CMPS 142 First Homework, Winter 2016

4 Problems, 15 pts, due start of class Monday 1/18

This homework is to be done in groups of 2 or 3. Each group members should completely understand the group's solutions and *must* acknowledge all sources of inspiration, techniques, and/or helpful ideas (web, people, books, etc.) other than the instructor, TA, and class text. Each group should submit a single set of solutions containing the names and e-mail addresses of all group members. Although there are no points for "neatness", the TA may deduct points for illegible or poorly organized solutions.

1. (2 pts) Prove that for an arbitrary number of examples m and number of features n, and set of m examples with n (real-valued) features, that the Least Squares cost function $J(\theta)$ is a convex function of the n-dimensional parameter vector θ.

   Recall that

   $$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

   and the hypothesis $h_\theta(x)$ is $\theta \cdot x$. You may either show that the Hessian H is positive semi-definite, or that for any $t \in [0, 1]$ and any parameter vectors v and w, we have $tJ(v) + (1-t)J(w) > J(tv + (1-t)w)$ (which is the way I did it). It may be helpful to prove it first for single examples and then use the property that the sum of convex functions is also a convex function.

2. (5 pts) Linear Regression.

   Download and install the Weka package of machine learning algorithms. We will be using Weka later, so each student should do this problem and combine solutions. Use data file:

   ```
   %—
   @relation 'hw1regression'
   @attribute 'x1' real
   @attribute 'x2' real
   @attribute 'x3' real
   @attribute 'y' real
   @data
   3, 9, 2, 19
   6, 9, 1, 19
   7, 7, 7, 10
   8, 6, 4, 11
   1, 0, 8, -3
   %—
   ```

   The data is presented in the aarf format used by weka, the first lines are header information naming the data and the features, and each example is given as a row

with comma-separated features. Note that the x_0 constant features are not part of the data.

There are 3 features: x_1, x_2, x_3. The target is y. There are 5 instances. The first instance has features $x_1 = 3, x_2 = 9, x_3 = 2$ and target $y = 19$.

Note: This is artificial & slightly cherry picked data. For each instance i, the features $\mathbf{x}_i$ were generated randomly and then noise was added. The targets were generated by the following function before the noise: $y = 0 * x_{1\text{ (clean)}} + 2 * x_{2\text{ (clean)}} - 1 * x_{3\text{ (clean)}} + 3$. Your computed weight vector should be somewhat similar to this, but with so few examples and so much relative noise you shouldn't expect to get really close.

(a) Save the data to an arff file and run the linear regression algorithm in weka on the full training set. Report the model and "Root mean squared error". Note that testing on the training set means we will have a relatively small error.

(b) Suppose you had an unlabeled instance $\mathbf{x} = [3, 3, 5]$. What $\hat{t}$ prediction for t would the model from part (a) give?

(c) Stochastic Gradient Descent. Start with the weight vector $w = [1, 1, 1, 1]$ (using the add-a-dimension trick were the added x_0 bias component is 1). Step through stochastic gradient descent on the first two instances ($x = [1, 3, 9, 2], y =$ and $x = [1, 6, 9, 1], y = 19$). For your learning parameter, use $\eta = 0.1$, For your error function use squared error. Report the updated weights after each instance. Show your work for at least the first iteration.

(d) (3 pts) Compute minimizing the least squares criterion for the data in problem 1 using the formula in Chapter 2 of Ng's notes: $\theta = (X^\mathsf{T} X)^{-1} X^\mathsf{T} \mathbf{y}$.

Here X is the (5-row, 4-column) matrix whose rows are the unlabeled instances and whose columns are each of the features for the instance augmented by the fixed feature $x_0 = 1$ whose weight is the bias. How does this θ compare to the weights from part a? If you like, you may use your favorite math software for the matrix arithmetic but it is quite feasible to do directly. You should not compute the inverse by hand (Google "inverse matrix 4x4 calculator" for some tools).

(e) If the examples are re-ordered (so the rows of X and elements of $\mathbf{y}$ are permuted), what happens to the learned θ vector and why?

3. (6 pts) Consider the weighted linear regression problem, where each example has a positive real-valued importance $w^{(i)}$ along with the feature vector $\mathbf{x}^{(i)}$ and label $y^{(i)}$. The importance weights indicate how important it is to fit the labels on the various examples. For example, if some y measurements (or feature measurements) were taken under poor conditions, then our confidence in the accuracy of those training examples may be less than in the other examples. Conversely, we might be more confident in some examples if they have been carefully curated and double checked. This lack of confidence can be represented by assigning the less confident examples a lower $w^{(i)}$ weight.

Assume that the add-a-dimension trick has been used, so the first component of each $\mathbf{x}^{(i)}$ is one, and thus the bias/intercept term is taken care of by θ_0. Then a natural way to the solve the importance-weighted least squares problem is to minimize the criterion:

$$J(\theta) = \sum_{i=1}^{m} w^{(i)} \left(\theta \cdot \mathbf{x}^{(i)} - y^{(i)}\right)^2 \quad .$$

(a) Show that the above $J(\theta)$ can be written as

$$J(\theta) = (X\theta - \mathbf{y})^\mathsf{T} W (x\theta - \mathbf{y})$$

Where X is the design matrix (whose rows are the example features), $\mathbf{y}$ is the vector of labels (for all the examples), and W is a matrix related to the weights. Clearly describe what this matrix W is.

(b) In the standard least squares problem, the minimizer of $J(\theta)$ was $X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y}$. By finding the derivative ∇_θ for the $J(\theta)$ from part a), and setting it to zero, find a similar closed form for the θ minimizing the weighted linear regression $J(\theta)$.

4. (1 pt) (to be done individually) E-mail a picture of yourself (preferably a portrait-style head shot) to the TA, to help us match up names and faces.

Exercises (Not to be turned in):

- Verify that for 2×2 matrices, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.