Math 189r
Homework 1
September 19, 2016

There are 8 problems in this set. 4 of the problems (you choose except this first set must include problem 1 and/or 2) are due on September 12, and the rest of the problems are due on September 19. Feel free to work with other students, but make sure you write up the homework and code on your own (no copying homework *or* code; no pair programming). Feel free to ask students or instructors for help debugging code or whatever else, though. When implementing algorithms you may not use any library (such as `sklearn`) that already implements the algorithms but you may use any other library for data cleaning and numeric purposes (`numpy` or `pandas`). Use common sense. Problems are in no specific order.

1 (**regression**). Download the data at `https://math189r.github.io/hw/data/online_news_popularity/online_news_popularity.csv` and the info file at `https://math189r.github.io/hw/data/online_news_popularity/online_news_popularity.txt`. Read the info file. Split the csv file into a training and test set with the first two thirds of the data in the training set and the rest for testing. Of the testing data, split the first half into a 'validation set' (used to optimize hyperparameters while leaving your testing data pristine) and the remaining half as your test set. We will use this data for the remainder of the problem. The goal of this data is to predict the **log** number of shares a news article will have given the other features.

(a) (**math**) Show that the maximum a posteriori problem for linear regression with a zero-mean Gaussian prior $\mathbb{P}(\mathbf{w}) = \prod_j \mathcal{N}(w_j|0,\tau^2)$ on the weights,

$$\arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log \mathcal{N}(y_i|w_0 + \mathbf{w}^\top \mathbf{x}_i, \sigma^2) + \sum_{j=1}^{D} \log \mathcal{N}(w_j|0,\tau^2)$$

is equivalent to the ridge regression problem

$$\arg\min \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + \mathbf{w}^\top \mathbf{x}_i))^2 + \lambda||\mathbf{w}||_2^2$$

with $\lambda = \sigma^2/\tau^2$.

(b) (**math**) Find a closed form solution $\mathbf{x}^\star$ to the ridge regression problem:

$$\text{minimize: } ||A\mathbf{x} - \mathbf{b}||_2^2 + ||\Gamma\mathbf{x}||_2^2.$$

(c) (**implementation**) Attempt to predict the log shares using ridge regression from the previous problem solution. Make sure you include a bias term and *don't regularize the bias term*. Find the optimal regularization parameter λ from the validation set.

Plot both λ versus the validation RMSE (you should have tried at least 150 parameter settings randomly chosen between 0.0 and 150.0 because the dataset is small) and λ versus $||\boldsymbol{\theta}^\star||_2$ where $\boldsymbol{\theta}$ is your weight vector. What is the final RMSE on the test set with the optimal $\lambda^\star$?

(d) (**math**) Consider regularized linear regression where we pull the basis term out of the feature vectors. That is, instead of computing $\hat{\mathbf{y}} = \boldsymbol{\theta}^\top \mathbf{x}$ with $\mathbf{x}_0 = 1$, we compute $\hat{\mathbf{y}} = \boldsymbol{\theta}^\top \mathbf{x} + b$. This corresponds to solving the optimization problem

$$\text{minimize: } ||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2.$$

Solve for the optimal $\mathbf{x}^\star$ explicitly. Use this close form to compute the bias term for the previous problem (with the same regularization strategy). Make sure it is the same.

(e) (**implementation**) We can also compute the solution to the least squares problem using gradient descent. Consider the same bias-relocated objective

$$\text{minimize: } f = ||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2.$$

Compute the gradients and run gradient descent. Plot the ℓ_2 norm between the optimal $(\mathbf{x}^\star, b^\star)$ vector you computed in closed form and the iterates generated by gradient descent. Hint: your plot should move down and to the left and approach zero as the number of iterations increases. If it doesn't, try decreasing the learning rate.

2 (**MNIST**) Download the training set at `http://pjreddie.com/media/files/mnist_train.csv` and test set at `http://pjreddie.com/media/files/mnist_test.csv`. This dataset, the MNIST dataset, is a classic in the deep learning literature as a toy dataset to test algorithms on. The problem is this: we have 28×28 images of handwritten digits as well as the label of which digit $0 \leq \texttt{label} \leq 9$ the written digit corresponds to. Given a new image of a handwritten digit, we want to be able to predict which digit it is. The format of the data is `label, pix-11, pix-12, pix-13, ...` where `pix-ij` is the pixel in the `i`th row and `j`th column.

