Adavanced Predictive Modelling

Quiz 1, Fall 2017 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+2 pts) Explain briefly what you understand by the "task discovery" and "data discovery" aspects of the data mining process.

(from Brachman and Anand, 96) Task Discovery is one of the first steps of KDD. The client has to state the problem or goal, which often seems to be clear. Further investigation is recommended such as trying to get acquainted with the customer's organisation after having spent some time at the place and then to sift through the raw data (to understand its form, content, organisational role and the sources of data). Then the real goal of the discovery will be found.

Data Discovery complements the step of task discovery, and is where we decide whether the quality of data is satisfactory for the goal (what the data does or does not cover).

Q2. (3 pts). Generally speaking (so just don't say "find mean and variance of a Gaussian"), what is the maximum likelihood principle used for in data science?

The Maximum Likelihood principle is used to estimate the parameters for a model distribution that maximizes the likelihood of it producing a given data set. Here,the likelihood function gives the probability that data points {Xn} drawn i.i.d are distributed from a gaussian model, N(mu,sigma^2). Since distributions are defined by their parameters (ie, mean/variance), discovering these values allow us to define distributions which produce samples most like our data.

Q3 (4+2 pts).

(i) What are the key assumptions regarding the relation between the dependent variable (Y) and the independent variables (Xs) made in the standard (multiple) linear regression model?

(ii) How do these assumptions justify the use of "ordinary least squares" (OLS) to determine the parameters of the model? (just state, no need to derive anything).

Assumptions between Y and Xs in standard multiple linear regression model:
i) The conditional mean of Y is linear in the Xj variables.
The error terms (deviations from true response for each Xj)
- are normally distributed around 0
- independent from each other
- identically distributed (i.e., they have constant variance)
Y|X1...Xp ~ N(B0+ B1X1...+ BpXp, sigma2)

The error term captures the effect of omitted variables

ii) That the error terms are independent and normally distributed with constant variance, makes the Ordinary Least Squares estimate a maximum likelihood estimator for our model parameters since in this case minimizing the overall sum of squared errors (ie, residuals) is equal to maximizing the likelihood function of the data.

Q4 (2 pts). Why is “one-hot coding” of a categorical variable (i.e. where each distinct assignment/value of a categorical variable corresponds to a bit which is 1 if that assignment is encountered and 0 otherwise) problematic for linear regression models?

When the variable we one hot encode is one of our predictors (X1, X2, etc), “one-hot coding” introduces one binary variable per level, which makes the predictors linearly dependent and the model unidentifiable (ie feature matrix is non-invertible). Additionally, standard linear regression assumes the predictor and responses are continuous, and in this case the variables introduced are discrete.

If the one hot coded variable was your response, then you would be in a setting where you would need logistic regression in order to do classification since y.