Math 160 (De Loera) **Homework 4**
Due date: April 13, 2014

INSTRUCTIONS

All homeworks will have many problems, both theoretical and practical. Programming exercises need to be submitted via SMARTSITE using the assignment boxes.

Write legibly preferably using word processing if your hand-writing is unclear. Be organized and use the notation appropriately. Show your work on every problem. Correct answers with no support work will not receive full credit.

1. CHALLENGE 1: The goal is for you to understand the idea of *Support Vector Machines* (SVMs) and the power of linear optimization for classification of data. In this project, we will use linear programming for *breast cancer diagnosis.* The project will use the Wisconsin Diagnosis Breast Cancer Database (WDBC). The idea is to come up with a discriminant function (a separating plane in this case) to determine if an unknown sample is benign or malignant. In order to do this, you will use part of the data in the above database as a "training set" to generate your separating plane and the remaining part as a testing set to test your separating plane. Attributes 3 to 32 form a 30-dimensional vector representing each case as a point in 30-dimensional real space R^{30}. To generate the separating plane, a training set, consisting of two disjoint point sets B and M in R^{30} representing confirmed benign and malignant cases. The separating plane, to be determined by solving a single linear program in MATLAB or SCIP is based on the formulation proposed in class.

   (a) Formulate the problem as a linear program. Solve the problem using the M and B as a training set from the first 500 cases of the wdbc.data file available from
   `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`
   The last 69 points should be used as a testing set. Solve the linear program and print out the separating hyperplane, and the minimum value of the LP.

   (b) Test the separating plane on the 69 cases of the testing set. Report the number of misclassified points on the testing set. It is probably a good idea if you create an MATLAB file to do this.

   (c) We saw in class that there is a a *quadratic programming* formulation of the separation problem (to maximize margin distance). Formulate the quadratic program in SCIP and try to solve it that way. Is the separating hyperplane better?

2. PROJECT 2: Finding out the top features that determine what is an "spam email"

   You will write an algorithm in SCIP for deciding what are the most important features that distinguish spam email from regular truthful email. You can download the data files from `http://archive.ics.uci.edu/ml/machine-learning-databases/spambase/`

   In there you can find the data file spambase.data that contains information on 4601 emails (1813 are Spam!!) Each row contains 58 attributes: 58 (57 continuous, 1 nominal class label). The last attribute in the very last column of 'spambase.data' denotes whether the e-mail was considered spam (1) or not (0).

   Most of the attributes indicate whether a particular word or character was frequently occurring in the e-mail. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters. For more information see the documentation.

GOAL use the LASSO method to figure out what are the most significant of the characteristics that define the spam emails (e.g., maybe number of consecutive capital letters is overly excessive?). Can you run LASSO directly? You can rewrite the LASSO convex model as a linear program by adding some extra variables!

3. CHALLENGE 3: You need to write a SCIP model to solve the following geometric problem:

   Your problem is packing m of spheres in a box of minimal area. The spheres have a given radius r_i, and the problem is to determine the precise location of the centers x_i. The constraints in this problem are that the spheres should not overlap, and should be contained in a square of center 0 and half-size R. The objective is to minimize the area of the containing box.

   - Show that two spheres of radius r_1, r_2 and centers x_1, x_2 respectively do not intersect if and only if $||x_1 - x_2||_2$ exceeds a certain number, which you will determine.
   - Formulate the sphere packing problem as an optimization model. Is the formulation you have found convex optimization?
   - Using your model write SCIP code to solve the packing problem of five and six circular disks of the same radius inside a square of half-size R. What is the optimal size if the disks have radius 1? Do some drawings using MATLAB of the packings you discovered. Is the solution unique?

