HGDS 200 - Foundations of Data-Driven Analysis 2018/2019 - Handout 3
1/7/2019 - 1/14/2019 Name:

The following should be completed after reading Chapter 3 (Data Pre-processing) of *Applied Predictive Modeling.*

Consider the (obviously fake) model below for predicting risk of acute myocardial infarction within the next 5 years:

$$M(\text{isMale}, \text{LDL}) = f(-9 + 1 \cdot \text{isMale} + 0.05 \cdot \text{LDL})$$

where

- $f(x) = \frac{1}{1+e^{-x}}$ is the logistic function,
- isMale is 1 if the patient is male and 0 otherwise, and
- LDL is the patient's LDL cholesterol in mg/dL. A "healthy" LDL cholesterol is considered to be < 100, while a value of > 160 is considered to be high.

Answer the following questions about the model M, you may find that it is useful to plot $y = f(x)$ to gain an intuition about the logistic function.

(1) Given that the coefficient of the isMale is 20 times larger than the coefficient of LDL, which feature would you say is of greater "importance" to the model? What should we do to the features if we would like this comparison to be more meaningful?

(2) Which patient presents a greater risk: a male with a healthy LDL or a female with a high LDL?

(3) If we wished to more directly measure which feature (being male or having an LDL of 160 or more) presented a greater 5 year MI risk by fitting a logistic regression model on relevant patient outcomes data, how could we do that?

(4) It turns out that being male makes a patient more likely to have a high LDL (we say that isMale is positively correlated with LDL). How might this confuse the way that we can interpret the coefficients?

(5) As an extreme example imagine that LDL appeared twice in the model (that is, we now have two features with perfect correlation). The model now can be written as

$$M(\text{isMale}, \text{LDL}) = f(-9 + 1 \cdot \text{isMale} + \beta_1 \cdot \text{LDL}_1 + \beta_2 \cdot \text{LDL}_2)$$

what could the values of the new coefficients β_1 and β_2 be?

(6) Another consideration when assessing heart health is the patient's HDL cholesterol, which evidence suggests has a protective effect on the cardiovascular system (that is, a higher HDL is indicative of better heart health). If we fit a model with HDL cholesterol (as well as the existing isMale and LDL features), what sign would you expect the coefficient of the HDL term to have?

(7) If a new study suggested that HDL cholesterol was more protective for women than men, how might we adjust the model to account for this?

(8) Suppose that we would now like to model the risk of a hospital inpatient having an adverse event during their stay - such an adverse event could have multiple causes, but we'll define it as an event that results in a transfer to an ICU or in death. The features for this model will be: temperature, systolic BP, respiration rate, and heart rate. Can we model this event in the way described above? How might we rather model this?

(9) Respiration rate is often null for for hospital inpatients. Do you think that this information is structurally missing? How should it be imputed?