**Introduction to Data Mining**

***Quiz 2, Spring 2016 Total Marks = 15;*** *Time: 15 min.*

*Declaration:* By submitting this quiz for grading, I affirm that I have neither given nor received help from another examinee and acknowledge that **this is a closed-book, closed-notes test.**

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Q1. (3 pts)** Given D as a DataFrame of compatible size/values and the class lasso has been properly instantiated etc., what will the function lasso.predict(D) do?

**This predicts the dependent variable given the independent variables in D. The result is a vector of predictions.**

**Q2. (2+1= 3 pts).** Briefly explain what you understand by “Multi-Dimensional Scaling” and give an example (other than the one in the notes) where this can be a useful procedure.

[**Wikipedia**](https://en.wikipedia.org/wiki/Multidimensional_scaling) **has a very good explanation with examples.**

**Q3 (2+1 pts).** What will you expect will happen to the (i) bias and (ii) variance of a linear regression model as the size of the training set is increased (everything else remains the same)?

**Bias remains unchanged and variance decreases.**

**Q4. (3 pts)** What will be the Bayes Decision Rule for a 3-class classification problem?

**P(Class|data) = P(data|Class)p(Class) / p(data). Calculate this value for each class and choose the class with the highest probability. If you are confused on how Bayes Theorem works, you can view** [**this tutorial.**](http://nbviewer.jupyter.org/github/tfolkman/learningwithdata/blob/master/Bayes_Primer.ipynb)

**Q5. (2+1 = 3pts)** What gradient does SGD compute (i.e. gradient of what function and w.r.t. what variable(s)) ?

Why is this a stochastic procedure?

**SGD computes the gradient of the cost function with respect to the weights (or coefficients, or betas, …). For clarity, the cost function of linear regression is the mean squared error. It is stochastic because it calculates the cost function using only 1 observation at a time (as opposed to looking at all the observations). Since it is only using 1 observation, it isn’t perfectly computing the gradient (given all the data) and thus follows a more random – or stochastic – path to the global minimum (assuming a convex cost function).**