{k}\lambda_i a^i - \sum_{j=1}^{m}\gamma_j b^j\right) \\
\text{s.t.} \quad & \sum_{i=1}^{k}\lambda_i - \sum_{j=1}^{m}\gamma_j = 0 \\
& \lambda_i \leq C \quad i = 1,\ldots,k \\
& \gamma_j \leq C \quad j = 1,\ldots,m \\
& \lambda \geq 0, \quad \gamma \geq 0\ .
\end{aligned}$$

2 Using AMPL and LOQO for pattern classification (30 points)

The purpose of this exercise is give you some experience in using AMPL and LOQO to solve pattern classification problems. You are asked to use these software tools to solve 2-dimensional pattern classification problems for two data sets. The data sets for the pattern classification problem are given in the files `linearPC11.dat` and `linearPC13.dat`. Each of these data sets consists of 220 2-dimensional points. About half of the points have the property P, and the other points do not have the property P. Each row of the file corresponds to a different point. The columns of the file are as follows:

- **Column 1:** data point number
- **Column 2:** binary value: "1" if the point has the property P, "-1" if the point does not have the property P
- **Column 3:** x-coordinate of the point
- **Column 4:** y-coordinate of the point

(a) Construct an AMPL model to solve the pattern classification problem for the data sets `linearPC11.dat` and `linearPC13.dat`. Your AMPL model should have an objective function term for penalties, since the 220 points in one of the two data sets is *not* completely separable.

For this problem, please solve the dual problem of the pattern classification model, namely problem D3 given in the previous exercise (Question 1). Hand in a hardcopy of your AMPL model.

(b) Using $C = 1$, solve your model two different ways:

- First solve the model using LOQO.
- Second, solve the model using an active set algorithm of your own design. The skeleton AMPL code for such an active set algorithm is given in the file `skeleton.mod`. Write your own complete active set method code based on the `skeleton.mod` file and solve the pattern classification problem using your method.

Report the computation time for the two methods and submit a hardcopy of your completed `skeleton.mod`.

(c) Create a picture of your solution and hand it in. In order to create a picture, we have provided you with the MATLAB file `nicefigure.m` to create a picture of your solution. In order to use this MATLAB file, your output must be in the correct format.

Such a format is presented at the end of the file `skeleton.mod`. Study the last part of the `skeleton.mod` file to see how to write the output in the proper format. Also, in order to use the MATLAB file `nicefigure.m`, your output file must be named `answ`. Make sure that this is the name of your output file. Then use the MATLAB file `nicefigure.m` to create a picture of your solution. If you have any difficulties producing a picture using MATLAB, ask the TA for help.

(d) Repeat parts (b) and (c) using penalty values of $C = 10$, $C = 100$, and $C = 1,000$. How does the solution change as you change the penalty value? Why might this be the case?

Note: We suggest that you write your model so that it can handle a problem of arbitrary dimension n and an arbitrary number of points m. (This way, you can re-use your model in other parts of this homework assignment.)

3 Classifying Cancer Cells (25 points)

There are many thousands of applications of the pattern classification problem. One example that comes from medicine is to to create a linear classifier that will allow you to classify a benign or malignant tumor based on characteristics of the tumor cell nuclei.

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Ten real-valued features hace been computed for each cell nucleus:

(a) **Radius** (mean of distances from center to points on the perimeter)

(b) **Texture** (standard deviation of gray-scale values)

(c) **Perimeter**

(d) **Area**

(e) **Smoothness** (local variation in radius lengths)

(f) **Compactness** (perimeter^2 / area - 1.0)

(g) **Concavity** (severity of concave portions of the contour)

(h) **Concave points** (number of concave portions of the contour)

(i) **Symmetry**

(j) **Fractal dimension** ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.

The file `bcdat1.dat` contains the data for 569 cells. 357 of the cells are benign, and 212 are malignant. The first column is the cell sample number. The second column is the cell classification for malignant (1), or benign (-1). The next 30 columns are the cell attributes. For instance, field 3 is Mean Radius, field 13 is Radius Standard Error, field 23 is Worst Radius.

(a) Construct and run a pattern classification model in AMPL to create a linear classifier for these two types of cells, based on the 30 attributes. We suggest that you use a large value of the penalty parameter C in your model. What is the resulting linear classifier? Note that you may find that the data sets are not completely separable.

(b) The file `bcdat1U.dat` contains the attributes of 5 unidentified cell samples. Use your linear classifier from part (a) to classify these 5 samples. What is your classification of each cell in the sample of 5?

(c) What is your confidence level that you have classified the unknown cells correctly?

4 Non-linear Classification (25 points)

The purpose of this exercise is to give you some experience in nonlinear pattern classification.

You should have found that the benign and malignant cells in the previous problem were not completely separable using a linear classifier. It has been demonstrated in the literature that this data set is separable with a Multi-Surface Method linear algorithm. Basically that means that a piece-wise linear function can be used instead of a single linear separator. This finding suggests that the data set should also be separable by a smooth nonlinear classifier.

(a) Following the lecture material, propose a nonlinear transformation to a higher dimension.

Hint: You might want to try a quadratic or cubic transformation. There are several mappings that will work successfully. You may not need to use the full dimensionality of the non-linear classifier.

(b) Construct an AMPL model that will separate the 560 cell into two classes.

(c) Use your non-linear classifier for the data set `bcdat1U.dat`. What is your classification of each cell in the sample of 5? Did any of your classifications change?