Question 1 - Visualization and Analysis of Penguin Dataset

The Palmer penguin dataset consists of 344 records of physical attributes of three species of penguin living on three islands in Antarctica (Table 1) [ref]. In this report, consideration is given to data cleaning, including missing values, correcting data imbalance and standardization. The data are explored through visualization and then analysis is carried out to compare the performance of a small number of AI approaches. In addition to a simple baseline method, the analysis consists of k-Nearest Neighbour (*k*nn) and random forest classification, *k*-means unsupervised clustering and a novel and interesting classification approach that uses insights gained from visualizations to define a classifier that is formed of a small number of separate two-dimensional support vector machine (SVM) classifiers.

Table 1. The attributes of the Palmer penguin dataset

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Type** | **Values in the dataset** |
| species | categorial | Adelie, Chinstrap, Gentoo |
| island | categorial | Torgersen, Biscoe, Dream |
| bill length | numerical | 32.1mm - 59.6mm |
| bill depth | numerical | 13.1mm - 21.5mm |
| flipper length | numerical | 172mm - 231mm |
| body mass | numerical | 2700g - 6300g |
| sex | categorial | Male, Female |

**Data cleaning - missing values, data imbalance and standardization**

In the dataset, 11 records have missing values. Two of these records are missing values for all of the numerical attributes and as well as a value for the *sex* feature. As the physical features of the male and female of each species differ considerably, the missing sex value made substituting the missing numerical attributes with statistically determined values cannot be justified. Consequently, these two records were deleted.

The remaining nine records with missing values all have no value for the *sex* attribute only. As can be seen in Table 2, the physical attributes of the male and female of each species are different and hence it is reasonable to consider assigning a *sex* to those records missing this attribute. The process followed was that each of the four numerical attributes was independently standardized (to zero mean and unity standard deviation). It was then confirmed using a Shapiro-Wilk test that each numerical attribute exhibits a normal distribution [ref]. This confirmation is needed for a robust application of a Z-test, and this was performed with respect to the relevant species population to assess separately both the hypothesis that the missing sex value is Male and that it is Female [ref]. It was found that two of the records could be imputed as Male and three as Female at the 95% confidence level and these were retained in the dataset with an imputed sex attribute value. The remaining four records failed under both hypothesis tests and these were removed from the dataset. The cleaned dataset was used for the subsequent visualization and data analysis presented here and consists of 338 records made up of 147 Adelie penguins (74 male, 73 female), 68 Chinstrap penguins (34 male, 34 female) and 123 Gentoo penguins (62 male, 61 female).

**Table 2.** Mean values for the numerical attributes for each sex of the species in the dataset, with means following standardization of the dataset shown in parentheses

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Species** | **Sex** | **bill length** | **bill depth** | **flipper length** | **body mass** |
| Adelie | Female | 37.3mm (-1.22) | 17.6mm (0.24) | 187.8mm (-0.93) | 3368g (-1.04) |
| Adelie | Male | 40.4mm (-0.64) | 19.1mm (0.97) | 192.4mm (-0.61) | 4043g (-0.20) |
| Chinstrap | Female | 46.6mm (0.48) | 17.6mm (0.22) | 191.7mm (-0.65) | 3527g (-0.84) |
| Chinstrap | Male | 51.1mm (1.31) | 19.3mm (1.07) | 199.9mm (-0.07) | 3939g (-0.33) |
| Gentoo | Female | 45.6mm (0.30) | 14.2mm (-1.48) | 212.7mm (0.84) | 4680g (0.60) |
| Gentoo | Male | 49.5mm (1.01) | 15.7mm (-0.73) | 221.5mm (1.47) | 5485g (1.60) |

**Better to show stdev?**

**Imbalanced data**

If a dataset is imbalanced, AI approaches may be biased in predicting classes that are more commonly found in the training data. The imbalance can be reduced either by deleting records of classes that are overrepresented in the dataset, or using ag statistical approaches to generate additional records of classes that are underrepresented. The Palmer penguin dataset is somewhat imbalanced, with the number of Chinstrap records being around half of that of either Adelie or Gentoo, which are present in similar numbers. The importance of imbalance depends on the analysis method applied and it is known that all the methods adopted in the current work are generally little affected by imbalanced data. The facts that the dataset is not greatly imbalanced and that the methods adopted are known to be little affected by imbalanced data [refs], no modifications were made to reduce imbalance.