(a) (**logistic**) Restrict the dataset to only the digits with a label of 0 or 1. Implement L2 regularized logistic regression as a model to compute $\mathbb{P}(y = 1|\mathbf{x})$ for a different value of the regularization parameter λ. Plot the learning curve (objective vs. iteration) when using Newton's Method *and* gradient descent. Plot the accuracy, precision ($p = \mathbb{P}(y = 1|\hat{y} = 1)$), recall ($r = \mathbb{P}(\hat{y} = 1|y = 1)$), and F1-score ($F1 = 2pr/(p + r)$) for different values of λ (try at least 10 different values including $\lambda = 0$) on the test set and report the value of λ which maximizes the accuracy on the test set. What is your accuracy on the test set for this model? Your accuracy should definitely be over 90%.

(b) (**softmax**) Now we will use the whole dataset and predict the label of each digit using L2 regularized softmax regression (multinomial logistic regression). Implement this using gradient descent, and plot the accuracy on the test set for different values of λ, the regularization parameter. Report the test accuracy for the optimal value of λ as well as it's learning curve. Your accuracy should be over 90%.

(c) (**KNN**) Solve the same problem posed in part (b) but use K-Nearest Neighbors instead of softmax regression and vary k instead of λ. Only try 3 values for k $(1, 5,$ and $10)$ and the ℓ_2 norm as your metric. Plot and report the same results as part (b).

3 (**Murphy 2.11**) Derive the normalization constant (Z) for a one dimensional zero-mean Gaussian

$$\mathbb{P}(x;\sigma^2) = \frac{1}{Z}\exp\left(-\frac{x^2}{2\sigma^2}\right)$$

such that $\mathbb{P}(x;\sigma^2)$ becomes a valid density.

4 (**Murphy 2.15**) Let $\mathbb{P}_{emp}(x)$ be the empirical distribution and let $q(x|\boldsymbol{\theta})$ be some model. Show that $\arg\min_q \mathbb{KL}(\mathbb{P}_{emp}||q)$ is obtained by $q(x) = q(x;\hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q,\mathcal{D})$ is the maximum likelihood estimate.

5 (**Linear Transformation**) Let $\mathbf{y} = A\mathbf{x} + \mathbf{b}$ be a random vector. show that expectation is linear:

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[A\mathbf{x} + \mathbf{b}] = A\mathbb{E}[\mathbf{x}] + \mathbf{b}.$$

Also show that

$$\text{cov}[\mathbf{y}] = \text{cov}[A\mathbf{x} + \mathbf{b}] = A\text{cov}[\mathbf{x}]A^\top = A\boldsymbol{\Sigma}A^\top.$$

6 (**Murphy 2.16**) Suppose $\theta \sim \text{Beta}(a, b)$ such that

$$\mathbb{P}(\theta; a, b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Beta function and $\Gamma(x)$ is the Gamma function. Derive the mean, mode, and variance of θ.

7 (**Murphy 8.3**) Gradient and Hessian of the log-likelihood for logistic regression.

(a) Let $\sigma(x) = \frac{1}{1+e^{-x}}$ be the sigmoid function. Show that

$$\sigma'(x) = \sigma(x)\left[1 - \sigma(x)\right].$$

(b) Using the previous result and the chain rule of calculus, derive an expression for the gradient of the log likelihood for logistic regression.

(c) The Hessian can be written as $\mathbf{H} = \mathbf{X}^\top\mathbf{S}\mathbf{X}$ where $\mathbf{S} = \text{diag}(\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n))$. Derive this and show that $\mathbf{H} \succeq 0$ ($A \succeq 0$ means that A is positive semidefinite).

8 (**Murphy 9**) Show that the multinomial distribution

$$\text{Cat}(x|\boldsymbol{\mu}) = \prod_{k=1}^{K}\mu_k^{x_k}$$

is in the exponential family and show that the generalized linear model corresponding to this distribution is the same as multinomial logistic regression.