Data Mining

Quiz 1, Spring 18 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 (2 pts). What do you understand by the statement that estimate of some numeric quantity “z” obtained by a certain procedure is “unbiased”?

This means that the expectation of z-estimate is equal to the true value of z.

Q2 (3 pts). What are two important problems that arise when two of the independent variables used in MLR are highly and positively correlated?

If features are highly correlated, the least squares optimization problem will be ill-conditioned (meaning the ratio of the largest singular value to the smallest singular value is large). This causes slow convergence and potentially numeric instability.

In terms of the uncertainty of the model, the standard error for highly correlated variables will be high. Consider the denominator (SSE_j) of the formula for standard error of the jth predictor in MLR: SSE_j will shrink to 0 as the correlation between two independent variables increases. This causes the standard error for the jth component to be large. (see slide 14 of “2 mlr-short.pdf”).

In addition, having highly correlated features reduces the interpretability of the model because changes in the dependent variable can be explained by multiple highly correlated features.

If features are exactly linearly dependent, X^TX will be non-invertible, so the closed form solution, $(X^TX)^{-1}X^Ty$, will be undefined.

Q3 (2+2+2 + 2 pts). (use back of page if needed) Suppose I want to use MLR to predict the price of a piece of cloth based on two independent variables, *size* and *quality*. The cloth pieces come in 3 distinct sizes (S) and exhibit 4 distinct levels of quality (Q). I am considering two choices:

Choice 1: encode both S and Q using integers (i.e. 0, 1, 2 for S and 0, 1, 2, 3 for Q).

Choice 2: encode both S and Q by using dummy coding.

I train both models on data representing 1000 pieces of cloth, and then use the trained models to predict the prices of another 1000 pieces (called the “test set”).

(a) What is the maximum number distinct values for price can I obtain for the test set? State your answer for each of the choices.

(b) Name one advantage of Choice 1 over Choice 2, and one advantage of Choice 2 over Choice 1

(c) If I code the lowest values for S and Q as zeros (in Choice 1), and also use this configuration as the baseline for choice 2 (which means they will result in all zeros for the corresponding dummy variables), will I obtain the same value of β_0 for the two trained models?

(d) What do you understand by the use of an “interaction term” in the formulation of an MLR solution?

Solution:

- 12 in both, as there are 12 unique combinations of size and quality regardless of encoding type
- The main advantage of dummy coding (Choice 2) is that you don't need to make an assumption on the quantitative nature of the variable. For instance, maybe the difference in price between the largest size and the medium size is roughly the same as the difference in price between the medium size and the small size, in which case either dummy coding or simple numeric coding will work fine. However, maybe the difference in price between the largest size and the medium size is much greater than the difference between the small size and the medium size. A linear regression will not be able to capture this relationship with a single β. Dummy coding discovers the exact relationship for you.

Numeric coding is still viable when you are confident that a feature definitely possesses a particular numeric relationship. More often, it is used when dummy coding creates too many features to be practical. If the training data is small, Choice 1 is less likely to overfit.

Interpretability was sometimes given as a difference – however, either coding could give a better interpretation depending on the situation.

- No. If the modelling assumption was perfectly valid, and there was a lot of training data to use, then the intercept term β_0 would be very similar between different coding types. However, if the numeric coding does not really fit the feature (as discussed above) or the observation-to-feature ratio is low and noise influences the parameters, the two β_0's can be very different.
- Interaction terms are additional features added as polynomial combinations of the existing features. For eg, x_1, x_2 with interaction terms will be x1,x2, x1*x2, x1^2, x2^2. For categorical variables, this would be a combination of the features. Eg (S=0,Q=0), (S=0,Q=1), (S=0,Q=2) etc

Q4 (2+2 pts). Briefly explain how the bias-variance tradeoff helps explain why you may be able to afford a more complex model (from the same family of models) when you have a larger training set.

With a larger training set, a higher variance (more complex) model is less likely to be overfit to the data. Accordingly, a lower bias, but higher variance model can be chosen when a larger training set is available.