Assignment 2

CS5824/ECE5424 – Fall 2020

Out: September 29, 2020
Due: October 09 (11:59pm)

For Question 1, submit both PDF and .ipynb file with all of your code and output. Both Colab and Jupyter are allowed. For Questions 2-4, submit a separate PDF file compiled from Latex. Please submit all three files through Canvas, no need to zip them all.

Late submissions incur a 0.5% penalty for every rounded up hour past the deadline for the first 24 hours and 0.6% penalty for every rounded up hour past the deadline for the second 24 hours. For example, an assignment submitted 5 hours and 15 min late will receive a penalty of ceiling(5.25) * 0.5% = 3%. A grade of zero will be given on the third late day.

Be sure to include your name and student number with your assignment.

1. **[40 pts]** Implement the following classification algorithm. Do not use any machine learning library, but feel free to use libraries for linear algebra and feel free to verify your results with existing machine learning libraries. Use the same dataset (handwritten digits) as for assignment 1 to train the algorithms.

   Logistic regression: let $\Pr(C_1|x) = \sigma(\mathbf{w}^T\mathbf{x} + w_0)$ and $\Pr(C_2|\mathbf{x}) = 1 - \sigma(\mathbf{w}^T\mathbf{x} + w_0)$. Learn the parameters $\mathbf{w}$ and w_0 by conditional likelihood maximization. More specifically use Newton's algorithm derived in class to optimize the parameters. 10 iterations of Newton's algorithm should be sufficient for convergence. Add a penalty of $0.5\lambda||\mathbf{w}||_\mathbf{2}^\mathbf{2}$ to regularize the weights. Find the optimal hyperparameter λ by 10-fold cross-validation. Additional input: vary λ from 0 to 4 with the step of 0.1 and use mean error.

   What to hand in:

   - **[10 pts]** Draw a graph that shows the cross-validation accuracy of logistic regression as λ varies. Report the best λ.
   - **[5 pts]** Report the accuracy of logistic regression (with the best λ for regularization) on the test set.
   - **[5 pts]** Print the parameters $\mathbf{w}, w_0$ found for logistic regression.
   - **[10 pts]** Briefly discuss the results:
     - Logistic regression finds a linear separator where as k-Nearest Neighbours (in assignment 1) finds a non-linear separator. Compare the expressivity of the separators. Discuss under what circumstances each type of separator is expected to perform best. What could explain the results obtained with KNN in comparison to the results obtained with regression?
   - **[10 pts]** A copy of your code.

2. **[20 pts]** A Bernoulli distribution has the following likelihood function for a data set $\mathcal{D}$:

$$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}, \tag{1}$$

where N_1 is the number of instances in data set $\mathcal{D}$ that have value 1 and N_0 is the number in $\mathcal{D}$ that have value 0. The maximum likelihood estimate is

$$\hat{\theta} = \frac{N_1}{N_1 + N_0}. \tag{2}$$

   (a) **[10 pts]** Derive the maximum likelihood estimate above by solving for the maximum of the likelihood. I.e., show the mathematics that get from eq:bernoulli to eq:mle.

   (b) **[10 pts]** Suppose we now want to maximize a posterior likelihood

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}, \tag{3}$$

   where we use the Bernoulli likelihood and a (slight variant[1] of α) symmetric Beta prior over the Bernoulli parameter

$$p(\theta) \propto \theta^{\alpha}(1-\theta)^{\alpha}. \tag{4}$$

   Derive the maximum posterior mean estimate.

3. **[20 pts]** As we know the Bayes rule

$$P(C_k|x) = \frac{P(x|C_k)P(c_k)}{P(x)}$$

Assume a Binary class classification problem where the lung cancer need to be classified as malign (C_1) or Begin (C_0). Classification can be done based on the features which include the observed symptoms which are captured in the data vector x.

   (a) **[10 pts]** Describe what does $P(C_0|x), P(x|C_0), P(C_0), P(C_1|x), P(x|C_1)$ and $P(C_1)$ mean in a few words.

   (b) **[10 pts]** Given

$$P(C_1|x) = \frac{1}{1 + exp(-w^T x)}$$

   where, w is a vector of weights that needs to be estimated. Find $P(C_0|x)$ and $\log \frac{P(C_1|x)}{P(C_0|x)}$

4. **[20 pts]** This question uses a discrete probability distribution known as the Poisson distribution. A discrete random variable X follows a Poison distribution with parameter λ if

$$Pr(X = k) = \frac{\lambda^k}{k!}e^{-\lambda} \tag{5}$$

where $k \in \{1, 2, \cdots\}$ Imagine we have a gumball machine which dispenses a random number of gumballs each time you insert a quarter. Assume that the number of gumballs dispensed is Poisson distributed (i.i.d) with parameter λ. Curious and sugar-deprived, you insert 8 quarters one at a time and record the number of gumballs dispensed on each trial:

...*Trial*	1	2	3	4	5	6	7	8
Gumballs	4	1	3	5	5	1	3	8

(6)

Let $G = (G_1, \cdots G_n)$ be a random vector where G_i is the number of gumballs dispensed on trial i:

   (a) **[10 pts]** Give the log-likelihood function of G given λ.

   (b) **[5 pts]** Compute the MLE for λ in the general case.

   (c) **[5 pts]** Compute the MLE for λ using the observed G.

[1] For convenience, we are using the exponent of α instead of the standard $\alpha - 1$.