CS4801: Principle of Machine Learning
Assignment 1

Due on 22nd August

This homework consists of problems covering application of regression, linear algebra, probability, regression and little bit of Bayesian classification. A few instructions to make life easier for all of us:

- Conceptual assignment need not to be submitted.
- We will have a short quiz on the question of this assignment (5 points) for 15 minutes on 22nd August 17:30 pm.
- In short quiz please write concisely and clearly. There are points for intermediate steps, but not in "talking problems to death."
- For programming assignment please submit your code and a short discussion on your observation (preferably PDF and latex) from your experiments. Put all codes and report in a single zipped file and name it as <First-name><Last-name>.zip. Then submit it in moodle.
- Deadline for programming assignment is 17:00 pm 22nd August 2017.

1 Conceptual Exercises

Exercise 1 : Application

(a) (1.5 points) Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in prediction. Finally, provide n(number of sample) and p(number of features).

i. We collect a set of data on the top 500 companies in India. For each company we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

ii. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

iii. We are interested in predicting the % of change in the INR in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the INR, the % change in the US market, the % change in the British market, and the % change in the German market.

(b) (1.5 points) Write one real life problem (other than the above three problems) for each of the following machine learning models. Also explain why you will choose to apply this ML technique to solve that specific problem.

i. Ridge regression

ii. LASSO

iii. logistic regression

Exercise 2: Linear Algebra

(a) (3 points) Any real, symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ has the decomposition $A = \mathbf{U}\Sigma\mathbf{U}^T$, where $\Sigma \in \mathbb{R}^{n \times n}$ is a diagonal matrix with the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\mathbf{A}$ on the diagonal and $\mathbf{U}$ is an orthogonal matrix in $\mathbb{R}^{n \times n}$, that is $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}$, which contains the corresponding eigenvectors (more precisely: an orthogonal basis of the eigenspace of the corresponding eigenvalue).

i. What are the eigenvalues and eigenvectors of $\mathbf{A}^k$ (matrix product with itself) for $k \in \mathbb{N}$.

ii. Let $\mathbf{U} \in \mathbb{R}^{n \times n}$ be an orthogonal matrix - show that $\|\mathbf{U}^T\mathbf{x}\|_2 = 1$ for any vector $\mathbf{x} \in \mathbb{R}^n$ with $\|\mathbf{x}\|_2 = 1$.

iii. Use the previous result to show that $max_{\|\mathbf{x}\|_2=1}\mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda_{max}$, where λ_{max} is the largest eigenvalue of $\mathbf{A}$. [hint: $max_{\|\mathbf{x}\|_2=1}\mathbf{x}^T\mathbf{A}\mathbf{x}$ means $max\mathbf{x}^T\mathbf{A}\mathbf{x}$ such that $\|\mathbf{x}\|_2 = 1$ (a constraint)]

[Hint: The last part can be solved using the decomposition of $\mathbf{A}$ and then doing a variable transformation $\mathbf{y} = \mathbf{U}^T\mathbf{x}$. Then one gets a very simple optimization problem for $\mathbf{y}$.]

Exercise 3: Multivariate Analysis

(a) (3 points) Compute the derivative $\triangledown_{\mathbf{x}} f$ for following functions $f : \mathbb{R}^d \to \mathbb{R}$ with respect to $\mathbf{x}$, where $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{d\times d}$, $B = (b_{ij}) \in \mathbb{R}^{m\times d}$

i. $f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w}\rangle = \mathbf{x}^T\mathbf{w} \quad \Rightarrow \quad \triangledown_{\mathbf{x}} f = \mathbf{w}$,

ii. $f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle = \mathbf{x}^T\mathbf{A}\mathbf{x} \quad \Rightarrow \quad \triangledown_{\mathbf{x}} f = \mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x}$,

iii. $f(\mathbf{x}) = \|\mathbf{B}\mathbf{x}\|_2^2 \quad \Rightarrow \quad \triangledown_{\mathbf{x}} f = 2\mathbf{B}^T\mathbf{B}\mathbf{x}$,

Exercise 4: Basic Probability

(a) (1 point) X and Y are two random variables and $Y = mX + c$ where m and c are not random variables. Prove that the correlation coefficient of X and Y is 1. [Hints : correlation coefficient is $corr(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$]

(b) (2 points) Derive the maximum likelihood estimation for parameter μ of following distributions

i. $\mathbf{x}$ be a p-dimensional real vector with multi-variate Gaussian distribution, i.e.,

$$P(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$$

where $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p\times p}$ are parameters (mean and covariance) of the distribution. Find the maximum likelihood estimation for μ.

ii. $\mathbf{x}$ be a p-dimensional binary (0 or 1) vector with a multivariate Bernoulli distribution.

$$(x|\mu) = \prod_{i=1}^{p} \mu_i^{x_i}(1 - \mu_i)^{(1-x_i)}$$

where $\mu \in \mathbb{R}^p$ is parameter vector, μ_i being the probability of $x_i = 1$. Find the maximum likelihood estimation for μ.

