Doug Woodward

CS613 HW 2

|  |  |
| --- | --- |
| # | Answer |
| 1 | 1. For a large sample size and a small number of predictors it is likely that a more flexible model would yield better accuracy. A flexible model can take advantage of the large sample size and will be unlikely to overfit with a larger sample size. 2. In the case of many predictors and a small number of samples, an inflexible model can be prefered to prevent overfitting. 3. With non-linear relationship, a more flexible model is preferred as an inflexible model will be unable to account for the nonlinear relationship. 4. With high variance in error terms, we know there is a lot of noise in the data. Therefore an inflexible model is desired as a flexible one would capture that noise. |
| 2 | 1. This would be a regression problem due to the need obtain a quantitative response of what factors are the largest contributors to performance, ie the correlation between each variable and outcome. n = 500, p = 4 2. This is a classification problem, the answer required is a qualitative nature, it should be Success or Failure. n = 20 and p = 13 as we do not include the “truth” variable of success/failure as part of our parameters. 3. This is a regression problem as it has a quantitative answer of what percent change will occur. n = 52, and p = 4 |
| 3 | b) Training MSE will usually decrease as the model is more flexible as it will contour exactly to the training data. Testing MSE will usually increase with flexibility as a more flexible model will overfit the training data. The bias decreases as a more flexible model can more accurately represent complex real life situations. Variance increases with increased flexibility as it refers to how different a model will be if given a different training set, therefore a more flexible model will fit closer to the data given and produce greater variance. Total error will be greatest at either tail end as at that point, either bias or variance will dominate the overall equation. |
| 4 | a) Classification would be useful for galaxy identification, as in fact it is used to analyze data from Hubble and the Sloan Digital Sky Survey. Predictors involved are the shape of the galaxy, the distribution of brightness, color range, and number of spiral arms. This is for prediction of what type of galaxy it is on the Hubble tuning fork.  Another application would be identifying software applications whose performance profile in a given setting is within normal bounds or outside acceptable limits. Predictors would be the memory consumption, cpu usage, I/O read and write, and network usage. By classifying application runs on a number of different units and different scenarios it can be inferred as to what applications are acceptable.  Classification can also be used in email spam detection in a predictive manner. Predictors would be words contained in the email, potentially email addresses used, and the response would be spam/not spam.  b) Regression could be used to predict the overall performance of a high school or college football recruit at the next level. Some factors considered would be height, weight, speed, vertical, strength of competition, grades, SAT scores, and other factors ranging from scores of highlight tapes to socioeconomic indicators. The response would be a probability score range of a floor and ceiling of potential of the recruit. That is, how good they would be on a range. This would be a predictive model.  Another regression application could be time series data for economic forecasting. Predictors could be aggregate consumption, GDP, exchange rates, and employment rates. With each of these, a single item, say GDP can be taken as the response, and then it can be inferred how a change in exchange rates relates to GDP growth, or how employment is tied to Federal rate hikes.  Applications can also be in predicting hurricane or storm strength. Predictors being ocean temperature, currents, air pressure, and plenty of other atmospheric indicators. Regression can then predict where the storm may move to, that is a continuous range of possible paths that are refined as more data is taken in.  c) Cluster analysis can be used for image grouping. Using a Kohonen Map, with attributes being the actual pixels and their values, the images can then be grouped into visually similar clusters.  Another common use of clustering could be for exoplanets. This is a rapidly growing field, and there are constantly new planets being discovered, as such, using clustering helps observers draw similarities between exoplanets.  Clustering is also commonly used to determine the optimal number and location for a set of stores within an area. |
| 5 | For regression, a more flexible model will perform well with a larger data set, as it will have a better representation of the actual problem space. The converse is true for a less flexible model, that is, with sparser data, it will fit to the problem with less sensitivity. For classification, having a flexible model can be problematic with large data sets, that is if the model is too flexible it will be adversely affected by outliers in a much greater sense than a regression problem. |
| 6 | A parametric model has a fixed number of parameters that are the only variables being adjusted for the model, they are typically built from a starting assumption about the form of the data, whereas a non parametric model makes no assumptions and can learn any functional form of the data. Non parametric models are preferable with a large data set and no starting assumption and a parametric model should be preferred when the data is known, or assumed to be modeled best in a given form, ie linear equation. Parametric is more likely to overfit if the data is more complex than the model. |
| 7 | 1. Using d(x,y) = sqrt((x0-y0)^2+(x1-y01)^2+(x2-y2)^2)    1. 3    2. 2    3. sqrt(10)    4. sqrt(5)    5. sqrt(2)    6. sqrt(3) 2. Green, as point 5 is the closest 3. Red, as the three closest points are 5, 6, 2, the latter two being Red. 4. If the Bayes decision Boundary is highly nonlinear we would want to use a smaller value of K, as with smaller values of K, a more complex Bayes boundary is drawn by KNN, as the most influential points are the local K, whereas a high value of K, produces a closer to linear boundary. |