unbalanced

----------

knn robust

logistic regression not

random forest robust

svm can be affected

k-means inbalance not relevant

**Visualization of the dataset**

|  |
| --- |
|  |
| **Figure 1**. The samples were taken from three islands; this attribute may be a cofounding factor |

Figure 1 shows the distribution of species between the three islands in the study. As Chinstrap and Gentoo penguins are found only on one island, there could be a concern that, when attempting to identifying species from their physical characteristics, these characteristics may be affected by an island’s environmental factors (such as predators, food supply or climate differences) and not just the inherent characteristics of the species. To assess whether the island inhabited is such a cofounding factor, the physical characteristics of the Adelie penguins can be compared, since they are found on all the islands in the dataset. A Shapiro-Wilk test was used to confirm that all of the Adelie numerical features are normally distributed and an ANOVA test confirmed that the physical characteristics of the Adelie penguins were not significantly affected by the island inhabited, meaning it is not likely to be a confounding factor in the dataset.

Pairwise scatterplots for the numerical features are shown in Figure 2. It can be seen that using *bill depth* in combination with either *flipper length* or *body mass* provides a separate cluster of Gentoo penguins (shown in green) allowing them to be identified. There is no pairwise combination of numerical features that completely separates the Adelie (orange) and Chinstrap (purple) clusters, but it can be seen that the distributions involving *bill length* provide good separation making this a good candidate for distinguishing between these species.

Figure 3 shows there is a difference in the body mass of the male and female samples for each of the three species. Differences between the sexes for the other three physical characteristics in the dataset were also apparent. Since narrower distributions are found if the sex of the species is considered rather than just the species itself, including sex will provide a finer grained distinction for species classification and this is knowledge can be used to improve performance of the analysis approaches as discussed in the analysis section.

|  |  |
| --- | --- |
|  |  |
| **Figure 3**. All numerical features show a significant statistical difference between the male and female measurements, exemplified in the body mass boxplot above. The plots show median values and Q1 and Q3 quartiles. Outliers are indicated if they lie outside the range Q1-1.5IQR to Q3+1.5IQR, where IQR=Q3-Q1 |
| **Figure 2**. Pairwise distributions of numerical features. Gentoo can be distinguished from the other species, but Adelie and Chinstrap may not be separable |

Talk about slope of line in ‘surprising plot’, this is almost vertical and may become more so as the number of samples is increased.

**Analysis**

Predicting the penguin species from the given features is a classification problem. A baseline classification method is first implemented, providing a reference for the performance of other methods. This report considers two supervised approaches, namely *k*nn and random forest. An unsupervised *k*-means approach is also taken, but as the clusters found are unlabelled, they need to be related to the species so that the classifying accuracy can be determined. An interesting and usual approach is also described that uses a combination of the insights found from visualizations. This led to the identification of a short sequence of two-dimensional linear classifications using Support Vector Machines (SVMs) that were able to perform at least as well as the other classification methods. Regression approaches were not considered as although categorical values could be assigned numerical values, performance is likely to poor unless they have recognizable ordinal counterparts, which is not the case for this dataset.

For all of the approaches considered, the dataset is divided into a combined training and validation set (80% of the dataset) and a test set (20% of the dataset). *A number of pseudo random sets were used?* During training, a ‘rotating set’ of validation data (20% of the whole dataset) is extracted from the combined set and used to tune metaparameters to maximize performance and to reduce overfitting. Once suitable metaparameters have been determined, the classification results are obtained using the test set. All the methods required metaparameter tuning and we took a grid approach in which a small set of values for each metaparameter is selected from a suitable range of reasonable values [ref]. *This can allow the identification of suitable values and perhaps a second stage in which a smaller range is concentrated upon.*

Not just a single test set….

The relative performances of the approaches are assessed using accuracy, measured as a percentage of correct predictions in the test set *and confusion matrices, so that the number of misidentifications for specific species can be seen and this may be useful in tuning models?* Methods that used precision either directly or as part of the metric (such as Recall and F1-Score) were not considered as these are generally more useful if the cost of false negatives is high, which was not found in the current work.