4. THEORY PROBLEMS: CONVEXITY vs NON-CONVEXITY As we saw in class, in Data Science and Machine Learning, convex sets and functions play an important modeling role. Here are some problems to help you think more carefully about the properties of convex sets functions:

   - Let C be a nonempty subset of R^n, and let λ_1 and λ_2 be positive scalars. Show that if C is a convex set, then $(\lambda_1 + \lambda_2)C = \lambda_1 C + \lambda_2 C$. Show by example that this need not be true when C is not convex.
   - Show that for x, y positive scalar real numbers $ye^{x/y} = \max_{a>0} a(x+y) - y \cdot a \cdot ln(a)$. Use this to prove that the function $ye^{x/y}$ is convex inside the positive orthant. Let $f(x) = \ln(e^{x_1} + \dots e^{x_n})$. Is this convex?
   - In this problem you need to test whether the following functions are convex or not:
     - The function $s_k : R^n \to R$ which is defined as $s_k(x) = \sum_{i=1}^{k} x_{[i]}$ where $x_{[i]}$ is the i-th largest component of the vector x. HINT: Explore what happens with examples $n = 3, k = 2$.
     - For $n = 2k - 1$ odd Consider the function $\phi : R^n R$ with

       $$\phi(x) = \frac{1}{n} \sum_{i=1}^{n} |x_i - med(x)|$$

       where $med(x)$ is the median of the components of x. HINT: Use s_k.
   - Let F be a closed compact convex set in R^n, show that for all u there is a unique point in F which is closest to u.
   - Consider the optimization problem

     $$\min \; 2x_1 \arctan(x_1) - \ln(x_1^2 + 1) + x_2^4 + (x_3 - 1)^2$$

     $$\textbf{subject to: } x_1^2 + x_2^2 + x_3^2 - 4 \leq 0, \quad x_1 \geq 0$$

.

Prove that this problem is a convex optimization problem. Use the Karush -Kuhn-Tucker theorem to find an optimal solution.

- Consider the optimization problem

$$\begin{gathered}\min\ x_3\\ \textbf{subject to: } (x_1-1)^2+x_2^2+x_3^2-1\geq 0,\\ (x_1+1)^2+x_2^2+x_3^2-1\geq 0,\\ x_1^2+x_2^2+x_3^2-4\leq 0\end{gathered}$$

.

  - Is the feasible region convex?
  - What is the convex hull of the feasible region?
  - Solve the above optimization problem over the convex hull.

- Find the maximum value of the function $f(x,y)=x^2+y^4+xy$ inside the convex hull of the points $(-1,1)$, $(-1,2)$, $(-2,2)$, $(-3,1)$. Is the maximum unique? What about the minimum value of $f(x,y)$? Does this function have a unique global minimum in R^2.
- Let y be an arbitrary point in $\mathbb{R}^n$ and let us define the function "distance to y" by:

$$d_{\mathbf{y}}(\mathbf{x})=\|\mathbf{x}-\mathbf{y}\|_2^2,\quad \mathbf{x}\in\mathbb{R}^n$$

Show that the function $d_{\mathbf{y}}$ is *strictly convex*. A function f defined over $\mathbb{R}^n$ is strictly convex if and only if for any pair of points $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$ with $\mathbf{x}\neq\mathbf{y}$ and any $\lambda\in(0,1)$:

$$f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y})<\lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y})$$

- Let us consider the polyhedron $P=\{\mathbf{x}\in\mathbb{R}^n\,|\,A\mathbf{x}\leq\mathbf{b}\}$ for some $A\in\mathbb{R}^{m\times n}$ and $\mathbf{b}\in\mathbb{R}^m$. In this problem, we are interested in the following linear program (LP)

$$\max_{\mathbf{x}\in P}\mathbf{c}^\intercal\mathbf{x} \tag{1}$$

We define the *recession cone* P^o associated with P by:

$$P^o\stackrel{\text{def}}{=}\{\mathbf{d}\in\mathbb{R}^n\,|\,\forall\mathbf{x}\in P,\forall\lambda\geq 0,\ \mathbf{x}+\lambda\mathbf{d}\in P\}$$

  a. Show that $P^o=\{\mathbf{d}\in\mathbb{R}^n\,|\,A\mathbf{d}\leq 0\}$.
  b. Show that P^o is a convex set.
  c. Show that the linear program (1) above is unbounded if and only if there exists $\mathbf{d}\in P^o$ such that $\mathbf{c}^\intercal\mathbf{d}>0$.