(c) (1 point)$X \in \mathbb{R}^+$ is a non-negative random variable. Prove that, for any positive real number $a > 0$

$$P(X \geq a) \leq \frac{E[X]}{a},$$

where $E[X]$ is expectation of X.

Exercise 5: Regression

(a) (6 points) Derive the solution for following optimization problem

i. Least Square Regression solves

$$E(\mathbf{w}) = (\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}).$$

Show that the solution is

$$\mathbf{w}^{ls} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

ii. Ridge Regression solves

$$E(\mathbf{w}) = (\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2.$$

Show that the solution is

$$\mathbf{w}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

iii. Lasso

$$E(\mathbf{w}) = (\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) + \lambda\|\mathbf{w}\|_1.$$

Show that the solution is

$$\mathbf{w}^{lasso} = sign(\mathbf{w}^{ls})(|\mathbf{w}^{ls}| - \lambda)^+.$$

[hints : $sign(\mathbf{w})$ gives sign of the element of vector $\mathbf{w}$, i.e., $sign([9, -8, 0]) = [1, -1, 0]$. The modulo function $|\mathbf{w}|$ gives absolution values , i.e., $|[9, -8, 0]| = [9, 8, 0]$ and $(x_i)^+ = max\{0, x_i\}$]

(b) (4 points) Ordinary Least Square (OLS) regression solves

$$E(\mathbf{w}) = (\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}).$$

In gradient descent approach in every iteration the parameter $\mathbf{w}$ is updated using following equation:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \alpha\frac{\partial E(\mathbf{w})}{\partial\mathbf{w}}|_{\mathbf{w}=\mathbf{w}^t} = \mathbf{w}^t - \alpha\mathbf{X}^T(\mathbf{X}\mathbf{w}^t - \mathbf{y}).$$

Derive the update equation of parameter $\mathbf{w}$ to solve following problem using gradient descent approach.

i. Ridge Regression : $E(\mathbf{w}) = (\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$

ii. Weighted least square regression minimize weighted average of square loss of all data points, where weight for the i^{th} data point is r_i then weighted least square regression solves $E(\mathbf{w}) = \sum_{i=1}^n r_i(\mathbf{x}_i^T\mathbf{w} - y_i)^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$

(c) (2 points) Consider the regression problem where the input $\mathbf{x} \in \mathbb{R}^p$ and the output $y \in \mathbb{R}$. Assume that the likelihood is specified in terms of the unknown parameter $\mathbf{w} \in \mathbb{R}^p$ as $p(y|X = \mathbf{x}, \mathbf{w}) \sim \mathcal{N}(\langle \mathbf{w}, \mathbf{x} \rangle, g(\mathbf{x}))$, where $g : \mathbb{R}^p \to \mathbb{R}^+$ is a known non-negative function. We are given the training sample $(\mathbf{x}_i, y_i)_{i=1}^n$ drawn independently according to the joint probability measure P on $X \times Y$. Assume a prior distribution on $\mathbf{w} : \mathbf{w} \sim \mathcal{N}(0, \Lambda)$, where Λ is a diagonal matrix with the diagonal entries given by $\lambda_{ii} \geq 0$). Find the maximum a posteriori estimate of w?

Exercise 6: Classification

(a) (2 points) Consider two non-negative numbers a and b and show that, if $a \leq b$, then $a \leq (ab)^{\frac{1}{2}}$. Use this result to show that, if the decision region of a two-class classification problem are chosen to minimize the probability of mis-classification, this probability will satisfy

$$p(mistake) \leq \int \{p(\mathbf{x}, \mathcal{C}_1) p(\mathbf{x}, \mathcal{C}_2)\}^{\frac{1}{2}} d\mathbf{x}$$

(b) (2 point) Derive the update equation for gradient descent approach to solve following problems

i. Logistic regression: $E(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n log(1 + e^{-y_i(\mathbf{x}_i^T \mathbf{w})}) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2$

(c) (2 points) The error of a binary classifier which guesses completely randomly is 0.5. What is the error of a random k-class classifier for $k > 2$ labels.

i. Random guesser G knows that there are k labels, and for each example, selects a label out of $\{1, ..., k\}$ uniformly at random. What is the error of G ?

ii. Now suppose we have a more sophisticated random guesser Z who knows that w_1 fraction of the data distribution has label 1, w_2 fraction has label 2, and so on. For each example, Z also selects a label out of $\{1, ..., k\}$ at random, but he selects label 1 with probability w_1, label 2 with probability w_2 and so on. What is the error of Z?