Mean of random sets 1 to 10, from best result of each of the grid

Talk about implementation – Python, loops to use different random training sets etc…. grid method

Perhaps a section on this…

***Baseline method***

A baseline is the simple prediction used to assess performance improvements that can be achieved by the main algorithms being considered. If the performance cannot be improved significantly above that of the baseline, then this may indicate that the approach being considered is not suitable or that the problem is particularly intractable. It also gives a basis for the comparison of principal methods being considered. In a classification problem, the baseline used is often to select the class that has the most observations, in regression the mean or the median and in unsupervised learning perhaps a random assignment of data values to clusters is made. In this work the Adelie penguins are the most common (147/388), giving an accuracy of 37.9%. Other methods in this study will be compared with this baseline figure.

***Classification method 1 - k-Nearest Neighbour (kNN)***

Training for *k*NN algorithm simply involves storing the training dataset. Classification then requires finding the distances (typically Euclidean or Manhatten?) from the test data value to the training examples. The *k* nearest neighbours to the test data value are found and the most common class label among the *k* neighbours is the predicted label.

*k*NN is known to be biased in favour of features with smaller standard deviations[ref], so the data were standardized before analysis. The metaparameters considered during training and selected for testing are shown in Table 3.

Comment?

**Table 3.** Metaparameters used in training the methods. The values shown in bold are those that most consistently produced the best accuracy results during validation and so selected for testing.

|  |  |  |
| --- | --- | --- |
| **Method** | **Parameter** | **Values considered** |
| *k*nn | number of nearest neighbours *k* | 3, 4, 5, 6, 7, 9 |
| weight function for prediction | uniform, distance |
| Minkowski distance power parameter | 1, 2 |
| random forest | number of trees in the forest | 50, 100, 200 |
| maximum depth of the trees | no maximum, 10, 20 |
| minimum number of samples to split node | 2, 5, 10 |
| minimum number of samples at leaf node | 1, 2, 4 |
| *k*-means | number of clusters *k* | 2, 10 |
| centroid initialization method | k-means++, random |
| number of runs with different centroid seeds | 5, 10, 20 |
| maximum number of iterations | 10, 20, 50 |

*200 too large for number of trees?????*

*Do these cover all the parameters?*

*Need to match my text – gini , entropy random forest?*

*k-means 0 Euclid*

The results obtained for the classification are shown in Table 4. Comments?

**Table 4.** Classification results - obtained from 10 random sets of training/valid and following training and using the parameters identified in Table 3

|  |  |
| --- | --- |
| **Method** | **Accuracy** |
| baseline, most numerous species | 37.9% |
| kNN, all features | 82.35% |
| kNN, no island, no body mass | 97.65% |
| random forest, all features | 98.97% |
| random forest, no island, no body mass | 98.97% |
| k-means, all features | 97.06% |
| k-means, no body mass\* | 97.06% |
| Surprising, no island, no body mass | 98.00% |
| Surprising, no body mass | 99.32% |

***Classification method 2 - Random forest***

A random forest is a collection of decision trees; each tree being trained separately on a subset of the training data using a subset of features. Decision tree training builds a set of nodes each defining a test on selected feature value to determine which output branch to follow. The CART (classification and regression tree) algorithm is often used to determine which feature to use at the node and one of a range of possible methods (such as information gain or Gini impurity) is used to assess the quality of the split.

The performance of the random forest is known to be largely unaffected by the standard deviation of the data and so no standardization was performed. The metaparameters shown in Table 3 were tuned using a *rotating???* training and validation set (describe params) and the results are shown in Table 4 using the best metaparameters.

Comments?

***Unsupervised method – k-means***

*k*-means is an unsupervised clustering method, but it can be used for classification by aligning identified clusters with classes. In the first cycle, *k*-means randomly selects *k* data points as cluster centroids in the training data space and assigns the remaining data points to the cluster whose centroid is closest (using Euclidian distance or another method). The means of the data points in each cluster are calculated and then used as the centroids in the next cycle. The cycle repeats until the clusters no longer change or a defined maximum number of iterations is reached.

