Samuel Looper
1003044181
2019-02-27

ROB313 Assignment #2

1. Assignment Objectives

The purpose of this assignment is to evaluate the effectiveness of Generalized Linear Models with various permutations in the regression estimation of the standard data sets provided in the class material. The different permutations of the algorithm each have their own function calls, and include parameter passing to allow for the optimization and parameter sweep for values such as the regularization constants or basis function values. These functions accept a training set with feature vectors (x) and corresponding labels (y), a testing/validation set of feature vectors, and the various parameters that can be modified and optimized. First, a non-kernelized version is tested against the Mauna Loa set, and an optimal regularization parameter is found. Next, a kernelized GLM, utilizing Cholesky factorization, is created and tested against the Rosenbrock set. The computational cost and translational properties of the kernel are evaluated. Finally, radial basis function is used as the GLM kernel, and is evaluated on all 3 regression data sets. In addition to the coding portion, two proofs are solved on paper, scanned and included to the report.

2. Algorithm Overview

The initial Generalized Linear Model builds on the linear model from the first assignment. First, the phi matrices for the training and testing sets are constructed using the selected basis functions: linear, quadratic, and of the form x sin(x) and x cos(x), which were chosen based on the shape of the graph. Then, the matrix undergoes the GLM algorithm, inverted using SVD from the numpy library.

A kernelized GLM was used to abstract away the feature space and put the computation in terms of inner products through the Gram Matrix K. In conjunction, a Cholesky factorization was utilized, and the runtime of this algorithm was compared to the previous. In addition, the kernel was analyzed in terms of translational invariance by creating parameterized plots of the kernel output values.

3. Results

3.1 Non Kernelized General Linear Model

Based on an initial analysis of the Mauna Loa set, it was concluded that the data approximates some combination of a sinusoid, a linear/quadratic function, and a function of the form x sin(x) & x cos(x). Thus, these functions were selected as the basis function for the GLM. In addition, the frequency of the sinusoidal components was analyzed. It was estimated to have a period of 0.055, when the training set was plotted with respect to its 1D feature vector. The lambda value parameter sweep was done using the validation set, while the final result was plotted with the testing set

Optimal Lambda Value: 5

Testing RMSE: 0.085866

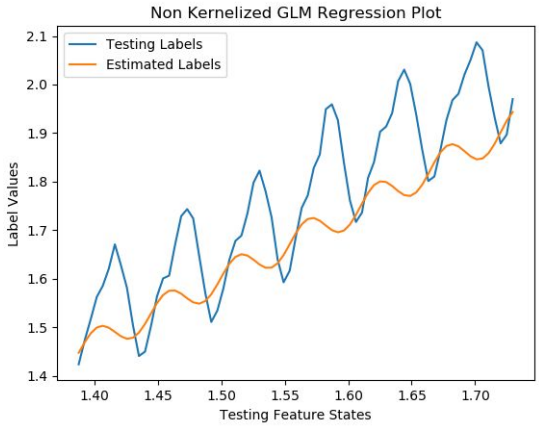

Fig 1: Result of the Mauna Loa regression

3.2 Kernelized General Linear Model

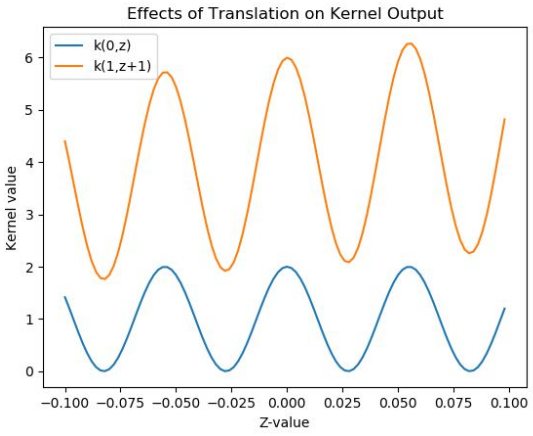

Fig 2: Kernel Output Analysis

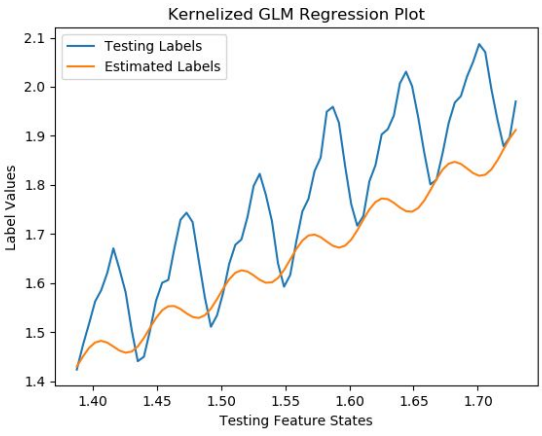

Fig 3: Result of the Mauna Loa regression

Next, the base algorithm was modified and its performance is reevaluated. The basis function space was collapsed using a kernelized implementation, and the weights were calculated using inner products in the dual representation using the Gram Matrix. The kernel was found by calculating the inner product of the basis function vectors for two arbitrary feature inputs. The kernel was then simplified by completing the square and scaling the linear inputs to create a single polynomial kernel term, and the double angle theorem was used to simplify the sinusoid function multiplications, yielding a final kernelized inner product expression:

