1 Logistic Regression (26 Points)

(a) Consider a binary logistic regression model, given n training examples $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)$, please write down the negative log likelihood (as loss function):

$$\mathcal{L}(\mathbf{w}) = -\log\left(\prod_{i=1}^{n} P(Y = y_i|\mathbf{X} = \mathbf{x}_i)\right).$$

(b) Use Gradient Descent Method to find the update rule for $\mathbf{w}$. Will this solution converge to a global minimum? Provide your reasoning.

(c) Let us extend the binary logistic regression model to handle multi-class classification. Assume we have K different classes, and the posterior probability for class k is given by:

$$P(Y = k|\mathbf{X} = \mathbf{x}) = \frac{\exp(\mathbf{w}_k^\top \mathbf{x})}{1 + \sum_1^{K-1} \exp(\mathbf{w}_t^\top \mathbf{x})}, \qquad \text{for } k = 1, ..., K-1$$

$$P(Y = k|\mathbf{X} = \mathbf{x}) = \frac{1}{1 + \sum_1^{K-1} \exp(\mathbf{w}_t^\top \mathbf{x})}, \qquad \text{for } k = K \tag{1}$$

Please write down the negative log likelihood $\mathcal{L}(\mathbf{w}_1, ..., \mathbf{w}_K)$, where we can simplify the multi-class logistic regression expression above by introducing an additional fixed parameter $\mathbf{w}_K = \mathbf{0}$.

(d) Please use the negative log likelihood $\mathcal{L}(\mathbf{w}_1, ..., \mathbf{w}_K)$ from (c) to compute the with gradient respect to w_i, i.e. $\frac{\partial \mathcal{L}(\mathbf{w}_1, ..., \mathbf{w}_K)}{\partial \mathbf{w}_i}$, and provide the update rule with Gradient Descent for $\mathbf{w}_i$.

2 Linear/Gaussian Discriminant (24 Points)

(a) Consider the Gaussian Discriminant Analysis, given n training examples $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$, with $y_n \in \{1, 2\}$, where

$$p(x_n, y_n) = p(y_n)p(x_n|y_n) = \begin{cases} p_1 \frac{1}{\sqrt{2\pi}\sigma_1} \exp(-\frac{(x_n - \mu_1)^2}{2\sigma_1^2}) & \text{if } y_n = 1 \\ p_2 \frac{1}{\sqrt{2\pi}\sigma_2} \exp(-\frac{(x_n - \mu_2)^2}{2\sigma_2^2}) & \text{if } y_n = 2 \end{cases}$$

Please write the the log likelihood function $\mathcal{L}(\mathcal{D})$, and use MLE to find $(p_1^*, p_2^*, \mu_1^*, \mu_2^*, \sigma_1^*, \sigma_2^*)$ that maximizes $\mathcal{L}(\mathcal{D})$.

(b) Consider a set of samples from two classes c_1 and c_2. Suppose $p(\mathbf{x}|y = c_1)$ follows a multivariate Gaussian distribution $\mathcal{N}(\mu_\mathbf{1}, \mathbf{\Sigma})$, and $p(\mathbf{x}|y = c_2)$ follows a multivariate Gaussian distribution $\mathcal{N}(\mu_\mathbf{2}, \mathbf{\Sigma})$ $(\mu_\mathbf{1}, \mu_\mathbf{2} \in \mathcal{R}^D, \mathbf{\Sigma} \in \mathcal{R}^{D \times D})$. Show that $p(y|\mathbf{x})$ follows a logistic function, i.e., $p(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^\top \mathbf{x})}$ for some $\boldsymbol{\theta}$.

3 Programming - Linear Regression (50 points)

In this assignment, you will study the linear regression as well as some standard feature selection techniques. You are allowed to use MATLAB or Python scripts. For MATLAB, you will find R2013b in http://software.usc.edu/matlab/. Without specific description, you are not allowed using Matlab toolbox functions like *knnclassify*, *knnsearch*. For Python, we only allow Python 2.7 and we strongly recommend to use Anaconda2 (https://www.continuum.io/downloads) for Python 2.7 and you are allowed using other libraries in Anaconda2 except machine learning packages such as *scikit-learn*, unless specified otherwise. Your script should be executable under the Anaconda2 environment and we will grade your code in the same environment. You can build your own functions or modules, however, you should be careful to include them into your submission. If you use other packages not included in Anaconda2 and we fail to run your code, we won't regrade it after installing required packages. You should implement all optimization algorithms by yourself. Below, we describe the steps that you need to take to accomplish this programming assignment.

3.1 Dataset

Loading We will use the *Boston Housing Data Set* from UCI's machine learning data repository. You can download the dataset from https://archive.ics.uci.edu/ml/datasets/Housing. If you are using Python, you are encouraged to use the *sklearn.datasets* module to load the data. See http://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html for example.

The dataset contains 506 data points, where each data point has 13 features attributes and 1 target attribute. The task is to predict the value of the target using the values of 13 features. If you load the data from the UCI website, the last attribute ($MEDV$) is the target. If you load the data using *sklearn.datasets* module, the target attribute is already separated into the *target* field.

Training/Test Splitting You need to separate the 506 data points into training/test sets. For the purpose of grading, we define the following splitting procedure. The test set consists of all $(7i)$-th data points (where $i = 0, 1, 2, \ldots, 72$), and the rest will be used as training set. This will create a test set of size 73 and a training set of size 433.

