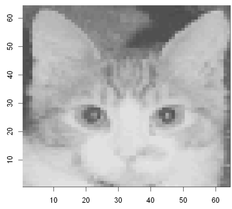
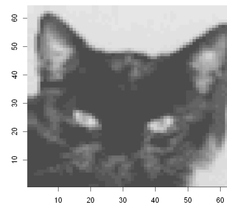
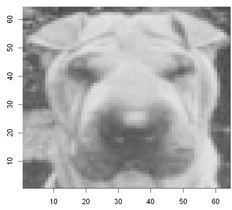
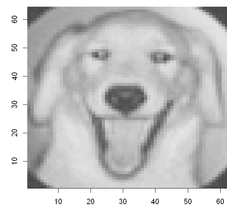
Project 3 - part of the exam from June 2022

Data set

For this exam you will use the Cats and Dogs data set: 99 images of Cats and 99 images of Dogs. Each image is 64 by 64 pixels so it's pretty low resolution. The data set is of dimension 198 by 4096 (64 times 64), i.e. raster scans of the images. Label 0 denotes a cat and Label 1 a dog. The csv files containing the images and labels are on Canvas. Here are 4 examples:

In [29]:
```
CATSnDOGS <- as.matrix(read.csv("CATSnDOGS.csv"))
Labels <- as.matrix(read.csv("Labels.csv"))
#
rotateM <- function(x) t(apply(x, 2, rev)) # the images are raster scans. Here, I just resort them fo
                                           # default image command in R to plot them with the right o
#
library(repr)
options(repr.plot.width=12, repr.plot.height=6)
#
set.seed(1000012)
ssc<-sample(seq(1,198)[Labels==0],2,replace=F)
ssd<-sample(seq(1,198)[Labels==1],2,replace=F)
par(mfrow=c(1,2))
image(seq(1,64),seq(1,64),rotateM(matrix(CATSnDOGS[ssc[1],],64,64)),col=gray.colors(256),xlab="",ylab
image(seq(1,64),seq(1,64),rotateM(matrix(CATSnDOGS[ssc[2],],64,64)),col=gray.colors(256),xlab="",ylab
image(seq(1,64),seq(1,64),rotateM(matrix(CATSnDOGS[ssd[1],],64,64)),col=gray.colors(256),xlab="",ylab
image(seq(1,64),seq(1,64),rotateM(matrix(CATSnDOGS[ssd[2],],64,64)),col=gray.colors(256),xlab="",ylab
```

For the exam questions, we will treat pixel values as features, ignoring the fact that these are images. What I mean, you are not expected to use image analysis tools etc. You can use the spatial (co)localization of the pixels for interpretation of results or to guide your methods. Visualizing e.g. feature importance or selection or mislabeled observations etc can be done using image representations for example.

IMPORTANT

Wherever the exam task contains a question ("Does X differ from Y?") it is *not sufficient* to answer "Yes" or "No" - you need to explain how you arrived to this conclusion, which results provide you with these answers, as well as a statement trying explain *why* the results are what they are, preferable discussing the latter in the context of this data set (e.g. pixels, part of images, image characteristics). Likewise, when a task is phrased as "Can you do X to get Y?" I, of course, mean for you to attempt to answer this question by doing something, not just hypothesizing about it.

The analysis tasks in each question are partly sequential - use your findings from earlier tasks to guide you with the latter ones. That also means you can write code that can be re-used for each subtask so plan ahead a bit for each question.

To answer the questions it is for the most part inadequate to perform the task once on the full data set. Make use of re-sampling techniques to support your findings.

Question 1 (60p) - subset

1a (10p)

Are the cats and dogs well separated, i.e. can you obtain good classification test accuracy performance on this data set? *Compare at least 5 classifiers of different character.*

Are there any images that are consistently mislabeled by the classifiers (use resampling to ascertain)? Why do you think these are difficult images to classify? Do the classifiers struggle with the same observations?

Are the errors balanced or is one class more difficult to classify correctly?

1b (10p)

Identify the most important pixels for classification - are these easy to identify or is there uncertainty in the selection of important features? *Compare at least 3 methods for selecting/identifying features. Note, does not need to be the same methods as in 1a.* Can you think of a way to improve on the stability of selection/identification? (Of course, here you should explore methods other than feature importance in RF and GBM since you already did that for project 2).

Can you explain why these pixels are selected/are deemed important?

Do the identified predictive features differ between the classification methods?

1c (10p)

Cluster the data set. Do the clusters agree with the class labels? Does changing the number of clusters have an impact on the overlap with the class labels?

What characteristics of the data does the clustering pick up?

Please think carefully about how you go about this question - how do you choose your input (e.g. do you perform preprocessing or not) to the clustering algorithm? what method and settings? Is it as clear cut how to make these choices if you *didn't know the labels*?

THEMES

Odd group numbers do theme 1 and even groups theme 2.

Theme 1:

Question 2 (60p) - subset

We will continue to work with the Cats and Dogs data set but now use simulations to investigate how feature selection and classification can be affected by various kinds of noise contamination. Here the focus is on cats vs dog prediction.

Since the tasks below are simulation based you will need to repeat the task for multiple runs and also compare to the noise free results for reference.

2b (10p)

The images are 64 by 64. You can view the images as built from 16 pixel blocks of size 16 by 16 - corner blocks, a next to corner blocks or interior blocks. Your task is to use a statistical method to choose *one* of these 16 by 16 blocks (the same block for all images) as input into the classification methods (You can do this manually or through group-lasso).

Is there a block of that size that results in good classification performance?

2c (10p)

Turn half of the images upside down.

What's the classification performance now (compared to noise free, uncontaminated data)? Which features are important now? Are upside down images over-represented among the misclassified observations.

2e (10p)

Finally, we turn to the pixels. The 64 times 64 (4096) pixels vary across images of cats and dogs. Explore the data set from the pixel perspective - that is, treat the pixels as 4096 observations to be clustered in 198-dimensional space.

Are there pixel clusters? What do they represent? How many clusters? Do these differ for cat and dog images? Is your answer to this question sensitive to the choice of metric and/or method?

Theme 2: simulation studies

By now, we have introduced wide variety of classification methods, from kNN and DA methods, to logistic regression and lasso regression, to kernel based methods like KRR and SVMS, ensemble methods like RF and GBM and finally neural networks. In project 2 you explore RF and GBM side-by-side. For this theme, explore all the above mentioned methods, and any other you may want to add, side-by-side in a simulation study.

Your task is to come up with a setting where the methods will have advantages and disadvantages over each other based on sample size. For small sample sizes, even if the class separation is complex, simple methods may still win out, and vice versa.

You can set up the simulation in any way you want.

Please run the simulation multiple times across at least 3 sample sizes and discuss your findings.