$$k(x,z) = (1+x^T z)^2 + (1+x^T z)\, cos\, (\omega\, (x^T z))$$

In addition to kernelization, the Singular Value Decomposition was replaced with a Cholesky factorization. In the end, the memory and runtime complexity were significantly poorer with this iteration of the algorithm. On the one hand, with a small feature space, kernelizing the algorithm will actually lead to calculations in a much larger vector space as N >> M. For this reason as well, the N^3 term in the Cholesky's runtime complexity will dominate over the $11N(D+1)^3$ term in the SVD runtime.

Note: The algorithm had exactly the same RMSE error and optimal regression parameter as the previous algorithm

3.3 Radial Basis Function Kernel GLM

Finally, the kernelized version was modified, using a radial basis function Gaussian kernel rather than the previously derived kernel. In this case the kernel parameter as well as the regression parameter were swept, and the algorithm was applied to 2 regression sets and 1 classification set. The optimal parameters were found using the validation set, the RMSE value was calculated on the estimate for the testing set, and the results are summarized in the table below.

	Optimal λ	Optimal θ	Testing RMSE or Accuracy
Mauna Loa Data Set	0.001	1	0.124479
Rosenbrock Data Set	0.001	2	0.193240
Iris Data Set	1	0.5	0.870968

Fig 4: Table summarizing optimal parameters and associated success metric for 3 evaluated data sets

3.4 & 3.5 Proof Questions

Rob313 Assignment #2

Samuel Looper (1003044181)

④ $\underset{w \in \mathbb{R}^D}{\operatorname{argmin}} \underbrace{\|y - Xw\|_2^2 + w^T \Gamma w}_{=\circledast} = \text{arg}$

$\circledast = (y - Xw)^T (y - Xw) + w^T \Gamma w = y^T y + w^T X^T X w - 2 y^T X w + w^T \Gamma w$

$\frac{d}{dw} \circledast = [(X^T X)^T + (X^T X)] w + -(2 y^T X)^T + (\Gamma^T + \Gamma) w = 0$

$= 2 X^T X w - 2 X^T y + 2 \Gamma^T w = 0 = X^T (Xw - y) + \Gamma^T w = 0$

$\therefore \underset{w}{\operatorname{argmin}} \underbrace{X^T (Xw - y) + \Gamma w = 0}_{\text{minimizer}}$ is the same as $X^T y = (X^T X + \Gamma) w$

$w = (X^T X + \Gamma)^{-1} X^T y$

⑤ $\underset{\alpha \in \mathbb{R}^M}{\operatorname{argmin}} \underbrace{\sum_{i}^{N} (y^{(i)} - f(x^{(i)}, \alpha))^2 + \lambda \sum \alpha_i^2}_{\circledast} = \operatorname{argmin} \|y - \hat{y}\|_2^2 + \lambda \alpha^T \alpha$

$= \operatorname{argmin} (y - \hat{y})^T (y - \hat{y}) + \lambda \alpha^T \alpha$

$\hat{y} = [f(x^{(1)}, \alpha) \dots f(x^{(N)}, \alpha)]^T = [k(x^{(1)})^T \alpha \dots k(x^{(N)})^T \alpha]^T$

$= [k(x^{(1)}) \dots k(x^{(M)})]^T \alpha = k^T \alpha = k \alpha$

$\therefore$ we need to $\underset{\alpha \in \mathbb{R}^M}{\operatorname{argmin}} [y - k\alpha]^T [y - k\alpha] + \lambda \alpha^T \alpha$

$\circledast = y^T y + \alpha^T k k \alpha - 2 \alpha^T k y + \lambda \alpha^T \alpha$

$\therefore \frac{d}{d\alpha} \circledast = 0 = 2 k k \alpha - 2 k y + 2 \lambda \alpha$

$k y = (k k + \lambda I) \alpha \qquad \alpha = (k k + \lambda I)^{-1} k y$

where k is calculated from training set, y is from training set, λ can be found by k-fold cross validation & $(kk + \lambda I)^{-1}$ by SVD or QR factorization iff kk is full rank

ifferent to regularization w.r.t. w since $w^T w = \alpha^T k \alpha$ & the extra k term odifies the equation