CS536 Homework 2

March 5, 2018

1 Problem

Consider the problem of modeling the concept described as $p(X, Y)$, where Y are the class labels. You seek to find the best predictive model $Y \approx g(X)$ for this concept. Show that:

1. The optimal Bayes rule $g^*(x)$ (i.e., the rule that results in the minimum Bayes error) also minimizes the median absolute loss (i.e., the L1 norm) $\mathbb{E}[|g(X) - Y|]$.
2. The L2 norm $\mathbb{E}[(g(X) - Y)^2]$ is minimized by the posterior $\eta(X) = P(Y|X)$.
3. Suppose you design the randomized rule $g(x) \sim \eta(x)$ (i.e., randomly pick a decision from the posterior $\eta(x)$). Show that this randomized rule does *NOT* lead to an error lower than the deterministic optimal Bayes rule.

2 Problem

Consider a two-class classification case where we now constrain the probability of error of one class to be fixed, $\epsilon_1 = P(error|\omega_1) = \epsilon$. Show the minimizing the error of the other class, $\epsilon_2 = P(error|\omega_2)$, results in the rule

$$\text{decide } x \in \omega_1 \text{ if } \frac{P(\omega_1|x)}{P(\omega_2|x)} > \theta,$$

where θ is selected so that the constraint above is satisfied.

Hint: Use a Lagrange multiplier to show that this problem is equivalent to minimizing the quantity

$$q = \theta(\epsilon - \epsilon_1) + \epsilon_2.$$

3 Problem

Let $p(\mathbf{x}|\omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma^2\mathbf{I})$ for a two-category d-dimensional problem with equal class priors.

1. Show that the minimum probability of error is given by

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-u^2/2} du,$$

   where $a = ||\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1||/(2\sigma)$.

2. Let $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2 = [\mu_1 \dots \mu_d]^t \neq \mathbf{0}$. What happens with the probability of error as $d \to \infty$? Hint: use the inequality

$$\int_a^\infty e^{-u^2/2} du \leq \frac{1}{a} e^{-a^2/2}.$$

4 Problem

Let $X \sim \mathcal{N}(0,1)$ and $Y = W \cdot X$, where $P(W = +1) = P(W = -1) = 0.5$. Clearly, X and Y are not independent. Show that:

1. $Y \sim \mathcal{N}(0,1)$.
2. $cov[X,Y] = 0$. Thus X and Y are uncorrelated but dependent, even though they are both Gaussian. Hint: you can use the rule of iterated expectation $\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY|W]]$.
3. Sample four sets of 100 random points X and construct the corresponding 100 points $Y|X$. Draw the scatter plots of $\mathcal{D}_i = \{(X_n, Y_n)\}_{n=1}^{N=100}$ for those four sets, $i = 1, \ldots, 4$.

   Note: Use `numpy.random.seed(0)` before sampling points for the four sets. (But do this only once, not for each set.) To sample Gaussian points use `numpy.random.randn` and use `numpy.random.rand` to create samples of $Y|X$.
4. Now find the optimal least sqaures linear regressors $Y \approx w \cdot X + w_0$ for the four datasets. Overlay those regressors on the scatter plots above. Comment on differences between the four cases and explain why there are differences, if any.

5 Problem

5.1 Linear Regression for News group Classification

You will implement Linear Regression for classification problem on the NEWS_DATASET. ('alt.atheism','talk.religion.misc','comp.graphics','sci.space') . You are going to use the 4 classes among the original 20 classes. You will encode group labels based on the two schemes. One is discrete scalar label and the other one is one-hot vector. For the group "3" example, the encoding can be one-hot vector representation like "[0,0,1,0]" or it can be represented by "2" (discrete labels). For each of the two cases, we will learn the linear regression model $y = w^T \cdot \phi(x)$ and define an classifier function C(y) mapping the y to the categorical encodings(discrete scalar label or one-hot vector). Also, you will compare their performance based on the two design cases of $\phi(x)$.

5.1.1 DATA FILE : train_PCA.pkl and test_PCA.pkl

The files, Train_PCA and Test_PCA, are arrays that each shape is (2034,1500) and (1353,1500). PCA(Principal Components Analysis) is applied to the previous Tf-idf features samples (2034, 33810) and we used only 95% engery principal components and the reduced dimension is 1500.

