MA2823: Foundations of Machine Learning **Homework 1**

Due October 10, 2016.

1 Homework Problem

Question 1

We want to predict the blood glucose level of an adult based on his/her sex (male, female), weight, height and cholesterol. We know the values of all variables for 10,000 adults, but only for 500 of those adults is their blood glucose level available. (i) What kind of machine learning problem is it? (ii) Are the 9,500 instances with missing blood glucose level useful for us?

Solution: (i) We can apply semi-supervised regression. It is a regression problem because the response variable (blood glucose level) is real-valued and not categorical. The categorical variable (sex) is used as input, not output, thus does not alter the problem formulation. The learning follows semi-supervised approach because only 5% of the observations are labeled. (ii) Yes, they are very useful because they allow us to learn the association between the input variables (sex, weight, height and cholesterol). For example it is expected (for healthy individuals) to have a larger weight if they are taller, without this having any effect on blood glucose level. We assume this because it makes sense, but a computer algorithm needs to be trained for it. For this purpose a large number of pairs weight/height is required but not any labeled samples. Once we have learned the expected variation in a healthy population, we can detect deviations as risk factors. For example, high cholesterol and overweight are risk factors for diabetes which results in high blood glucose level. Here, the labeled samples are useful in the learning process.

Question 2

What is the VC dimension of a plane in 3D? Please elaborate on your response.

Solution: The VC dimension of hyperplanes in $\Re^d$ is $d+1$, thus for a plane in $\Re^3$, the VC-dimension is 4. It is easy to understand why it can shatter 4 points except for one exception: all points to lie on a plane and the points within the same class to be anti-diametric. However in this case the rule does not apply, because we are interested if all possible labellings of some n-points can be shattered.

Let's now understand why it cannot shatter more than 4 points. If there are 5 points with 3 of them being in the same class, we can form the surface that goes through those 3 points. If the other 2 points lie on opposite sides, there is no way to shatter them using a plane, thus the VC dimension cannot be 5.