CS530—Fall 2021—Quiz 3

This is an open-book quiz. You can use any material that you want. But you must of course take this quiz without the help of anyone else. Please type your name below to indicate that you understand the instructions above and will abide by them.

Your name: Toby Chappell

Below we describe 6 datasets. For each dataset, explain the advantages and disadvantages of using the following algorithms on it. Then, if you had to use only one algorithm on each dataset, explain which you would use for each dataset. There is not necessarily always one best answer. So, your explanation sincerely matters.

The algorithms are (1) Logistic Regression, (2) K-Means, (3) Random Forest, (4) K-Nearest Neighbor, (5) Support Vector Machine (with your choice of kernel), (6) Principal Component Analysis, and (7) Linear Discriminate Analysis.

(Hint: In your answer, think of the different assumptions that each algorithm makes as well as any other characteristics of those algorithms.)

1. A hospital dataset contains 12,000 patients' medical data. The dataset has 213 features; each feature is related to the demographic and medical history of the patient. Of those, 152 are continuous (e.g., height, weight, sugar levels) and 61 are categorical features (e.g., blood type, birth gender). You also have the labels of the risk that each patient would suffer from a heart attack over the next year (1 to 5, 1 being the lowest), compiled from estimates of experienced physicians. You are asked to build a classifier to detect whether a patient is highly at risk (>=4 rating) of a heart attack over the next year.
2. Logistic Regression: One advantage to using Logistic Regression would be that you could actually determine the probability of whether a patient is high risk of having a heart attack over the next year. A disadvantage would be that you the categorical variables could not factor into your model (you could perhaps find a way to convert from a categorical to a continuous variable for some features but it would involve a lot more extra work on the dataset).
3. K-Means: An advantage to K-Means is that it attempts to find a structure in the data without any assumptions (other than the data is linked by a distance metric). The downside to using K-Means is that since even though we have the labels for the dataset, we would not be able to factor that into our model.
4. Random Forest: A huge advantage for random forest is that it would select random features and samples from the dataset for each tree, thus decorrelating the data. A disadvantage would be that random forest can perform slower depending on the parameters given.
5. K-Nearest Neighbors: K-NN does not make any assumptions about the dataset and does not require a model on a training set. However, k-NN is fairly slow and with a large dataset such as this it would be very time consuming to run.
6. SVM: SVM would work very similarly to logistic regression and better if the data was linearly separable or nonlinear. A disadvantage however would be that we would not have access to probabilities which would be useful for this situation.
7. PCA: PCA does not make any assumptions and would work very well for feature selection for a dataset with as many features as this. This would however make the resulting features far more difficult to interpret.
8. LDA: An advantage to LDA is that it would result in dimensionality reduction and be able to make the data more easily classifiable. However, the two classes of interest would be required to have the same variability which is not guaranteed.

Best algorithm: I believe random forest would work best here since you could make use of both categorical and continuous data. In addition, it would help determine the specific features that should be selected in determining the optimal model.

1. You are trying to determine the prices of houses in various cities within Los Angeles County. The features in your dataset are square feet, number of bedrooms, number of bathrooms, lot size, city name, zip code, previous selling price (the price the house fetched when last purchased), year bought (when it was last purchased), etc. The dataset you acquired contains this information about each house in Los Angeles County. And you are trying to develop an app that would predict at what price to list a house given its features.
2. Logistic Regression: Logistic Regression could work well for this dataset since the classes would be related to each other and you could even determine the probability. However, you could not make use of categorical variables which could be very informative in this case (location for instance is very indicative of price). You would need to also create a true label to predict on (perhaps previous price plus inflation) which would make the predictions slightly less reliable.
3. K-Means: K-Means would be useful since there are no true labels to the data. However, it may be tricky to work with since you would have to specify the number of classes you want in addition to the fact that K-Means will only show there is a distinction between classes rather than the price of that class.
4. Random Forest: Random Forest could do well here since it could take advantage of both categorical and continuous variables. In addition, Random Forest can be used for regression unlike most of the other algorithms. However, you would need to also create a true label to predict on (perhaps previous price plus inflation) which would make the predictions slightly less reliable.
5. K-Nearest Neighbors: K-NN cannot be used for regression.
6. SVM: SVM cannot be used for regression.
7. PCA: PCA could work here but it would make results harder to interpret and could not deal with categorical variables.
8. LDA: LDA cannot be used for regression.

Best algorithm: Random Forest would probably be best, with Logistic Regression as a close second, since it could take advantage of both categorical variables and continuous variables. Furthermore, Random Forest can use Regression unlike most of the other algorithms.