Note: Make sure you perform the splitting correctly. The rest of the tasks will depend on the correct splitting.

Data Analysis To understand the data at hand is an important step toward solving the machine learning task. You need to plot the histograms of all numerical attributes and calculate the Pearson correlation with the target value. It should be done on the training data. For numerical attributes, the histogram should be plotted using 10 bins with equal width.

Data Preprocessing Standardize all features so they are centered at 0 and the standard deviation is 1. Remember that you should only use the training data. You need to remember the transformation so that you can apply them on the test data for evaluation.

3.2 Linear Regression

Let $\boldsymbol{x}_i$ and y_i be the feature vector and target value of the i-th data point, respectively. You need to implement linear regression and ridge regression algorithms and apply them on the Boston housing data. Discuss the performance of these algorithms.

Linear Regression Train a linear regressor on the training data, report the mean squared loss (MSE) on both training and test sets. The MSE is defined as

$$\text{MSE}\left(\boldsymbol{y}^{\text{true}}, \boldsymbol{y}^{\text{pred}}\right) = \frac{1}{N}\sum_{i=1}^{n}\left(y_i^{\text{true}} - y_i^{\text{pred}}\right)^2,$$

where $\boldsymbol{y}^{\text{true}}$ is the ground truth and $\boldsymbol{y}^{\text{pred}}$ is the prediction.

The linear regressor parametrized by $(\boldsymbol{w}, b)$ is defined by the following optimization problem

$$\min_{\boldsymbol{w},b} \frac{1}{N}\left(f(\boldsymbol{x}_i, \boldsymbol{w}, b) - y_i\right)^2$$

where

$$f(\boldsymbol{x}_i, \boldsymbol{w}, b) = \boldsymbol{w}^\top \boldsymbol{x}_i + b.$$

Hint: there is analytically solution to the above optimization problem.

Ridge Regression Train a Ridge regressor on the training data, report the mean squared loss (MSE) on both training and test sets.

The linear regressor parametrized by $(\boldsymbol{w}, b)$ is defined by the following optimization problem

$$\min_{\boldsymbol{w},b} \frac{1}{N}\left(f(\boldsymbol{x}_i, \boldsymbol{w}, b) - y_i\right)^2 + \lambda||\boldsymbol{w}||_2^2$$

where

$$f(\boldsymbol{x}_i, \boldsymbol{w}, b) = \boldsymbol{w}^\top \boldsymbol{x}_i + b.$$

Try $\lambda = 0.01, 0.1, 1.0$, and report the results.

Hint: there is also analytically solution to the above optimization problem.

Ridge Regression with Cross-Validation Finding the best λ plays an important role in using the Ridge regression. You should using 10-fold cross-validation (CV) on the training set to evaluate the choice of λ. Select λ from $[0.0001, 10]$. Report the CV results for different λ and their MSE on test set.

3.3 Feature Selection

In this section, we will solve the same problem but using fewer features. We will select 4 features out of the 13 features to predict the price. Discuss the difference between different approaches.

Selection with Correlation Try the following schemes:

(a) Select the 4 features with the highest correlation with the target (in absolute value).

(b) First select one feature with the highest correlation with the target (in absolute value), train a linear regressor. In the remaining features, find another feature with the highest correlation with the residue of previous regressor. Include this feature in the linear regressor and update the residue. Proceed until you find all 4 features. The residue is defined as the difference between the true target value and the predicted value.

Report the results for both schemes.

Selection with Mutual Information Select the 4 features with the highest mutual information with the target. For numerical features, you should discretize them into 10 bins, where each bin contains the same number of data points. The correct bin assignment can be found by sorting the value of that feature. Report the results.

Selection with Brute-force Search Try all combination of 4 features, report the best combination of features and the results.

3.4 Polynomial Feature Expansion

Now, you should expand the 13 features through polynomial expansion. That is, create new features by multiplying the old features together, i.e., $x_i * x_j$ for $i, j = 0, 1, \ldots, 12$. Create the new expanded data with $104 = 13 + (1 + 13) * 13/2$ features, properly standardize the new features, apply the linear regression model and report the results.

Submission Instruction: You need to provide the followings:

- Provide your answers to problems in hardcopy. The papers need to be stapled and submitted into the locker#19 on the first floor of PHE building.
- Provide your answers to problems in `*.pdf` file, named as `CSCI567_hw2_fall16.pdf`. You need to submit the homework in both hardcopy (at the collection locker #19 at the PHE building 1st floor by **5:00pm** of the deadline date) and electronic version as `*.pdf` file on Blackboard. If you choose handwriting instead of typing all the answers, you will get 40% points deducted.
- Submit ALL the code and report via Blackboard. The only acceptable language is MATLAB or Python2.7. For your program, you MUST include the main script called `CSCI567_hw2_fall16.m` or `CSCI567_hw2_fall16.py` in the root of your folder. After running this main file, your program should be able to generate all of the results needed for this programming assignment, either as plots or console outputs. You can have multiple files (i.e your sub-functions), however, the only requirement is that once we unzip your folder and execute your main file, your program should execute correctly. Please double-check your program before submitting. You should only submit one `*.zip` file. No other formats are allowed except `*.zip` file. Also, please name it as `[lastname]_[firstname]_hw2_fall16.zip`.

Collaboration You may collaborate. However, collaboration has to be limited to discussion only and you need to write your own solution and submit separately. You also need to list with whom you have discussed.