COMS 4771 HW3

Due: Sun Nov 12, 2017 at 11:59pm

You are allowed to work in groups of (at max) three students. These group members don't necessarily have to be the same from previous homeworks. Only one submission per group is required by the due date. Name and UNI of all group members must be clearly specified on the homework. You must cite all the references you used to do this homework. You must show your work to receive full credit.

1 **[Low-dimensional information-preserving transformations]** (*hashing the cube*) You have a collection of nonzero distinct binary vectors $x_1, \ldots, x_m \in \{0,1\}^n$. To facilitate later lookup, you decide to hash them to vectors of length $p < n$ by means of a linear mapping $x_i \mapsto Ax_i$, where A is a $p \times n$ matrix with 0-1 entries, and all computations are performed modulo 2. Suppose the entries of the matrix are picked uniformly at random (ie, each an independent coin toss)

(i) Pick any $1 \leq i \leq m$, and any $b \in \{0,1\}^p$. Show that the probability (over the choice of A) that x_i hashes to b is exactly $1/2^p$. (Hint: focus on a coordinate $1 \leq j \leq n$ for which $x_{ij} = 1$.)

(ii) Pick any $1 \leq i < j \leq m$. What is the probability that x_i and x_j hash to the same vector? This is called a *collision*.

(iii) Show that if $p \geq 2\log_2 m$, then with probability at least $1/2$, there are no collisions among the x_i. Thus: to avoid collisions, it is enough to linearly hash into $O(\log m)$ dimensions!

(question credit: Prof. Sanjoy Dasgupta)

2 **[Bayesian interpretation of ridge regression]** Consider the following data generating process for linear regression problem in $\mathbb{R}^d$. Nature first selects d weight coefficients $w_1, \ldots, w_d$ as $w_i \sim N(0, \tau^2)$ i.i.d. Given n examples $x_1, \ldots, x_n \in \mathbb{R}^d$, nature generates the output variable y_i as

$$y_i = \sum_{j=1}^{d} w_j x_{i,j} + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$ i.i.d.

Show that finding the coefficients $w_1, \ldots, w_n$ that maximizes $P[w_1, \ldots, w_d | (x_1, y_1) \ldots, (x_n, y_n)]$ is equivalent to minimizing the ridge optimization criterion.

3 **[Spam classification Competition!]** You'll compete with your classmates on designing a good quality spam classifier.

(i) Signup on http://www.kaggle.com with your columbia email address.

(ii) Visit the COMS 4771 competition at: https://www.kaggle.com/t/73829f0970444b1eae914130e6db48ca and develop a spam classifier.

(iii) Your pdf writeup should describe your design for your spam classifier: What word embedding, preprocessing techniques and classifier you used? Why you made these choices? What resources you used and were helpful? What worked and what didn't work?

Evaluation criterion:

- You must use your (and your group members') UNI as your team name in order to get points. For example:
  - If you have two group members with uni: `ab1234` and `xy0987`,
  - the teamname should be: `ab1234_xy0987`
- Your team must get an accuracy score of $\geq 75\%$ on the test-dataset to get full credit.
- Top ten teams (across both sections) will get extra points!