Assignment 3

Machine Learning COMS 4771

Spring 2014, Itsik Pe'er

Assigned: Feb 12th Due: Class time, Feb 19th

Submission: Your submission folder on Courseworks. Submit folders for `Assignment03.Problem01, Assignment03.Problem02` and optionally `Assignment03.Problem03` as in previous assignment. Further print `Assignment03.Problem03` if you need it for the quiz.

1) Simulation:

a) Write a function to simulate *N* uniformly-drawn points within bound polyhedra. The function `SimPolyHedra` would proceed by a 1st stage of uniformly drawing (possibly >*N*) points within a box that is bounding the desired polyhedra, followed by a 2nd stage of filtering, so that the function outputs only the first *N* of those points that fall within the polyhedra. The 2nd stage would retain a point iff it is in any of $p \geq 1$ convex polyhedra that are provided as input. The input arguments are:

`N:` The number of points to simulate

`Bounds:` A real $D\times 2$ matrix, each of its rows specifying the (lower,upper) bounds of a *D*-dimensional box from which all points are drawn at the 1st stage

`Polyhedra:` A cell-array of real matrices $M^1,\ldots,M^p$. M^i is of size $f_i\times(D+1)$. It defines a convex polyhedron of f_i faces, as all the vectors x in $\boldsymbol{R}^D$ such that $M^i\begin{bmatrix}1\\x\end{bmatrix} \geq \vec{0}$. Each face is thus defined by a row of M^i interpreted as a hyperplane in $\boldsymbol{R}^D$. Note that M^i may be unbounded, by devfined faces, as we only consider its intersection with the `Bounds` box.

Output:

`X:` A real $N\times D$ matrix, each of its rows specifying a vector that is inside the `Bounds` box and inside at least one of the polyhedra M^i . To generate those you draw potential *D* dimensional vectors `x` whose transpose could serve as rows of `X` . You would then check whether to include each row, i.e. whether any for each such `x` there exists at least one M^i such that the $M^i\begin{bmatrix}1\\x\end{bmatrix} \geq \vec{0}$ condition is satisfied.

[20 points]

b) Use the above to write a function `SimTanzania()` that simulate points of particular colors in the Tanzanian flag (see attached Tanzania.pdf). This flag spans the axis-bounded rectangle between (0,0) and (1.5,1.0), and has 5 regions in green, yellow, black, yellow and cyan, separated by the lines $y=\frac{2}{3}x-0.24, y=\frac{2}{3}x-0.16, y=\frac{2}{3}x+0.16, y=\frac{2}{3}x+0.24$. Draw points inside the planar rectangle of the flag, and save 4 text files, each of N=50 rows and D=2 columns of numbers in text, specifying 50 points of the appropriate color: `Tanzania_green.txt`, `Tanzania_cyan.txt`, `Tanzania_black.txt` and `Tanzania_yellow.txt` .

[5 points]

2) Probabilistic interpretation:
   a) Consider `SimPoly` of Assignment2, Question 1 as defining a probability space of potential outputs `y`. As such, it defines a probability density function $f(y)$ over possible values of `y`= y . Of course, this probability space is different for each input, so $f(y)$ depends on the inputs `RealThetas, sigma` and `x`. Denote it $f_{\theta,\sigma,x}(y)$ for `RealThetas`= θ, `sigma`=σ and `x`=x. Prove that the least-squares regression result $\theta = \theta^*$ maximizes $f_{\theta,\sigma,x}(y)$ for any σ, x and y.
   [15 points]
   b) Consider `SimLogistic` of Assignment2, Question 3 , with zero noise, as defining a probability space of potential outputs `y`. As such, it defines a probability function $P(y)$ over possible values of `y`= y . Of course, this probability space is different for each input, so $P(y)$ depends on the inputs `RealThetas` and `x`. Denote it $P_{\theta,x}(y)$ for `RealThetas`= θ and `x`=x. Prove that the logistic regression result $\theta = \theta^*$ maximizes $P_{\theta,x}(y)$ for any x and y.
   [15 points]

   Guidance: Neither 2a nor 2b requires computing derivatives. Both can be solved by the definition of θ^* as an ERM, so you are welcome to just use that.

3) Optional:
   Prepare a single sided, single page, 12-font English-letter (with potential notations in Greek) cheat sheet for the quiz scheduled for Feb 19th. This would be the only allowed material.
   [0 points]

Good luck!