Practical Excercise A

In this exercicse you will implement a classifier, apply it to a simple dataset, evaluate the results using several metrics, and (hopefully) improve the performance using a non-linear feature expansion. Where code is asked for you need only include the central commands (complete listings are unnecessary). You must also give an interpretation of what the numerical values and plots you provide mean. Why are the results the way they are? Emphasis should be on the understanding of the technical content and discussion of the results.

1 Exercises

a) Consider the Logistic Classification model (aka Logistic Regression, see below). Derive the gradients of the log-likelihood of the parameters, $\frac{\partial}{\partial \beta}\mathcal{L}(\beta)$ where $\mathcal{L}(\beta) = \log \mathsf{P}(y|X, \beta)$.

> Logistic Classification
>
> To assist you with the derivation in part (a), we briefly recapitulate Logistic Classification. In this model each datapoint's class label is assumed to be generated from a Bernoulli distribution in which the probability of a positive class label is given by a logistic function applied to weighted inputs,
>
> $$\mathsf{P}(y^{(n)} = 1|\tilde{x}^{(n)}) = \frac{1}{1 + \exp(-\beta^\intercal \tilde{x}^{(n)})} = \sigma(\beta^\intercal \tilde{x}^{(n)})$$
> $$\mathsf{P}(y^{(n)} = 0|\tilde{x}^{(n)}) = 1 - \sigma(\beta^\intercal \tilde{x}^{(n)}) = \sigma(-\beta^\intercal \tilde{x}^{(n)}).$$
>
> Here the inputs are augmented with a fixed unit input $\tilde{x}^{(n)} = (1, x^{(n)})$ which enables biases to be handled simply as $\beta^\intercal \tilde{x}^{(n)} = \beta_0 + \sum_{d=1}^{D} \beta_d x_d^{(n)}$.
> The probability of the dataset is a product of Bernoulli distributions,
>
> $$\mathsf{P}(y|X, \beta) = \prod_{n=1}^{N} \mathsf{P}(y^{(n)}|\tilde{x}^{(n)}) = \prod_{n=1}^{N} \sigma(\beta^\intercal \tilde{x}^{(n)})^{y^{(n)}} (1 - \sigma(\beta^\intercal \tilde{x}^{(n)}))^{1-y^{(n)}}.$$

b) Write pseudocode to estimate the parameters β using gradient ascent of the log-likelihood $\beta^{(new)} = \beta^{(old)} + \eta \frac{\partial}{\partial \beta}\mathcal{L}(\beta^{(old)})$ from X and y. You should use vectorised code: avoid having a loop over the rows of X. Include a short description detailing how you would choose the learning rate η.

c) The dataset for this practical can be found in the files `X.txt` and `y.txt`. You can load the data into python using

```
import numpy as np

X = np.loadtxt('X.txt')
y = np.loadtxt('y.txt')
```

This creates two variables given by

> `X` is a 1000×2 dimensional array containing two-dimensional input features for 1000 datapoints. Rows of `X` are denoted as column vectors $x^{(n)}$ below. Elements of `X` are denoted $x_d^{(n)}$.
>
> `y` is a 1000 dimensional binary vector containing the class labels. Elements of `y` are denoted $y^{(n)}$ below.

Visualise the dataset in the two-dimensional input space displaying each datapoint's class label. For this you can use the python function `plot_data` from the provided code. Discuss how well a classifier with a linear class boundary is likely to perform on these data.

d) Split the data randomly into training and test sets with 800 and 200 data points, respectively. Transform the pseudocode from the preparation exercise (a) into python code and use it to train the Logistic Classification method on the dataset. Report training curves showing the log-likelihood on training and test datasets per datapoint (averaged) as the optimisation proceeds, for this, you can use the functions `plot_ll` and `compute_average_ll` from the provided code. Visualise the predictions by adding probability contours to the plots made in part (c), for this you can use the functions `plot_predictive_distribution` from the provided code.

e) Report the final training and test log-likelihoods per datapoint. For the test data, apply a threshold to the probabilistic predictions so that those greater than $\tau = 1/2$ are assigned a positive predicted class label $\hat{y} = 1$ and those equal or below are assigned to a negative predicted class label $\hat{y} = 0$. Use these hard predictions to obtain and report the 2×2 confusion matrix:

		predicted label, $\hat{y}$	
		0	1
true label, y	0	fraction of true negatives $P(\hat{y} = 0 \mid y = 0)$	fraction of false positives $P(\hat{y} = 1 \mid y = 0)$
	1	fraction of false negatives $P(\hat{y} = 0 \mid y = 1)$	fraction of true positives $P(\hat{y} = 1 \mid y = 1)$

f) Expand the inputs through a set of radial basis functions (RBFs) centred on the training datapoints. The feature-expanded inputs now become $\tilde{x}_1^{(n)} = 1$ (to handle the bias terms as before) and $\tilde{x}_{m+1}^{(n)} = \exp\left(-\frac{1}{2l^2}\sum_{d=1}^{2}(x_d^{(n)} - x_d^{(m)})^2\right)$ (the RBF functions). In other words, the $(m+1)$th feature is given by a radial basis function centred on the mth training datapoint with width l. The dimensionality of the inputs should now be $N_{\text{train}} + 1$ where N_{train} is the number of training datapoints. For this, you can use the function `expand_inputs` from the provided code.

Train the Logistic Classification model on the feature-expanded inputs and display the new predictions for three choices of RBF width $l = \{0.01, 0.1, 1\}$. Visualise the predictions using probability contours as in part (d). For this, you can again use again the function `plot_predictive_distribution` with argument `map_inputs` given by `lambda x : evaluate_basis_functions(l, x, X_train)`, as indicated in the provided code. You will need to adjust the learning rate appropriately for each choice of length-scale, `l`.

g) Report the final training and test log-likelihoods per datapoint, the 2×2 confusion matrices for the three models trained in part (f). Compare the results to those obtained using the original inputs and explain your findings.

To avoid numerical errors when computing the log-likelihoods, you may find the following fact useful: $\log(\sigma(\beta^\intercal \tilde{x}^{(n)})) \to \beta^\intercal \tilde{x}^{(n)}$ as $\beta^\intercal \tilde{x}^{(n)} \to -\infty$.