



Assignment 2 (10 pts)

Due February 28, 2018

- Q1) 2pts** Implement a generalized linear model (GLM) that minimizes the least-squares loss function (using the SVD). By observing the structure of the one-dimensional `mauna_loa` training dataset, design a compact set of basis functions to construct a GLM for this dataset. Justify your design choices. Additionally, choose a regularization parameter λ , and use the validation set to verify your model. Finally, use both the training and validation sets to predict on the test set, plot the prediction relative to the test data, and present the test RMSE. Comment on the performance of your model.
- Q2) 2pts** Considering the same basis functions used in Q1, construct a kernelized GLM from the dual perspective. Select a positive regularization parameter, and construct the model using the Cholesky factorization. Comment on the computational cost and memory requirements of the dual approach versus the primal approach. Also, visualize your designed kernel by plotting $k(0, z)$, and by plotting $k(1, z + 1)$ where $z \in [-0.1, 0.1]$. Is your kernel translation-invariant (stationary)? Make your argument based on your plots of the kernel.
- Q3) 2pts** Construct a radial basis function (RBF) model that minimizes the least-squares loss function. Use a Gaussian kernel considering the grid of lengthscales $\theta = \{0.05, 0.1, 0.5, 1, 2\}$, consider the grid of regularization parameters $\{0.001, 0.01, 0.1, 1\}$, and construct the model using the Cholesky factorization. Select the hyperparameters across the grid of possible values by evaluating on the validation set. Construct the model on the datasets `iris`, `rosenbrock` (with `n_train=1000`, `d=2`), and `mauna_loa`. Use both the training and validation sets to predict on the test set, and format your results in a table (present test RMSE for regression datasets, and test accuracy for classification datasets).
- Q4) 2pts** Derive the minimizer of a linear regression model with a least-squares loss and general Tikhonov regularization whose training problem is defined as follows

$$\arg \min_{\mathbf{w} \in \mathbb{R}^D} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \mathbf{w}^T \mathbf{\Gamma} \mathbf{w},$$

where $\mathbf{\Gamma} \in \mathbb{R}^{D \times D}$ is a symmetric positive semi-definite matrix.

- Q5) 2pts** Considering the GLM

$$\hat{f}(\mathbf{x}, \boldsymbol{\alpha}) = \sum_{i=1}^N \alpha_i k(\mathbf{x}, \mathbf{x}^{(i)}),$$

derive a computational strategy to estimate $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \dots, \alpha_N\}^T \in \mathbb{R}^N$ by minimizing the objective function $\sum_{i=1}^N (y^{(i)} - \hat{f}(\mathbf{x}^{(i)}, \boldsymbol{\alpha}))^2 + \lambda \sum_{i=1}^N \alpha_i^2$. Compare this expression for the weights to those we derived in class using the dual representation. Are they different or the same? Explain why.

Submission guidelines: Submit an **electronic copy** of your report in **pdf** format, and **documented** python scripts. You should include a file named “README” outlining how the scripts should be run. Upload a single **tar** or **zip** file containing all files to Quercus. You are expected to verify the integrity of your **tar/zip** file before uploading. Do not include (or modify) the supplied *.npz data files or the **data_utils.py** module in your submission. The report must contain

- Objectives of the assignment
- A brief description of the structure of your code, and strategies employed
- Relevant figures, tables, and discussion

Do not use scikit-learn for this assignment, the intention is that you implement the simple algorithms required from scratch. Also, for reproducibility, always set a seed for any random number generator used in your code. For example, you can set the seed in numpy using `numpy.random.seed`