**Practice questions ML Basics resit**

Note that a few practice questions can never cover everything. These open questions and the way I write down the answers can help you refine how you answer questions on the test. Good luck!

Q1. The reason we don’t use the mean-squared error for logistic regression is that it is not convex. Below:  
1. Make two small plots conceptually showing a convex versus a non-convex cost function.  
2. Explain why having a non-convex cost function is problematic for gradient descent.

Q2. Write down the weighted sum (z(l)j) that a neuron calculates. There’s three things you can tweak about this weighted sum in (almost) all neurons in the network to lower the cost function. What are these three things? Write down what they are, and the partial derivatives of the weighted sum with respect to each of these three quantities.

Q3. High-dimensional space is wonky. Explain in your own words:  
1. How/why (Euclidian) distances become virtually meaningless in high-dimensional space  
2. How/why adding more features to your supervised ML algorithm won’t improve performance (unless you drastically increase training data amount)

Q4. This question pertains to classifier performance and comparing it. Explain:

1. Why we cannot just use accuracy to gauge how well our classifier performs in a classification task.
2. Why we compare classifiers using something like the ROC AUC, or other metrics that average over thresholds.

Q5. I have the following data:

|  |  |
| --- | --- |
| **Gene 1 expression** | **Gene 2 expression** |
| 12 | 38 |
| 3 | 49 |
| 9 | 67 |

|  |
| --- |
| **Label** |
| Healthy |
| Diseased |
| Healthy |

1. Write down how you would calculate the logistic regression hypothesis function using matrix multiplication. Use @ for matrix multiplication. You don’t have to calculate the outcomes. Use sigmoid() for the sigmoid function.
2. Why don’t we just use linear regression for a problem like this?
3. Write down the cost function of this logistic regression and its partial derivative w.r.t. theta\_2.
4. I want to compare the logistic regression classifier with other classifiers. I care most about positive samples in a background of many negative samples (think cancer screening). What metric could I use?

Q6. K-fold cross-validation is the basic way we obtain an estimate of our generalisation performance, which is the thing we care about. Describe in your own words what this procedure implicitly assumes about the data your classifier will see once deployed, and modifications so that validation performance is more indicative of performance in the real world.

Q7. I want to optimise the parameters of a univariate linear regression (y=theta\_0 + theta\_1 \* x) using gradient descent. Write down in *pseudocode* how you would calculate the partial derivatives and take one step in the right direction to minimise the cost function.   
Pseudocode means: the logical steps you are taking should be correct, but code syntax doesn’t have to be correct. Note: I for sure won’t ask this for (linear algebra) backpropagation.