1. A dataset contains 2065 measurements of 4 different kinds of plants, including their height, the diameter of the stem, their age, and the length of the leaves. You need to classify future plant entries into those same 4 categories.
2. Logistic Regression: Logistic Regression does not perform very well with more than 2 classes, especially when classes do not exhibit a hierarchy such as in this case.
3. K-Means: An advantage again of K-Means is that it attempts to find a structure in the data without any assumptions (other than the data is linked by a distance metric). The downside would be that even though we have the labels for the dataset, we would not be able to factor that into our model.
4. Random Forest: An advantage for random forest is that it would select random features and samples from the dataset for each tree, thus decorrelating the data. A disadvantage would be that random forest can perform slower depending on the parameters given.
5. K-Nearest Neighbors: K-NN does not make any assumptions about the dataset and does not require a model on a training set. However, this can also be a disadvantage for a large dataset like this since you would have to run the algorithm over the entire dataset with every new entry.
6. SVM: SVM could do very well on this dataset since it has high generalizability. However, SVM does not extend very well to more than 2 classes.
7. PCA: PCA does not make any assumptions and would work very well for feature selection for a dataset with as many features as this. This would however make the resulting features far more difficult to interpret.
8. LDA: An advantage to LDA is that it would result in dimensionality reduction and be able to make the data more easily classifiable. However, the four classes of interest would be required to have the same variability which is not guaranteed.

Best algorithm: The best algorithm for this dataset would be LDA since it would lead to dimensionality reduction. In addition, the assumption that the classes would exhibit the same variabilities would have to be tested but if that assumption is met then it could lead to an optimal separation.

1. You give 10,000 people a questionnaire to fill out about their experiences during COVID. Each questionnaire contains 27 questions that the participants answer on a 7-point Likert scale (e.g., “To what extent are you occupied with thoughts regarding becoming infected with COVID-19? From 1 (not at all) to 7 (all the time)”, “How has your financial situation changed since March 2020? From 1 (changed markedly for the worse) to 7 (changed markedly for the better)”). A domain expert tells you that these data should generally subdivide into 3 groups in which samples are more similar to each other than different among groups. And you decide to test this.
2. Logistic Regression: Logistic Regression does not work since there are no labels
3. K-Means: An advantage to K-Means is that it attempts to find a structure in the data without any assumptions (other than the data is linked by a distance metric). Given the fact that we are totally unaware of what the true value our classes would even be, this would help us significantly.
4. Random Forest: Random Forest would not work since there are no labels.
5. K-Nearest Neighbors: K-NN does not work since there are no labels.
6. SVM: SVM would not work since there are no labels.
7. PCA: PCA would work in this case but just might not be necessary since there are only 27 columns.
8. LDA: LDA could not work since there are no labels.

Best algorithm: K-Means would 100% be the best algorithm in this case. It does not require labels for each entry and the goal of the algorithm is to find structure in the data. To test the hypothesis, you could run K-Means and see if the class distinctions are significant.

1. You have a dataset containing information about 1998 customers: demographic data with 101 features, which includes age, income, education, etc. The dataset also contains a label: whether the customers previously bought the product you are promoting. After performing PCA, the first 2 principal components give you 90% of the variance. You visualize the first 2 principal components in a 2D plot and find that the data points of customers who bought the product look roughly as though they reside on an ellipse and are generally well separated from the dataset reflecting the people who did not buy the product, which more often than not cluster inside the ellipse. You were just given a new set of data containing 3487 samples of demographic information of new customers. You want to build a classifier that would predict the probability that new customers will buy your product to better focus your advertising campaign.
2. Logistic Regression: Logistic Regression would work really well here since it could provide the probabilities easily.
3. K-Means: K-Means would not really be giving us what we want here since it doesn’t use our true labels and it wouldn’t give us probabilities.
4. Random Forest: Random Forest would not work that well since the data is easily separable.
5. K-Nearest Neighbors: K-NN would work well for classification but it would be difficult to get the probabilities we need.
6. SVM: SVM would work better than Logistic Regression since the data is easily separable, however, it could not get the probabilities we need.
7. PCA: PCA should be used since it is already known to make the data easily separable (need to assume the new dataset exhibits the same distribution but that assumption should most likely be met).
8. LDA: LDA would work well for classification but it would be difficult to get the probabilities we need.

Best algorithm: Logistic Regression would work the best since it is the only classifier to provide a probability for an event to occur. In addition, PCA should be used as a preprocessing step.

1. You have a set of 1745 cardiac patients, with 7 continuous features for each patient (age, weight, blood pressure, etc.). You also know which of them survived open-heart surgery. Your job is to predict which future patients will survive open-heart surgery. Using various visualizations, also using dimensionality reduction, you take a look through the feature space of the dataset, and you realize that the separation manifold between the two classes is likely to be highly nonlinear. Further, as far as you can tell, the classes are not that easily separable.
2. Logistic Regression: Logistic Regression would not work since it cannot classify nonlinear data.
3. K-Means: K-Means would not be useful here since labels are provided.
4. Random Forest: Random Forest would work very well here since it can classify nonlinear data. However, since there are only 7 features decorrelating the data is not as large of an advantage here compared to larger datasets.
5. K-Nearest Neighbors: K-NN could perform fairly well since it is able classify nonlinear data. However, it is difficult to determine the optimal k.
6. SVM: SVM would work very well here since you could use a kernel in order to classify the highly nonlinear data. The disadvantage would be that the optimal kernel would need to be determined.
7. PCA: PCA would not really be necessary since there are only 7 features.
8. LDA: LDA would not work here since the data is nonlinear.

Best algorithm: SVM would most likely perform best here since they work well on highly nonlinear data using a kernel (probably RBF). This is because SVM is able to project data into different dimensions in which the data becomes linearly separable (or nearly linearly separable).