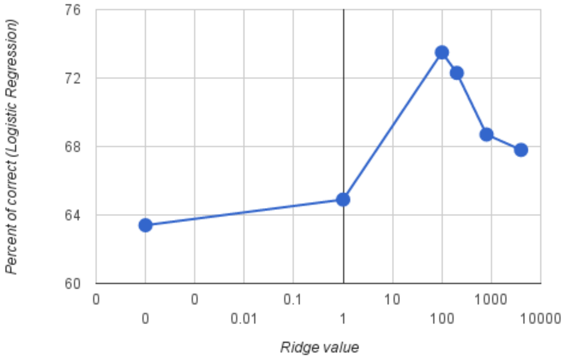
Fig. 1. Ridge & Percent of correct with Logistic Regression



1 Exploration of the dataset

1.a

Accuracy of classifier:

- SimpleLogistic: 64
- Logistic: 66.8

The difference between `SimpleLogistic` and `Logistic` are **XXX**

Using `InfoGainAttributeEval`, **XXX - fill in the result**. The reason for different performance in those 2 classifiers are **XXX**

1.b

The role `ridge` parameter are **XXX**

Compare regularization to feature selection **XXX**

Interpret the result **XXX**

A graph can be seen in Fig. 1. *X-axis* is drawn in log-scale in order that any trend is fully visible.

1.c

XXX Explore the effect of `gamma`. See Fig. 1

XXX Explore the effect of `complexity parameter`. See Fig. 1

1.c

This procedure does not guarantee to find the values of `gamma` and `c` that lead to the highest percentage correct (PC). Since **XXX**

Fig. 2. Gamma & Percent of correct with SVM Classifier

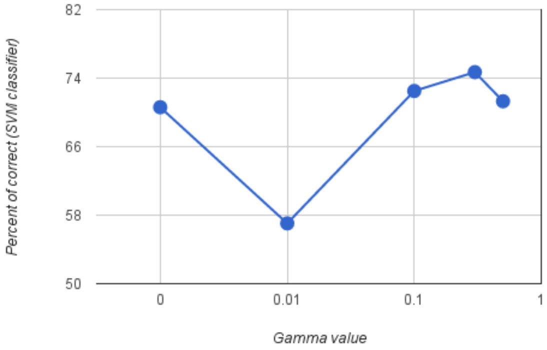
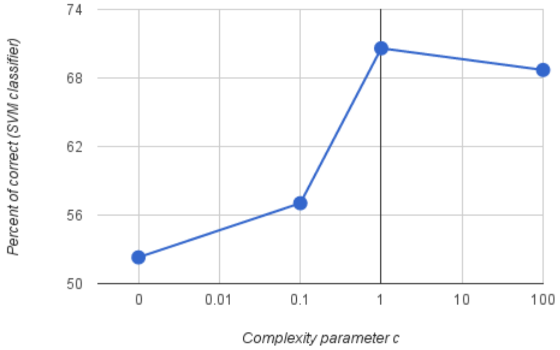


Fig. 3. Gamma & Percent of correct with SVM Classifier



1.d

Look at the list of the best 50 features, there are 3 *class indicator* variables in that list. The class indicator variable `is_bird` is ranked quite high in the list, at position 5. `is_cat` and `is_aeroplane` follows with position 18, and 22 respectively. **OOO**

The `SimpleLogistic` was trained on `train_images_partA`, there are 2 versions: (i) dataset with `imaId` removed, (ii) dataset with `imgId` and all the class indicator variables (except `is_person`) are removed. Then the classifier are tested on the validation set with the appropriate attributes removed. PC result:

- Remove `imgId`, keep all *class indicator variables*: 76.46
- Remove `imgId`, remove all *class indicator variables* but `is_person`: 69

XXX Relate the result to the observed feature ranking.

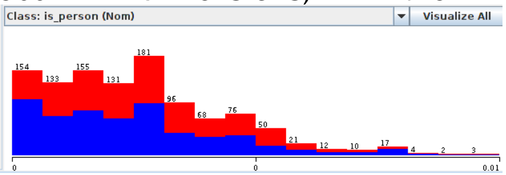
It **would/would not XXX** be easy to make use of the results in practice. And the reasons are **XXX**

2 Mini Challenge

Dataset exploration

Validation set

- Look at validation set, the only class indicator variable is `is_person`.
- Looking at the histogram of 500 dimensions, the most



Preprocessing data for training set