TRINITY COLLEGE DUBLIN

School of Computer Science and Statistics

Week 9 Questions ST3009: Statistical Methods for Computer Science

For each problem, explain/justify how you obtained your answer in order to obtain full credit. In fact, most of the credit for each problem will be given for the derivation/model used as opposed to the final answer.

Question 1. You're asked to write a programming that finds the minimum of the function $f(x) = x^2 - 1$.

(a) Give code (matlab or python are both fine) that uses gradient descent to find the approximate minimum. Note: the derivative $df/dx = 2x$.

(b) Gradient descent has a learning rate parameter α. Run your gradient descent code with $\alpha = 1$, $\alpha = 0.1$ and $\alpha = 0.01$ and for each plot how the function value $f(x)$ changes at each update. Discuss.

(c) An alternative to gradient descent is to pick a point at random in the vicinity of the current point x_k until a new point x_{k+1} is found that causes function $f(x)$ to decrease i.e. such that $f(x_{k+1}) < f(x_k)$. Modify your code to implement this random strategy and give an example plot of how $f(x)$ changes at each update.

(d) Discuss how the randomised approach in (c) compares with gradient descent approach in (b).

Question 2. Your task is to write a machine learning algorithm that classifies hotel reviews as positive (on balance the reviewer likes the hotel) or negative (on balance the reviewer dislikes the hotel). You have been given a set of reviews manually labelled as positive or negative to use as training data. This data has been preprocessed to construct a dictionary consisting of the N distinct words used in the set of the reviews. Using this dictionary a feature vector is constructed for each review. This vector has N entries, one for each word in the dictionary, and the value assigned to an entry is the number of times the word appears in a review.

(a) Using this data describe how you would use logistic regression to classify the hotel reviews.

(b) What statistical assumptions are made by this logistic regression classifier?

(c) Discuss how you could use bootstrapping to estimate a confidence interval for the prediction of the logistic regression classifier for a specified review. Hint: the randomness in the prediction is due to the training data, so consider resampling the training data.