For *k*-means clustering, standardization or normalization of features is often recommended. This is because *k*-means clustering relies on calculating distances between data points to assign them to clusters. *k*-means operates with numerical values not categorical data. So, we only look at the numerical values.

The performance of the classification could potentially be improved by training a set of clusters for male penguins and another for female penguins. The *sex* feature can be used to determine which set of clusters to use for classification.

Using the approach described above, the metaparameters shown in Table 3 were tuned using a *rotating???* training and validation set (describe params). The best metaparameters were then used in the training. Parameters other than *k*??

*K – the elbow & silhouette methods were implemented, but we realize a value of* k needs to at least equal to the number of classes! In fact – what did results show?

|  |  |
| --- | --- |
|  |  |
| **(a)** in estimating *k*, the elbow method uses the change in slope of the ‘inertia’ (here *k*=3) and the silhoutte method uses the score closest to 1 (here *k*=2) | **(b)** the clusters can be mapped to species according to majority voting and the assignment to classes are shown by polygon colours (here for *k*=10) |
| **Figure 4**. Two methods were considered for determining the value of *k* for the *k* means clustering and the assignment of clusters to species can be seen in lower dimensional visualizations | |

***Something surprising and unusual - a novel combined visualization and analysis (CVA) approach***

In the novel CVA approach, visualization of pairwise combinations of the numerical data, combined with a short sequence of simple two-dimensional linear classifiers based on SVMs was found to be able to produce results of accuracy at least as good as the approaches investigated in the previous section.

The CVA approach requires greater manual effort in gaining a deeper understanding of the nature of the dataset. This is in contrast to ‘black box’ classification approaches that are often applied with little underlying insight into the nature of the data or knowledge of the method adopted. The drawback of the approach taken is that it is not applicable generally as it may not always be feasible or possible to extract the necessary insights from visualizations. Also, it will become more difficult to apply the CVA approach as the number of features is increased.

|  |  |
| --- | --- |
|  |  |
| **(a)** *bill depth* and *flipper length* allow the Gentoo species to be distingushed from the remaining species | **(b)** *bill length* and *bill depth* allow the Adelie and Chinstrap species to be distingushed |
| **Figure 5**. The samples were taken from three islands; this attribute may be a cofounding factor | |

Perhaps could be continued with further features????

Why does removing a feature improve performance?

REMOVING features can improve performance…

random forest, no island,98.82%

Confusion matrix to show that confusion is between Adelie and ?.

\*categorical data not included

**Conclusions**

Nice things about my method. Careful to carry out AI in such a way that it is robust. Good idea generally, not just because of my method.

The SVM approach is shown to produce accuracy results at least as good as the other analysis methods applied in this work, although clearly the approach is specific to this particular application.

The SVM approach used here is not using the data in an ‘n-dimensional’ way, just using a number of segments of fewer dimensions (typically 3 or 4). While this approach has found success here, it is unlikely to be practical for datasets having larger numbers of dimensions. The manual effort required is also likely to be prohibitive in many cases.

You should consider how to visualize the data and which algorithms to

try. Nothing you do will be completely successful, this coursework is

not here to judge your final accuracy but the care you bring to your

investigation. Here are some thing you should consider:

\begin{itemize}

\item The kind of algorithm to use, for example whether to classify, regress or cluster.

\item The metric to use to measure the performance of the model.

\item What sort of baseline to compare the model to.

\item How to choose the hyperparameters of your model.

\end{itemize}

For good marks you should include some graphs that illustrate

properties of the data and you should compare two classification

algorithms, both to each other and to a baseline model. The algorithms

you pick do not need to be unusual, for example $k$nn classification

would be perfectly good, though, of course, for full marks this would

include some consideration of how to pick $k$ and how to measure the

distance, though, as you know, no approach to chosing $k$ is every

going to be completely satisfactory. In addition, you should include

either some exploratory regression or unsupervised learning; for

regression you might regress two properties and examine whether the

regression parameters are the same for each penguin type; unsupervised

learning could use $k$-means, for example. You do not need to do both

regression and unsupervised learning.