(d) (2 points) Consider the following two data distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ over labeled examples. There is a single feature, denoted by X which takes values in the set $\{1, 2, 3, 4\}$ and a binary label $Y \in \{0, 1\}$. $\mathcal{D}_1$ is described as follows:

$$\begin{aligned} Pr(X = i) &= \frac{1}{4}, \quad i \in \{1, 2, 3, 4\} \qquad (1) \\ Pr(Y = 1|X = i) &= 1, \quad i \in \{1, 4\} \\ Pr(Y = 0|X = i) &= 1, \quad i \in \{2, 3\} \end{aligned}$$

D_2 is described as follows.

$$\begin{aligned} Pr(X=i) &= \frac{1}{4}, \quad i \in \{1,2,3,4\} \\ Pr(Y=1|X=i) &= \frac{i}{10}, \quad i \in \{1,2,34\} \end{aligned} \tag{2}$$

i. Consider the following classifier

$$h : h(x) = 1 \text{ if } x > 1.5 \text{ and } 0 \text{ otherwise.}$$

What is the true error of h when the true data distribution is D_1?

ii. Suppose our classifier is

$$h_t : h(x) = 1 \text{ if } x > t \text{ and } 0 \text{ otherwise}$$

. Find t which minimizes the true error of h_t when the true data distribution is D_1. What is the true error of this classifier?

iii. Repeat parts **i.** and **ii.** for the data distribution D2.

2 Programming Exercises

Exercise 1 : File read and write

(1 point) Download iris data from https://archive.ics.uci.edu/ml/machine-learning-databases/iris/. This is perhaps the best known database to be found in the pattern recognition literature. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant (Iris Setosa, Iris Versicolour, Iris Virginica). It contains 4 numeric features and the predicted attribute is class of iris plant.

Write a program (without using any machine learning related in built library) to create another file named as *iris-svm-input.txt* which contain the same data but in the following format

- Each line of the output file contains information corresponding to one sample in te following format

  label index1:value1 index2:value2

- *index* indicates the index of feature and *value* indicates the numerical value of that feature for the sample.
- Skip the *index* if the corresponding *value* is zero
- Consider class labels as follows
  - Iris-setosa : 1
  - Iris-versicolor : 2
  - Iris-virginica : 3
- example : class 1, the feature vector (0.7,1,1,0,2) translates to

  1 1:0.7 2:1 3:1 5:2

2.1 Exercise 2 : Regression

In the following we want to learn a function, $f : \mathbb{R} \to \mathbb{R}$, given some training data $x \in \mathbb{R}, y \in \mathbb{R}$ using least squares,lasso and ridge regression and also we will optimize parameters for Lasso and Ridge regression using 5-fold cross validation.

Write following two functions (do not use scikit, sklearn library function but you can use numpy matrix operation):

(a) (point 1) Implement least square regression with help of matrix inversion

- w = LeastSquares(Featurematrix, $\mathbf{y}$):
  - input: Featurematrix matrix $\phi \in \mathbb{R}^{n \times p}$ and the outputs $\mathbf{y} \in \mathbb{R}^n$ (column vector)
  - output: weight vector w of least squares regression as column vector

(b) (point 2) Implement **stochastic gradient descent** algorithm with step size 0.1 to solve ridge regression

- w = RidgeRegression(Featurematrix, $\mathbf{y}$, λ):
  - input: Featurematrix matrix $\phi \in \mathbb{R}^{n \times p}$, the outputs $\mathbf{y} \in \mathbb{R}^n$ (column vector) and the regularization parameter $\lambda \in \mathbb{R}^+$
  - output: weight vector w of least squares regression as column vector

(c) unzip *regressiondata.zip*. It contains following files

- x : training feature
- y : training label
- xts : test feature
- yts : test label

(d) The relation between x and y can be non-linear. Hence to catch non-linear relationship we will generate Featurematrix=$[1, \mathbf{x}, \mathbf{x}^2, \ldots, \mathbf{x}^{10}]$

(e) (2 points) model selection :

i. Find optimal hyper-parameter λ for ridge regression using 5 fold cross validation on training data. Find out the optimal λ from

$$\lambda \in \{2^{-10}, 2^{-9}, 2^{-8}, \ldots, 2^0, 2^1, \ldots 2^{10}\}.$$

ii. Discuss with increase of regularization parameter how the training error, validation error and test error change. Also report test error for least square regression. Plot three graph showing these three kinds of error where x-axis will show λ and y axis will show corresponding error. [for better visualization use log scale along x axis]

iii. For this experiment use your own implementation of RidgeRegression and LeastSquare. If your implementation for least square and ridge regression is not ready please use any library function to complete this experiment.]