Question 1 - Visualization and Analysis of Penguin Dataset

The Palmer penguin dataset consists of 344 records of the physical features of three species of penguin living on three islands in the Antarctic (Table 1) [ref]. In this report, consideration is given to data cleaning, including missing values, correcting imbalance and standardization; the data are explored through visualization and then analysis is carried out to compare the performance of different AI approaches. In addition to a simple baseline method, the analysis consists of two classification methods (k-Nearest Neighbour and random forest), a clustering method (k-means) and an additional 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 linear classifiers. The 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.

Table 1. The features of the Palmer penguin dataset

|  |  |  |
| --- | --- | --- |
| **Feature** | **Type** | **Values** |
| 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 - consideration of missing values, imbalanced data and standardization**

There are 11 records in the dataset that have missing values. Two records have none of the numerical features and are missing a value for the *sex* feature. Substituting the missing numerical values with a statistically determined value was considered, but as the physical features of the male and female of each species differ considerably, this approach is unlikely to provide an updated record that is consistent with the original dataset. Consequently, these two records were deleted.

The remaining nine records with missing values all have no value for *sex*. As shown in Table 2, the physical attributes of the male and female of each species are different and so it may be possible to use this difference to assign a sex to the records with missing values. The approach taken was, for those records with missing values, standardize separately each of the numerical attributes (so they each have zero mean and unity standard deviation). For each record with a missing value for *sex*, a two-sample t-test was performed for each of the four numerical features to assess separately the hypothesis that the missing *sex* value is *Male* and that the missing *sex* value is *Female*, both at the 99% confidence level. It was found that one of the records could be imputed as *Male* and five as *Female* and these were retained in the dataset with the modified sex attribute. The remaining four records that failed under both hypothesis tests were removed from the dataset. The cleaned dataset had 339 records (148 Adelie, 68 Chinstrap and 123 Gentoo samples).

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

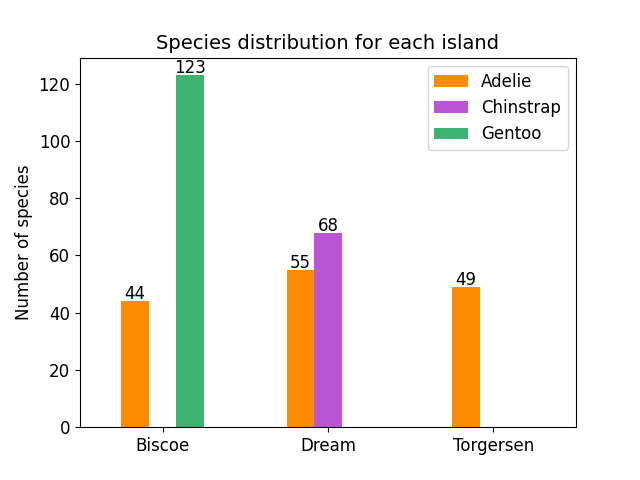
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **species** | **Sex** | **bill length** | **bill depth** | **flipper length** | **body mass** |
| Adelie | Female | 37.3mm (-1.22) | 17.6mm (0.24) | 187.8mm (-0.93) | 3368g (-1.04) |
| Adelie | Male | 40.4mm (-0.64) | 19.1mm (0.97) | 192.4mm (-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) |

**Imbalanced data**

If a dataset is imbalanced, it can be necessary to either delete samples from the classes that are overrepresented, or use statistical approaches to generate. Has the effect of the AI method not working well as some classes are overrepresented. The Palmer penguin dataset is somewhat imbalanced, with the number of Chinstrap samples being around half of that of either Adelie or Gentoo, with these two species having a similar number of samples. The importance of imbalance depends on the analysis method applied and it is known that the methods adopted in the current work, namely kNN, random forest and k-means are generally little affected by imbalanced data. The facts that the dataset is only marginally imbalanced and that the methods adopted are known to be little affected by imbalanced data, no modifications were made to reduce imbalance in this work.

**Investigation of the island as being a confounding 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. In this dataset, the Adelie penguins are present on all three islands and so any effect on physical characteristics that is peculiar to one island would be likely to affect this species. Following a Shapiro-Wilk test to confirm that all the statistical distributions of all the Adelie numerical features are normal, an ANOVA test was carried out, and this confirmed there was no statistically significant difference of the physical characteristics of the Adelie numerical features between the islands. Consequently, the island inhabited by the penguins was unlikely to be a confounding factor affecting the correct identification of the penguin species.

****

**Figure 1**. Species distribution between islands

Outliers?

**Visualization of the dataset**

**Analysis – Metaparameters, metrics**

The problem is to determine the species is principally one of classification. This report considers two such supervised approaches, namely knn and random forest, but an unsupervised approach is also taken, but the clusters found 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 that led to the identification of a short sequence of two-dimensional linear classifications (using SVM). Regression approaches were not considered as although categorical values could be assigned numerical values, performance is likely to best if they have recognizable ordinal counterparts and this is not the case here.

Accuracy & confusion matrices

The majority of learning methods involve decisions to be made about the values of metaparameters in the. We took a grid approach in which a small set of values is selected across a wide range of possibles. This can allow the identification of suitable values and perhaps a second stage in which a smaller range is concentrated upon.

The dataset is divided into a training set, validation set and a test set. During training, validation sets are used (see grid\_search) to tune metaparameters and reduce overfitting during training. Once suitable multiparameters have been determined, the results are obtained using the test set

Scaling – standradization needed ?

scaling – knn yes

kmeans (somehow worse) - 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.

svm?

Random forest – no need to scale for decision trees

What about for my method?

The main metric that will be used is accuracy, namely what percentage of the penguins’ species are correctly predicted by the model. This supported by confusion matric, 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 is not specifically relevant to the current work.

**Baseline?**

For the baseline could use a simple majority - could be sufficient?

Use the island? Probs get about 65%???

**Analysis**

Three well-known analysis methods were implemented, plus a novel method described in the next section

**Something surprising and unusual - a combined visualization and analysis approach**

In this 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.

While this approach may take more effort, in that deeper understanding of the nature of the dataset needs to be obtained, the approach is not a ‘black box’ (classification approaches are frequently treated in this way with little underlying knowledge of either the data or the methods being adopted. The drawback of the approach taken is that it is not applicable generally as it may not always be feasible to extract the necessary insights from just pairs of combinations and the nature of the data may not be revealed without resorting the multi-dimensional approaches of the better know classification methods considered in the previous section. Also, it will become more difficult to use this approach as the number of features is increased.

Perhaps could be continued with further features????

**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.

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}