You should make sure any assessment is not restricted to the data used

in train models or decide on metaparameters. In your report you should

explain your decisions. You code will not be marked for elegance, but

it should run correctly; it is expected you will use Python, but any

of Python, Julia or R is fine. Do not include screenshots of graphs,

they should be imported directly; resize them to the correct size

before importing them, if the labels are tiny the graphs will not be

marked. Make sure figure captions are descriptive, it is better to

have some overlap between figure captions and the main text than to

have figure captions that are not reasonably self-contained.

As a rough guide to marking:

\begin{itemize}

\item Initial description of the data, including some graphs or other approaches to visualisation. 6 marks.

\item Either unsupervised learning or regression. 6 marks.

\item Two algorithms should be tested, if only one algorithm is

included the 28 available marks will be halved.

\item Overall presentation (3 marks), including use of appropriate

sections, plots, diagrams, or tables to make your point. Do not

include code snippets in the report. Instead, describe in words or

equations what you are implementing. Format equations correctly.

\item Suitable choice of algorithms (4 marks).

\item Suitable choice of evaluation for algorithms (3 marks).

\item Comparison with a suitable baseline (3 marks) and a justification for which baseline to use.

\item A description of metaparameter selection (3 marks), if one

algorithm has not metaparameter, then explain that and note why not

and why this do or does not make it a better algorithm for these

data.

\item Describe and compare the results from your two algorithms,

include a description of how you implemented the algorithms. (6 marks)

\item There are some marks (6 marks) for something suprising and unusual.

\end{itemize}

\section\*{Question 2 - Ethical challenge facing us in data science and AI}

For two of these three types of ethical challenge facing us in data science and AI:

\begin{enumerate}

\item The protection of data, of the people whose data they are and participants in any study.

\item Avoiding the amplification of biases and regressive values implicit in historic dataset.

\item The safety of AI systems and the possible of existential threats from machines.

\end{enumerate}

describe what you think is a specific example of a challenge that

could arise or has arisen in the past. Obviously the three broad types

of challenge overlap, do not worry about the boundaries between these

types, but do try to address different types of threat in your

examples. Explain how the ethical problems could be addressed, or at

least made more transparent.

\subsection\*{Report}

Your report should be no longer than five pages, including any

references. It is expected that Question 2 would occupy about a fifth

of this space; use an 11 or 12pt font and do not try tricks like

expanding the margin to fit in more text, shorter is better than

longer.

Your report must be submitted in pdf and should be prepared in LaTeX;

overleaf is a good approach, but not required as long as LaTeX has

been used. As always when using LaTeX, give yourself over to defaults,

our expectation of what a document should look like has been

conditioned on LaTeX, so it is best not to try to override the look of

the document.

Avoid code snippets in the report unless that feels like the best way

to illustrate some subtle aspect of an algorithm; do always though

consider a mathematical description if possible. You will be asked to

submit code and it may be tested to make sure it works and matches

your report. It will not, however, be marked in and of itself.

\subsection\*{knn}

Perhaps use F1-score (there are others!) as the classes are imbalanced in number?

F1-score is a metric that considers both precision and recall. Precision measures the accuracy of positive predictions (TP/(TP+FP)), while recall (also known as sensitivity) measures the fraction of positives that were correctly identified (TP/(TP+FN))

F1-score is the harmonic mean of precision and recall and is calculated as follows:

F1 = 2x(PrecisionxRecall)/(Precision+Recall)

F1-score ranges from 0 to 1, where a higher value indicates better model performance. F1-score is particularly useful when classes are imbalanced because it considers both false positives and false negatives.

\section\*{Report template}

This is a report template, you don't need to use this template, but do

use it if it is helpful.

Here is an example of an equation:

\begin{equation}

\pi=4\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}\ldots\right)

\end{equation}

or

\begin{equation}

\pi=4\sum\_{n=0}^\infty\frac{(-1)^{n}}{2n+1}

\end{equation}

where $\pi$ can be written in line by using \$'s. Here is a vector:

\begin{equation}

\mathbf{x}=\left(\begin{array}{c}x\_1\\x\_2\end{array}\right)

\end{equation}

<https://machinelearningmastery.com/how-to-get-baseline-results-and-why-they-matter/>