5.1.2 Discrete Label Experiment

1. We will encode k^{th} data sample's label, c_k, into an element of the set $\{0, 1, 2, 3\}$. This implies that you can use the data samples' target labels as they are.
2. Your working files are :
   - hw2_student → Linear_Regression_Scalar.ipynb
   - hw2_student → cs536_2 → models → Linear_Regression_Classifier.py
3. In your model of $y = w^T \cdot \phi(x)$, specify the dimension(shape) of w , x , and y.
4. In this experiment, we will use the $\phi(x) = x$
5. What is your C(y) function choice? Draw your C(y) on the domain of y. Why do you choose the function?
6. Fill out the working files with your code as appropriate. For the cost, we will use the mean square error(MSE) with quadratic regularization, where n is the number of samples,

$$E(w) = \sum_{k=1}^{n} \frac{1}{2}\{c_k - C(w^T \cdot x_k)\}^2 + L||w||^2 \tag{1}$$

7. In the experiment, please compare the train and validation accuracies along with the different L values. Draw a plot $\log(L)$ vs train and $\log(L)$ vs validation accuracies together. What value L would you choose based on the plot? Why?

8. Based on the your choice of L, re-train the model and find the test accuracy.

5.1.3 One-Hot Vector Label Experiment

1. We will encode k^{th} sample's group label c_k into an element of the set $\{[1,0,0,0],[0,1,0,0],[0,0,1,0],[0,0,0,1]\}$.

2. Your working files are :

   - hw2_student → Linear_Regression_hotvector.ipynb
   - hw2_student → cs536_2 → models → Linear_Regression_Classifier.py

3. In your model of $y = w^T \cdot \phi(x)$, specify the dimension (shape) of w , x , and y.

4. In this experiment, we will use the $\phi(x) = x$

5. If you do multi-class classification (choose only one), then what is your C(y) function choice?

6. Fill out the working files with your code as appropriate. we will use the mean square error(MSE) with quadratic regularization, where n is the number of samples,

$$E(w) = \sum_{k=1}^{n} \frac{1}{2} ||c_k - C(w^T \cdot x_k)||^2 + L||w||^2 \tag{2}$$

   In the experiment, please compare the train and validation accuracies along with the different L values. Draw a plot $\log(L)$ vs train and $\log(L)$ vs validation accuracies together. What value L would you choose based on the plot? Why?

7. Based on the your choice of L, re-train the model and find the test accuracy.

8. Plot the ROC curves for each group.

5.1.4 RBF feature $\phi_m(x) = exp\frac{-(x-\mu_m)^2}{2\cdot\sigma_m^2}$ (m^{th} component feature)

1. We will transform 1500 dimensional PCA features to 1000 dimensional RBF feature set.

2. You will be given the K-means clustering results with 1000 clusters. You could construct 1000 RBF functions by using the clustering results. File : Kmean_cluster.pkl

3. You will implement your own name RBF_featre.py which creates train_RBF.pkl and test_RBF.pkl based on the RBF functions.

4. By using the new features (train_RBF.pkl and test_RBF.pkl) and the files(Linear_Regression_hotvector.ipynb and Linear_Regression_Scalar.ipynb), find your L and the test result for the RBF features.

5. Fill out the Table 1.

6. For the one-hot vector, plot the ROC curves for the RBF.

5.2 Analysis

1. We learned w based on the mean square error (MSE) formulation. What does the MSE assume about $P(C = c|w, \phi(x))$ as we learn w based on the maximum a posteriori estimation? Do you think it is the right approach? If you can design the $P(C = c|w, \phi(x))$ again, what probabilistic density would you use and why?

2. Please compare the test performance of the above two implementations 5.1.2 and 5.1.3. Why one is better than the other?

3. In Table 1, does the RBF function help to improve the inferior encoding scheme? If it does, why it can improve the one? If it does not, what remedy would you suggest?

Table 1: Test Accuracy and L

Basis function	*Encoding*	
	Discrete Scalar	**One-Hot vector**
$\phi(x) = x$	test accuracy and L	
$\phi_m(x) = exp\frac{-(x-\mu_m)^2}{2\cdot\sigma_m^2}$		

6 Problem: Optimal Bayesian Estimation

In this problem you are going to implement optimal Bayesian estimation model for classification of NEWS_DATASET base on different scenarios. Please open OptBayesEstim.ipynb and follow the instruction form there. All implementations and reports must be reported in that file. At the end of your work, just submit the Opt-BayesEstim.ipynbfile.